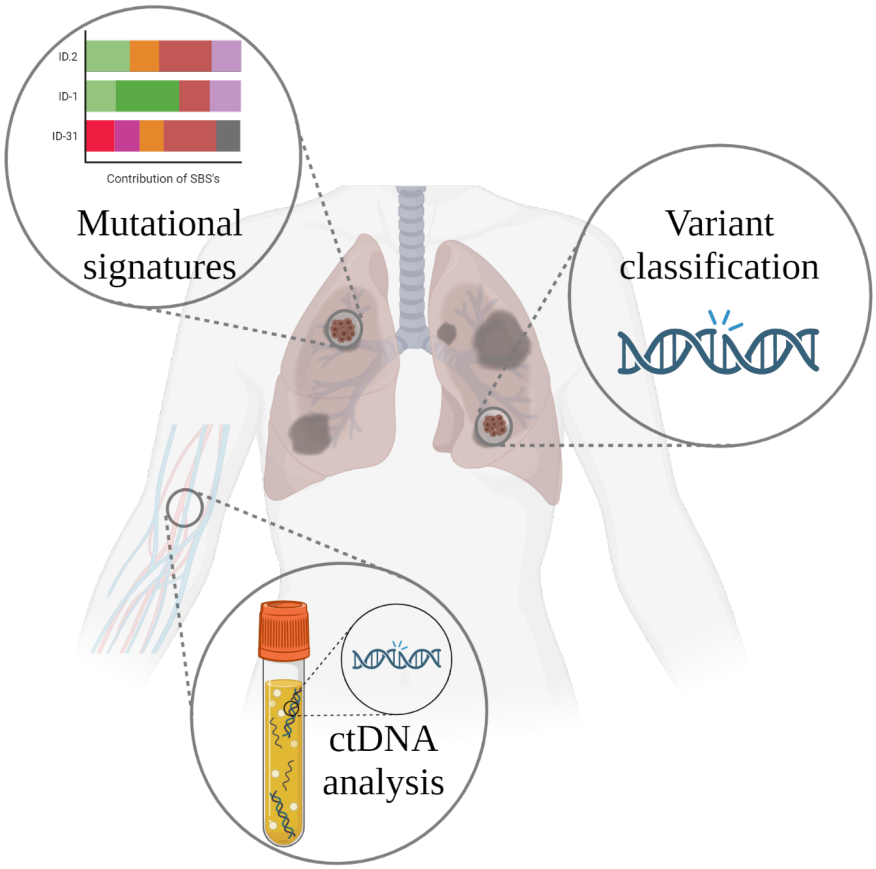
CHALMERS
UNIVERSITY OF TECHNOLOGY



Genetic profiling in non-small cell lung cancer

To predict response to immunotherapy

Master's thesis in Biotechnology

JOHANNA SVENSSON

DEPARTMENT OF LIFE SCIENCES CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Genetic profiling in non-small cell lung cancer

To predict response to immunotherapy

JOHANNA SVENSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Division of Systems and Synthetic Biology
Chalmers University of Technology
Gothenburg, Sweden 2023

Genetic profiling in non-small cell lung cancer
To predict response to immunotherapy
JOHANNA SVENSSON

Supervisor: Anna Rohlin, Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital and Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg

Examiner: Eduard Kerkhoven, Division of Systems and Synthetic Biology, Department of Life Sciences, Chalmers University of Technology

Master's Thesis 2023
Department of Life Sciences
Division of Systems and Synthetic Biology
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of lungs with tumours, giving an overview of analyses done in the genetic profiling. Created with BioRender.com. Typeset in LaTeX

Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Genetic profiling in non-small cell lung cancer
To predict response to immunotherapy
JOHANNA SVENSSON
Department of Life Sciences
Chalmers University of Technology

Abstract

Introduction. Treatment of non-small cell lung cancer (NSCLC) was revolutionised with immunotherapy. Particularly important is immune checkpoint blockade (ICB) targeting PD-1/PD-L1; nevertheless, two thirds are unresponsive to ICB. Better biomarkers are warranted besides the FDA approved tumour mutational burden (TMB). Genetic variants in a few selected genes have been suggested to predict response to ICB alone or in combinations as co-variants in both blood an tissue. This study aims to interpret variants in both blood plasma and tissue, and in addition analyse mutational signatures of the tumours, that might be used as biomarkers.

Material and Methods. The prospective study cohort includes n=50 stage III-IV NSCLC patients that received ICB as first- or second line of treatment. Blood and tumour tissue was sequenced with next-generation sequencing (NGS) with a panel of 591 cancer-associated genes. A comprehensive variant interpretation and classification approach was used to subclass somatic variants into 6 different categories based on standard workflows, in combination with several databases and prediction tools. In addition, mutational signatures were extracted using SigProfiler tools and analysed. For n=26 patients variants were also monitored in blood during treatment with ICB using ultrasensitive methods for variant identification.

Results and Discussions. In total 859 true variants were identified. These included 40 pathogenic, 96 likely pathogenic and 685 variants of unknown significance (VUS). The VUS:es were further subclassed into different categories to identify those with higher or lower driver properties and probability of pathogenicity. By using this approach 34 VUS++ and 75 VUS+ were identified. Frequently mutated genes, number of variants in different classes and their pathogenicity were related to ICB response, as was mutational signatures and levels of ctDNA at various timepoints.

Conclusion. Understanding the genetic landscape and identifying biomarkers of ICB are key considerations in development of personalised treatment. The approach of a thorough classification including subclassification of the VUS:es led to identification of variants that can potentially function as biomarkers, in combination with other. Mutational signature analysis lead to differentiation of tumour types. The analysis of the combination of mutational signatures and genetic variants further enhanced refinement of biomarkers of response to ICB. Monitoring variants in ctDNA is a molecular tool for early identification of response or progress during treatment in NSCLC.
Keywords: biomarker, ctDNA, ICB, mutational signature, NSCLC, variant classification.

Acknowledgements

First of all I would like to thank my supervisor Anna Rohlin that I got the opportunity for doing this project and believing in my ideas as well as teaching me everything I needed to know. For all interesting discussions about research and the BioLung project in particular. Without you, this project would not be possible; you have made the start of my journey towards becoming a scientist better than I could ever imagine, and for that I am incredibly thankful.

Thanks to everyone at Clinical Genetics and Genomics at Gothenburg University and Sahlgrenska University Hospital for warmly welcoming me to the department. A special thanks to Maria Yhr and Jennie Gaarder for teaching me to lab with blood and plasma and helping me out in the lab. Also, thanks Angelica Bergström for interesting and helpful discussions and handling of ctDNA data during the autumn. Thank you, Katarina Truvé for the bioinformatics of mutational signatures and Marcos Díaz Gay for instructions and evaluations of them. Nikita Dutta, Sukanya Raghavan, Andreas Hallqvist, Ella Äng Eklund and Anna Rohlin, along with the rest involved in the BioLung project, thank you for helping me out and letting me participate in this project. Last but not least, I would like to thank the patients, without them, this research would not be possible.

Even though this thesis have taken a lot of time and work, I have learnt so much and had a lot o fun while doing it, more than I could have imagined. Lastly, I would like to thank my friends and family for always supporting me.

Johanna Svensson, Gothenburg, May 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

cfDNA	Cell-free DNA
CNV	Copy number variant
CPV	Consensus pathogenic variant
CT	Computed tomography
ctDNA	Circulating tumour DNA
DBS	Doublet base substitutions
EMA	European Medicines Authority
FDA	U.S. Food and Drug Administration
FFPE	Formalin-fixed paraffin-embedded
GoF	Gain of function
HR	Hazard ratio
ICB	Immune check-point blockade
ID	Small insertions and deletions
IGV	Integrative Genome Viewer
Indels	Insertions or deletions
KM	Kaplan-Meier
LoF	Loss of function
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MM	Mutant molecules
NGS	Next generation Sequencing
NMF	Non-negative matrix factorisation
NSCLC	Non-small cell lung cancer
OS	Overall survival
PD	Progressive disease
PD-1	Programmed death protein 1
PD-L1	Programmed death-ligand 1
PFS	Progress-free survival
PR	Partial response
TMB	Tumour mutational burden
SBS	Single base substitutions
SD	Stable disease
SNV	Single-nucleotide variant
VAF	Variant allele frequency
VUS	Variant of unknown significance

Contents

List of Figures

List of Tables

1
Introduction

One of this and future generations' greatest issues to try to undertake is cancer. Together with an aging population, cancer cases are increasing, leading to a greater need for effective and well-functioning treatments. In fact, the second largest proportion of cancer-related deaths is due to lung cancer, and most patients suffer from non-small cell lung cancer (NSCLC) [1]. Although treatments have been revolutionised with the introduction of immunotherapy [2], this field of research is of importance to explore and understand further [1]. By analysing the variants in the tumour cells, the immunotherapeutic treatments can be designed for a patient group, eventually leading to more personalised treatment.

This thesis is a part of the BioLung project, which is a NSCLC patient cohort study that is conducted in collaboration between Sahlgrenska University Hospital and University of Gothenburg. The vast majority of patients in the BioLung cohort are diagnosed at advanced stages of NSCLC.

2
Theory

In the following section the most relevant theoretical background for the thesis is described. The background includes concepts as immunotherapy, somatic variants, biomarkers, circulating tumour DNA (ctDNA) and mutational signatures.

2.1 Non-small cell lung cancer

Lung cancer can be divided into two main subtypes, small cell lung cancer and NSCLC, where 80-85% of cases belongs to the latter subtype [3]. Worldwide and in the Nordic countries, among malignant tumours, lung cancer has the highest mortality and morbidity rate [2, 3], with 1-8% as 5-year overall survival (OS) for stage IV patients within the US [3]. Yearly, lung cancer contributes to over 12 000 deceased in the Nordics [2]. In NSCLC there are two main histological subtypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) [2]. Most patients are diagnosis when they already have entered stage IV with metastases. Few get diagnosed in early stages I or II, when the tumour has a limited spread; the rest are diagnosed when the tumour has a locally advanced spread at stage III. Even though NSCLC has a high mortality rate, diseases in stages I, II or III are able to cure, but the cure rates are low [2]. There are different treatments for NSCLC, where the main ones are surgery, chemo-, immuno- and radiation therapy [4]. However, for the later stages (III and IV) the tumour is inoperable, and surgery is no longer an option [2]. For most NSCLC patients in earlier stages (I-II) there are no indications for need of immunotherapy.

2.2 Clinical concepts and terms

Clinical concepts as response measurement with RECIST 1.1 criteria, and endpoints for clinical trials are presented.

2.2.1 Measurement of treatment response

RECIST 1.1 is an objective measure used to estimate treatment response [5]. RECIST 1.1 is based on measurement of solid tumours from imaging methods e.g. computed tomography (CT) scans. The tumour is measured where the diameter is at the longest and targeted lesions are both primary tumour and metastases. The reference value to compared change with is the sum of diameters of followed lesions,

often at baseline. Definitions and concepts of RECIST 1.1 criteria for targeted lesions are [5]:

- **Complete response (CR).** Total disappearance of targeted lesions.
- **Partial response (PR).** Targeted lesions have reduced by ≥30% from reference at baseline.
- **Progressive disease (PD).** An increase of ≥20% compared with the patient-overall minimal reference value. PD also require that the sum of diameters of targeted lesions have increased by ≥5 mm.
- **Stable disease (SD).** Change is not enough to be PR nor PD.

2.2.2 Endpoints in clinical trials

To test if cancer therapies are effective, clinical endpoints must be determined [6]. The most well-known endpoint is to use the objective measure OS, counted from baseline, such as treatment start, until death. However, OS has disadvantages that some might survive a long time even though early PD and includes non-cancer deaths even though positive clinical response. Nevertheless, prolonging life is what most treatments are used for, making OS the golden standard. Other measurements are also suggested such as progression-free survival (PFS). PFS uses one of two endpoints, either death or disease progression. The advantage of using PFS is to assess short-term, although, longer PFS does not consistently correlate to longer OS.

2.3 Immunotherapy in non-small cell lung cancer

Immunotherapy is a treatment where a person's own immune system is used. A great improvement in immunotherapy occurred with the introduction of antibodies called immune check-point blockade (ICB) [1]. ICB is especially important for NSCLC patients, with inhibitors targeting and blocking programmed cell death-ligand 1 (PD-L1) on tumour cell or programmed death protein 1 (PD-1) on immune cell, see figure 2.1. However, most NSCLC patients do not respond to ICB treatment [2], and for some patients ICB can cause immunotoxicity or lead to an accelerated disease progression [3]. The mechanism of action of PD-L1 is to block the immune response by binding to PD-1[7]. The blockade leads to the antibody being unable to bind the antigen and destroy the tumour cell [7]; hence, high PD-L1 expression is associated with suppression of the adaptive immune system [4]. Using ICB treatment with antibodies targeting PD-L1 or PD-1, the blockade to immune response is removed. Thereby, anti-PD-L1 or anti-PD-1 treatment leads to the immune cells being able to reach and destroy the cancer cells [3]. A high expression of PD-L1 is also to some extent associated with worse outcome in NSCLC [7].

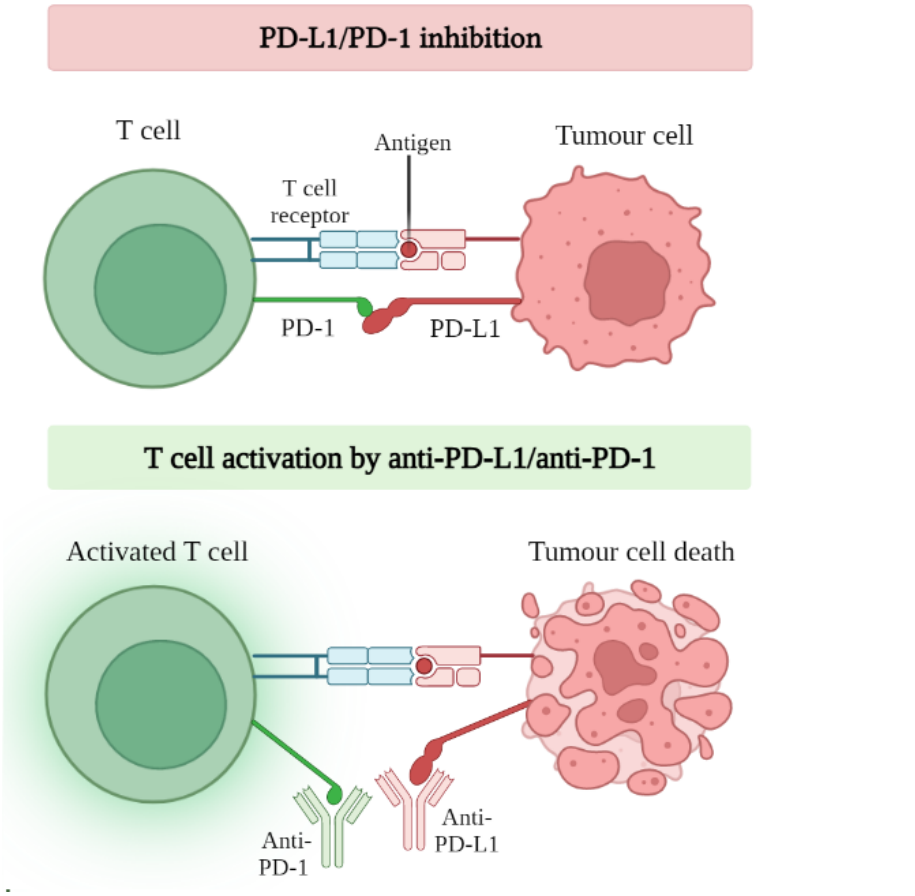


Figure 2.1: Mechanism of action for PD-1/PD-L1 inhibitors. The top half shows inhibition of PD-1 by PD-L1; the lower half shows how the T cell is activated by inhibition of PD-1/PD-L1, leading to destruction of the tumour cell. Created with BioRender.com.

2.3.1 Immune check-point blockade

Treatment decisions for NSCLC are based on factors as general health, lung function, stage at diagnosis, treatment history and PD-L1 expression [2]. PD-L1/PD-1 inhibitors are a group of monoclonal antibody treatments used in cancers and in the BioLung cohort four immunotherapeutic treatments with different active substances are used. Anti-PD-1 treatments are pembrolizumab and nivolumab and anti-PD-L1 antibodies are durvalumab and atezolizumab [2]. ICB is given intravenously every second to sixth week, depending on the active substance [8]. Side effects are common, can occur in many organs, e.g. liver, lungs and skin, and are characterised by autoimmunity. Depending on the seriousness of the side effects the treatments might have to be stopped or changed. Hence, two thirds of patients receiving ICB do not respond [2], since side effects are undesirable giving immunotherapy only to patients that will likely respond is preferable.

2.4 Genes and variants

2.4.1 Nomenclature

The standard terminology will be shortly described, to clarify any inconveniences of nomenclature used. In all cases, as the terms mutation, mutant and polymorphism have underlying assumptions of degree of pathogenicity, a genetic alteration will be referred to as a variant, in agreement with ACMG [9, 10]. Genes consist of coding-regions i.e. exons and non-coding regions i.e. introns, splice sites are present at ends of exons and essential in translation to mRNA [10]. Upstream of a gene refers to the 5'-end of the coding strand, while downstream of a gene is towards the 3'-end, see figure 2.2. When describing up- or downstream of an exon +/- is used, respectively; splice sites are present at flanking positions including -2/-1/+1 and +2 from exons.

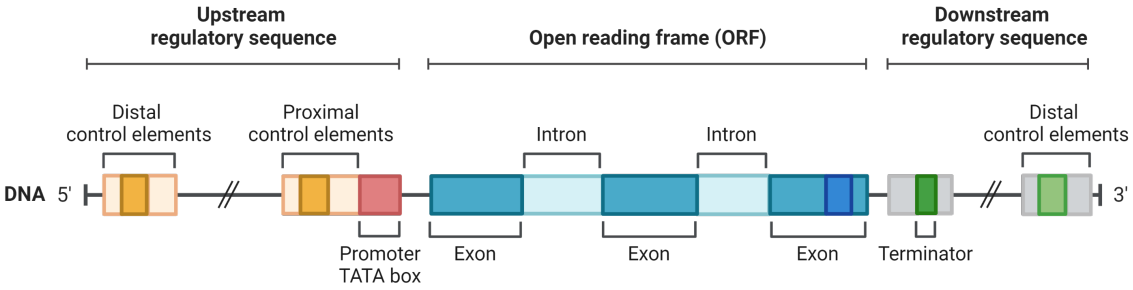


Figure 2.2: Schematic illustration of an ordinary gene. Created with BioRender.com.

Deletions or insertions can lead to the open reading frame shifting resulting in a frameshift, or the indel can alter whole triplets causing inframe variants. Loss-of-function (LoF) variants includes frameshift, stop-gain and splice site variants, non-LoF variants are synonymous, non-synonymous, inframe indels, startloss and stoploss variants. A protein change is denoted by p., e.g. the variant *KRAS*: p.G12C is a variant in the gene *KRAS* and indicate that the amino acid glycine (G) at position 12 is changed to cysteine (C).

2.4.2 Somatic variants

Somatic variants are, unlike germline variants, not congenital, but instead occur spontaneous during the lifetime [11]. Somatic variants can occur in all cells, except germ cells, and are associated with cancer and other developmental disorders. Nevertheless, most somatic variants are harmless. Although, if accumulated over a long period of time, a somatic cell can escape its intended function and uncontrollably clone itself, eventually causing cancer.

In the genome, genetic alterations can appear in many ways, which can be divided into one of two categories [12]. The first category containing changes to the DNA sequence such as single-nucleotide variants (SNVs) and small insertions or deletions (indels). The other category containing larger fractions and changes such as fusion

genes, copy number variants (CNVs) and large indels [12], also essential in cancer [13].

2.4.3 Functional effects of variants, oncogene and tumour suppressor gene

Genetic variants can lead to the protein losing its original, changing, or gaining a different function [11, 14], or the variant is silent and has no effect on the protein [11]. To gain knowledge in tumour development, it must be establish if the variant is present in an oncogene or a tumour suppressor gene [15]. One group of genes are called proto-oncogenes, if in which a variant occurs, turns into an oncogene [16]. The normal function of a proto-oncogene includes regulation of cell differentiation, division and death; all of which becomes uncontrolled if mutated into an oncogene. Essentially, an oncogene can make a healthy cell develop into a cancer cell [17]. In contrast, the normal function of a tumour suppressor gene is to protect healthy cells from growing uncontrollably and turning into cancer cells. If a variant occur in a tumour suppressor gene, the gene's function is predicted to be inactivated, resulting in a LoF [11, 14]. While, if a variant is in an oncogene, a gain-of-function (GoF) is expected for the gene. The change-of-function variant is rare but can also occur [14]. The functional effect of a variant in a gene can be different based on the characteristics of the gene.

A few numbers of genes exhibit both oncogenic- and tumour suppressor characteristics and is referred to as double agents [17]. Kinases and transcription factors are the most common protein types in which some proteins act as double agents in cancer. A variant in a double agent could either increase or decrease activity. In some proteins, the change of activity is dependent on the domain in which the variant occurs. However, which attribute each double agent gene exhibit needs to be determined for each tumour type [14, 17].

2.4.4 Variants in cancer

In cancer an accumulation of variants occurs, however not all variants are driving the development of the tumour [15, 14]. Variants are characterised as either drivers or passengers [15]. A driver is defined as a variant that is crucial in development of cancer [14], whereas a passenger variant is a variant without any phenotypical or biological effects [11]. The importance of classification is based on the complex nature of tumour development [14]. Since variants are accumulated, it is of importance to distinguish which drive and start the cancer development.

Understanding the variants in a tumour is necessary to be able to classify them accordingly. The classification system for each variant often includes the classes pathogenic, likely pathogenic, benign, likely benign or variant of unknown significance (VUS). The mentioned classes are widely used clinically and is based on the consensus The American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) guidelines and standards [9]. ACMG and AMP guidelines are widely used for interpretation of germline and somatic variants. Tools to predict functionality of variants is combined with knowledge-based

databases and included in ACMG and AMP criteria.

A hotspot is defined as a genetic position frequently mutated in cancer [18]. Driver variants in oncogenes are found in specific hotspot positions, for example in the *RAS*-family genes (*KRAS*, *HRAS* and *NRAS*) p.G12 is a hotspot [19]. Hotspot positions are not found at the same extent in tumour suppressor genes as in oncogene [18]. Tumour suppressor genes rather becomes harmful due to LoF variants or if present in certain exons or domains[19, 18, 20]. Proofreading genes *POLE1* and *POLD1* are examples of tumour suppressor genes where pathogenic or harmful variants are restricted to exons in a certain domain of the protein [20]. Also, hotspots are not necessarily excluded to exons, variants with driving properties have been found outside of coding regions, for example in *TERT* promoter regions [19, 18]. Oncogenic driver variants in NSCLC have been identified during the recent years and more are expected to be found and make treatable [21]. Specific driver variants have treatments approved by the U.S. Food and Drug Administration (FDA), the genes are *EGFR*, *MET*, *ALK*, *ROS1*, *BRAF*, *NTRK*, *RET*, *HER2*, *NRG1* and *KRAS*. Nevertheless, a third of the NSCLC cases seen has no known driving variant.

Some of the the mentioned oncogenic driver genes are mutually exclusive, meaning that the probability of them co-occurring is very low [22]. Why mutual exclusivity occur is not fully understood but believed to be dependent on protein interactions in tumour types. However, if one variant is found in an oncogene, it is believed that the signalling pathway is already switched on. There are two established hypotheses for mutual exclusivity, the first suggesting that two variants would not benefit the cell and the second is that expressing two activating oncogenes leads to cell death. Even though the oncogenic drivers, e.g. *KRAS* and *EGFR*, are thought to be mutually exclusive, there are few exceptions of cases where co-occurring variants are found [23].

2.5 Established biomarkers in non-small cell lung cancer

Biomarkers can be specified either as agnostic, predictive or prognostic, to mention some, and treatments can be decided upon the detected level of a specific biomarker [24]. An agnostic biomarker is an indicator regardless of tumour origin, a predictive biomarker predict the livelihood of response to a specific treatment and a prognostic biomarker indicates the outcome regardless of treatment. Currently, a few biomarkers are approved for prediction of treatment and diagnostics in NSCLC [2]. Nevertheless, there is conflicting evidence whether the used biomarkers are accurately able to predict response to ICB. The most used predictive biomarker for response to ICB is, as mentioned, tissue PD-L1 expression [24]. For example, radiation or chemotherapy treatment can affect and modify the expressed level of PD-L1 [7]. Although, PD-L1 expression is an unreliable biomarker [24], it is approved for clinical use for determination of treatment by the European Medicines Authority (EMA) [2].

A suggested agnostic biomarker for NSCLC is the tumour mutational burden (TMB), which refer to the number of mutated genes per megabase (Mb) of DNA in

the sequenced tissue sample [13]. TMB is however also a conflicting biomarker with varying results [13]; that is because there are different lab-specific calculations for this parameter [24]. For example, some calculations only consider non-synonymous variants while other consider all types of variants. Without standardisation, TMB can be seen as a biased biomarker; despite this, TMB is approved as an agnostic biomarker by FDA[2].

2.6 Circulating tumour DNA

Whole blood contains erythrocytes, leukocytes and thrombocytes that are encapsulated by plasma. The blood plasma contains, among other components, a small proportion of cell-free DNA (cfDNA) [25]. The cfDNA represent all DNA in the cells, additionally in cancer patients, some of the cfDNA comes from tumour cells and is known as ctDNA. CfDNA is found in very small concentrations, approximately 10-30 ng/mL [26]. The ctDNA can be used for quantitative analysis, leading to the possibility to analyse the prevalence of specific variants in ctDNA [25, 26]. There are different methods that can be used for detecting ctDNA, the methods have in common that they are ultrasensitive to be able to detect the small amounts of ctDNA [26, 27]. Usually, one of two approaches can be used, one is to monitor already known variants over time to e.g. follow response to treatment, while the other is to use a panel and screen for variants to identify a tumour or relapse.

In cancer patients, both cfDNA and ctDNA is found, the detected levels varies between patients and tumour type [26]. CfDNA leaks from cells undergoing cellular processes such as cell death and active secretion. Not only can cfDNA be found in blood, but also other bodily fluids, such as sweat and saliva. Characteristics of cfDNA is that its length is usually 146 bp, wrapped around a histone, connected by a 20 bp long DNA-linker. The fragments can be longer if it comes from other processes than apoptosis. A blood sample with cfDNA is a instantaneous measurement, since cfDNA is quite unstable with half-life is between 15-150 min [26]. The process of cfDNA clearance is a biological concept that is poorly understood, the clearance does however take place in spleen, liver and kidneys. Cf- or ctDNA clearance has no consensus definition but is study-specific and can be defined as the lack of detectable variants, given a decent coverage [28]. CtDNA clearance is associated with longer PFS and OS in NSCLC patients, the clearance possesses potential of being a predictive and prognostic biomarker. Also, driver variant clearance is connected to longer OS, but total ctDNA clearance is connected to even better OS.

As mentioned, two main different ctDNA approaches are suggested, screening a panel with several variants of known drivers or use specifically selected variants based on the variants detected in the tumour [29, 30]. As a third of NSCLC patient has an unknown driver [21], to select what variants to follow patient- or tumour-specific variants is suggested as an approach [29]. CtDNA analyses using panels with the most common hotspots has also shown clinical applicability for NSCLC patients but require optimisation [30]. However, the latter approach is suitable for minimal invasive molecular pathological testing for treatment considerations more than treatment monitoring.

CtDNA is reported using mutant molecules (MM) per mL plasma and variant

allel frequency (VAF) in percentage [31]. Both VAF and MM/mL plasma is affected by the input of cfDNA, which can lead to underestimation due to contamination of germline or wild-type DNA molecules. MM/mL plasma is supposed to give more accuracy in terms of tumour burden compared to ctDNA VAF. To calculate MM/mL plasma equation 2.1 is used, where MM/mL plasma is dependent on PCR input in ng and VAF in % is divided by 0.033 ng per haploid genome and the plasma volume V_{plasma} in mL. MM/mL plasma is always rounded down to the nearest whole value.

$$\text{MM/mL plasma} = \frac{\text{PCR input} \cdot \text{VAF}}{0.033 \cdot V_{plasma}} \tag{2.1}$$

Physical quantities MM/mL plasma and VAF have shown to be in analytical agreement, thus both should be reported as one might not be sufficient[31].

2.7 Mutational signatures

2.7.1 Biological impact and aetiology

Tumours with somatic variants often affect genes involved in cellular processes important in for example, replication or repair of DNA [32]. Somatic variants can be caused by external factors such as UV-exposure or tobacco smoking. Combinations of variant types can be generated from specific mutational processes; these are called mutational signatures. What distinguishes and characterises a mutational signature can be the type of variant, its context or distribution, the reconstructive ability or stage of the cancer [11]. Mutational signatures are sorted as either single base substitutions (SBS), doublet base substitutions (DBS), small insertions and deletions or CNVs [32]. Each reference signature consist of a unique pattern, of which the sample-specific signature are composed of. The reference mutational signatures are developed by the Wellcome Sanger Institute (Hinxton, UK), and is a project within Catalogue Of Somatic Mutations In Cancer (COSMIC) [33]. Some of the COSMIC signatures have a proposed aetiology related to factors such as tobacco smoking, age and DNA mismatch/repair which can be used for analysis [32]. Specific cancer types have some signatures higher expressed than other types. Apart from offering deeper understanding of mutational processes and cancer aetiology, mutational signatures are suggested to have potential to function as predictive biomarkers [34]. In addition, mutational signatures can be used to discover therapy sensitivities and explore biological consequences of driver variants.

2.7.2 Mathematical background and algorithms

The combinatorics behind mutational signatures are basically the same for DBS and SBS, thus will be described for SBS [35]. Generally, SBS can be sorted in 6 main combinations, being C-G, C-T, C-A, G-T, G-A, A-G or A-T base shifts, where e.g. variant C-G represent both C>G and G>C. To characterise the variants even further, it is determined where the variant is present in correlation to the 5'- and 3'-ends by the surrounding nucleotides. The types for SBS can then be determined to 96; for 6 subtypes, 4 possible 5' nucleotides and 4 possible 3' nucleotides ($6 \cdot 4 \cdot 4 = 96$).

96 are the most used types, since that is what COSMIC consists of, but without consideration of flanking bases 6 types are created, while 2 flanking bases on each side result in 1536 bases, to mention some [36]. Using additional flanking bases can lead to discovering novel mutational signatures and further understand mutagenic processes.

Moreover, SigProfiler is collection of bioinformatics tools to use for discovering and deciphering signatures, and the general algorithm will be shortly described [35, 37, 36]. Mathematically, each mutational signature is dependent on the number of variants i.e. exposure; and the mutational process which in turn depends on probability of the signature to be true [35]. The mutational catalogue of signatures can be expressed as a matrix, where the mutational catalogue, M, is the product of the mutational processes, P, multiplied with their exposure, E, see matrix notation in equation 2.2.

$$M = P \times E \tag{2.2}$$

Non-negative matrix factorisation (NMF) is a method applied and used on biological data due to its ability to extract complex information [35]. Using NMF, non-biological data and unidentified noise can be sorted out. When signatures are reconstructed then cosine similarities for each sample is given. The cosine similarity describe the average deciphering error, i.e. similarity between two non-negative mutational profiles A and B, see equation 2.3, where K is the number of variant types [35]. The cosine similarity varies from 0 to 1, where 1 is the exact same signature structure [35, 38]. A cosine similarity below 0.75 can happen by chance and when above 0.90, it is very likely to be true positives [37]. A low cosine similarity may be due to errors due to insufficient calling of variants, too few variants or the sample possessing a novel signature.

$$sim(A, B) = \frac{\sum_{k=1}^{K} A_k B_k}{\sqrt{\sum_{k=1}^{K} (A_k)^2} \sqrt{\sum_{k=1}^{K} (B_k)^2}} \tag{2.3}$$

SigProfilerExtractor is a completely unsupervised machine-learning method, meaning it can discover hidden patterns in large scale data [37], while SigProfilerAssignment is a supervised method which is based on already known patterns, i.e. the COSMIC signatures, that the data is then fit into (unpublished). COSMIC signatures has also been linked together into subgroups according to aetiology, to counteract overfitting, subgroups can be excluded. To avoid overfitting of the model, signatures with unlikely aetiology in terms of not arose from true biological processes, can be excluded. Excluding subgroups can also result in a biased solution if not excluded correctly, while overfitting might be an issue if keeping the subgroup. Overfitting is also avoided with NMF in the program itself every iteration through bootstrapping before NMF [35]. However, excluding subgroups can in turn also lead to biased results and not detecting outlying samples (unpublished).

2.7.3 Formalin-fixed paraffin-embedded samples

Formalin-fixed paraffin-embedded (FFPE) samples have some mutational signatures whose true aetiology cannot be connected to true biological processes, but rather

degradation of a tissue in FFPE [39]. The main reason for storing tumour material in FFPE is to preserve tissue morphology for clinical diagnosis. Most commonly, the FFPE variants are variants where C is mutated to T, which results in false positive mutational signatures. To try to increase the quality of the FFPE tissue, it can be chemically repaired before sequencing. In degrading FFPE samples two signatures are correlated, SBS1 and SBS30, these are biologically caused by deamination, i.e. removal of amino group. The true deamination can be mistaken for FFPE-induced deamination for SBS1; without chemically repairing the tissue before sampling, the variants can be mistaken for the rare signature SBS30. To determine true biological processes from FFPE-induced artefacts, the artefact variants can be computationally removed [39].

2.8 Bioinformatical workflow

The main steps of a bioinformatical workflow is seen in figure 2.3.

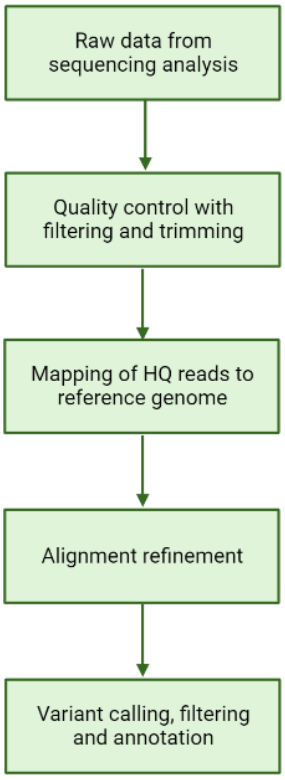


Figure 2.3: Bioinformatical workflow for a general sequencing run. Created with BioRender.com.

When a sequencing run is done, a lot of raw data is achieved [40]. Following base calling is done and each read is given a quality score reflecting a probability of the base being true. High quality reads reflect a score of above 30, with 99.9% certainty. Then various filters were applied, trimming of reads and mapping to reference genome were done, to in the end achieve vcf-files with the called variants.

2.9 Statistical methods

In cancer studies survival analyses are done to be able to compare outcomes and its contributing factors [41]. Clinical endpoints can be OS or PFS, characterised by events, i.e death or progress [6, 41]. Statistical methods based on normal distribution is not useful in the cases for survival analyses [41]. A phenomenon specific to survival analyses is called censoring and can be due to patients lack follow-up data, suffered a non-cancer related death or the event has not occurred. To correctly handle censoring, special methods are required. Special non-normal statistical methods are Kaplan-Meier (KM) survival analysis, logrank tests and Cox regression multivariate analysis of proportional hazard for contributing factors.

KM survival function is described in equation 2.4, where $S(t_j)$ is the probability of a patient being alive at time point t_j, d_j is number of events at time point t_j and n_j is the number of patients alive right before time point t_j [41]. Using KM survival analysis, a KM survival curve for cumulative survival probability over time is used to visualise survival [41].

$$S(t_j) = S(t_{j-1})(1 - \frac{d_j}{n_j}) \tag{2.4}$$

Comparison between groups in survival curves can be made statistically with a logrank test, a nonparametric test [41]. Logrank tests is used to calculate probability of the null hypothesis being true. The null hypothesis for logrank test is that the hazard ratio (HR) is equal to 1, when HR is defined as the relative survival between groups. Logrank tests is then used to estimate if a one group is significantly different compared to another group.

Cox regression is a statistical method for testing multivariate [42]. The Cox statistical model is presented in equation 2.5, where h(t) is the hazard function, x is a covariate, b is the size of the covariate, p is the number of covariates. Unlike logrank tests, Cox regression model takes the size of the effect and also consider a clinical assessment of the impact.

$$h(t) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + ... + b_px_p) \tag{2.5}$$

2.10 Aim

The thesis has an overall aim which is divided into two partial aims that combined are meant to fulfil the overall aim. Overall, the thesis aims to give insight in identification of biomarkers in blood and tumour tissue, related to its effectiveness in prediction ability, for use in clinical applications.

- The first partial aim is to predict and understand patients' response to ICB treatment. Individual variants as well as mutational signatures will together with additional information be used to analyse the response to treatment in tumour tissue samples.
- The second partial aim is to find biomarkers in the blood plasma samples, to investigate if they can be used to predict and understand the patient's response

to treatment. This aim will be fulfilled by interpreting results from ctDNA analyses.

3

Methods

The following section will describe the set-up and design of the study followed by the methods used. An overview of next generation sequencing (NGS) and bioinformatics will be given, followed by an in-depth variant interpretation and classification. How the mutational signatures were produced will also be described. The two different sequencing methods used for ctDNA will be presented, and finally how the pathway analysis was performed.

3.1 Study design

The BioLung cohort is an ongoing prospective observational study, started in 2019 and where subjects (i.e. patients) are still included. Patients are recruited from both Sahlgrenska University Hospital and Skövde Hospital. Clinical follow-up is done every three months using computed tomography scan (CT scan). If responsive to immunotherapy, the patient is treated at longest for 24 months. For this study, 9 months is used as a cut-off where responders and non-responders are evaluated. SD for (at least) 9 months together with CR and PR are responders, while PD within 9 months is a non-responder.

Liquid biopsies in form of blood samples are collected at every treatment visit for the first five treatment cycles, including baseline (A (baseline), B, C, D, and E); schematic diagram of sampling times is presented in figure 3.1. Depending on treatment, the cycles can be between 2-4 weeks. If a clinical progression of disease is detected, a sample will be collected at the patient-specific time point F, if no clinical progression is seen a follow-up sample will be collected 1 year from baseline. To clarify, F does not necessarily follow E. F can be taken after the 1-year sample or even before the first five treatment cycles are completed in case for an early progression.

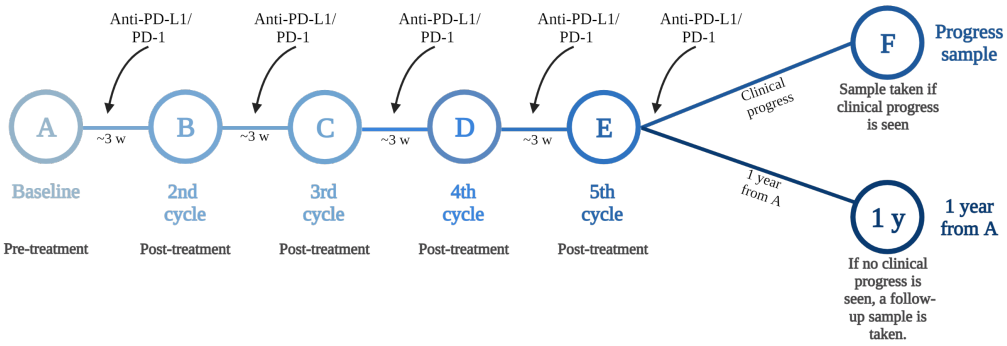


Figure 3.1: Flowchart over study design and sampling in the BioLung cohort study. Created with BioRender.com.

The tumour tissue was retrieved at diagnosis of the patient via a needle biopsy, at a patient-specific time point somewhere before A. The tumour sample was then analysed using a clinical NGS-panel that include hotspots with potential to influence treatment decisions. For a subgroup of patients, the samples are sent to Eurofins Genomics (Europe Sequencing GmbH, Ebersberg, Germany) for further analysis using a larger sequencing panel, more described in section 3.2. The biopsy is taken at time for diagnosis, following the time from tissue biopsy to inclusion in BioLung is not standardised based on the study, as the liquid biopsy.

The patient groups in the project are split into subgroups based on what samples are available. The main divider is if there are extended NGS data available, see figure 3.2. Further, the mutational signature analysis contain 50 patients, there after the variant classification is done, and the data from there is used to select variants to follow for ctDNA analysis. Where extended NGS data is missed, samples were selected for ctDNA analysis with variants based on the clinical NGS data.

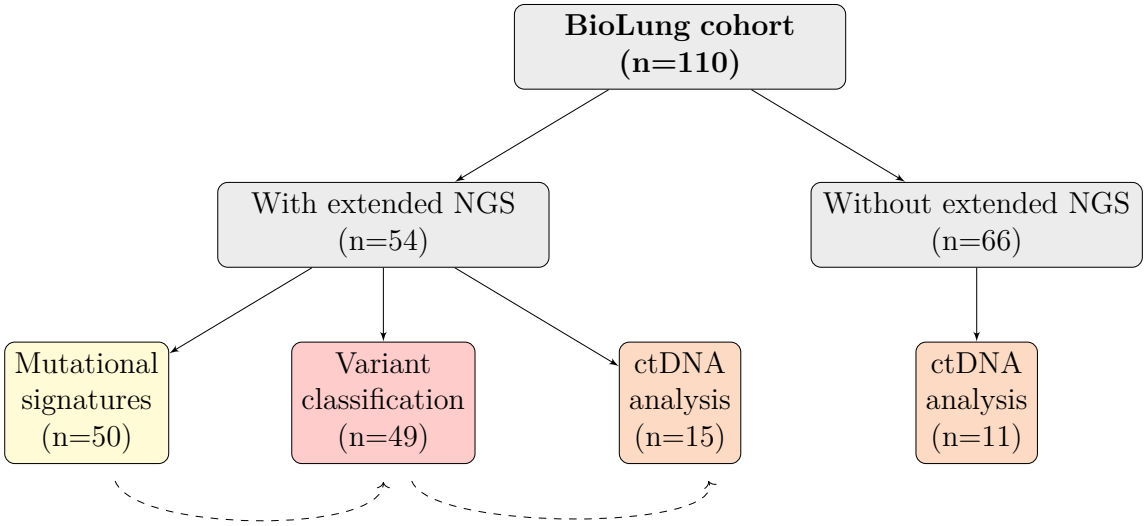


Figure 3.2: Overview of patient selection for the different parts of the project, where n is the number of patients included.

3.1.1 Metadata

Tumour biopsy samples from 53 patients have undergone NGS analysis. 2 of the 53 patients were then excluded from the studied since it was determined they were misdiagnosed with NSCLC; one originally had kidney cancer and the other had small cell lung cancer, resulting in 51 patients. 1 of the 51 patients moved to another country, hence were also excluded due to lack of follow-up data, see figure 3.3 for CONSORT diagram of patient selection to the NGS sub cohort.

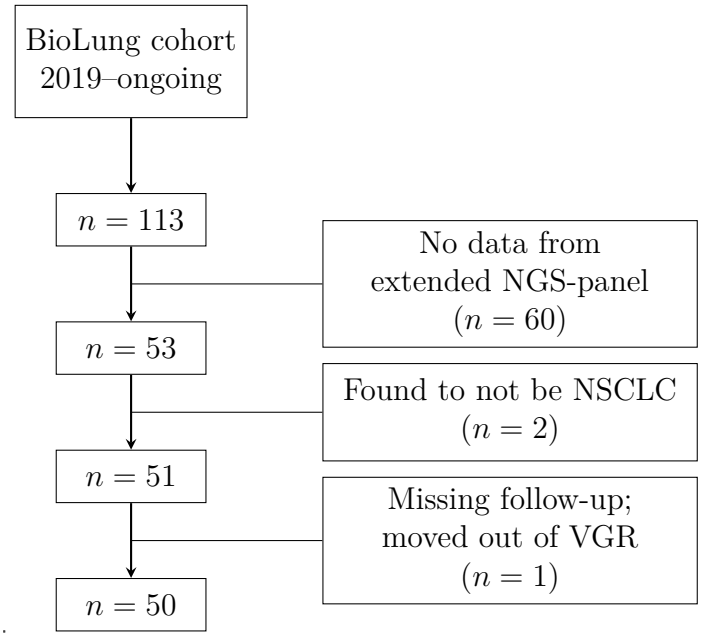


Figure 3.3: CONSORT diagram of patient selection for the analyses. Number of patients included are shown together with reasons for exclusions.

Table 3.1 shows the patient selection for variant classification as well as all patients together with parameters characterising the cohort, parameters such as histological diagnosis, PD-L1 expression and response to ICB. Note that only 49 patients are presented in the NGS sub cohort, while the CONSORT diagram states 50 patients for the same analysis. One patient was excluded during the project according to an analysis performed, the reason of exclusion will be explained in the results.

Table 3.1: Metadata for patients from the cohort. NGS, i.e. included in variant classification and mutational signatures of BioLung study and for the subcohort for ctDNA analysis

	n (%), NGS	**n (%), ctDNA**
Patients	49 (100%)	26 (100%)
Sex		
Female	27 (55%)	14 (54%)
Male	22 (45%)	12 (46%)
Age		
≥70	19 (39%)	11 (42%)
>70	30 (61%)	15 (58%)
Median	72	74
Range	(36–86)	(53–83)
Stage		
I	1 (2%)	2 (8%)
III	6 (12%)	8 (31%)
IV	42 (86%)	15 (58%)
Histodiagnosis		
LUAD	36 (73%)	21 (81%)
LUSC	12 (24%)	4 (15%)
NOS	1 (2%)	1 (4%)
PD-L1 expression		
High (≥50)	23 (47%)	12 (46%)
Low (<50)	26 (53%)	12 (46%)
Response to ICB		
Responder	28 (57%)	11 (42%)
Non-responder	21 (43%)	13 (50%)
Smoking history		
Smoker (current or previous)	44 (90%)	25 (96%)
Non-smoker	5 (10%)	0 (0%)

3.2 Next generation sequencing and bioinformatics

The material from tumours were extracted from a BioLung patient in the cohort and analysed clinically using NGS, as mentioned. Either the primary or metastasised tumour tissue is sampled during a needle biopsy, then stored in FFPE until sequenced. For samples in the NGS subgroup, the samples were sequenced by Eurofins Genomics. The sequencing panel used was INVIEW Oncoprofiling (Eurofins Genomics), which contains around 600 protein coding genes with a limit of detection of 1% on indels and SNVs. The FFPE tissue was sequenced using the genetic sequencer Illumina HiSeq (San Diego, CA, USA). Eurofins Genomics also provided vcf-files with the called variants after processed through their bioinformatical pipeline. During this project, Eurofins Genomics developed a new bioinformatic pipeline, thus all samples were run through both pipelines. Only high-quality reads (Q-score above 30) were considered and a vcf-file containing the variants was constructed. The vcf-file was then processed in-house through Alissa Interpret (Agilent, Santa Clara, CA, USA) to sort out most synonymous variants, variants more common than 1% in the population and variants with reads less frequent than 5%. Alissa Interpret returned an Excel-file listing the variants along with necessary information for further analyses.

3.3 Classification of somatic variants

The NGS data was further analysed to identify true variants in the genetic sequences. The analysis was done with aid from various databases and predictive tools, following a workflow for somatic variant classification, see figure 3.4 [14]

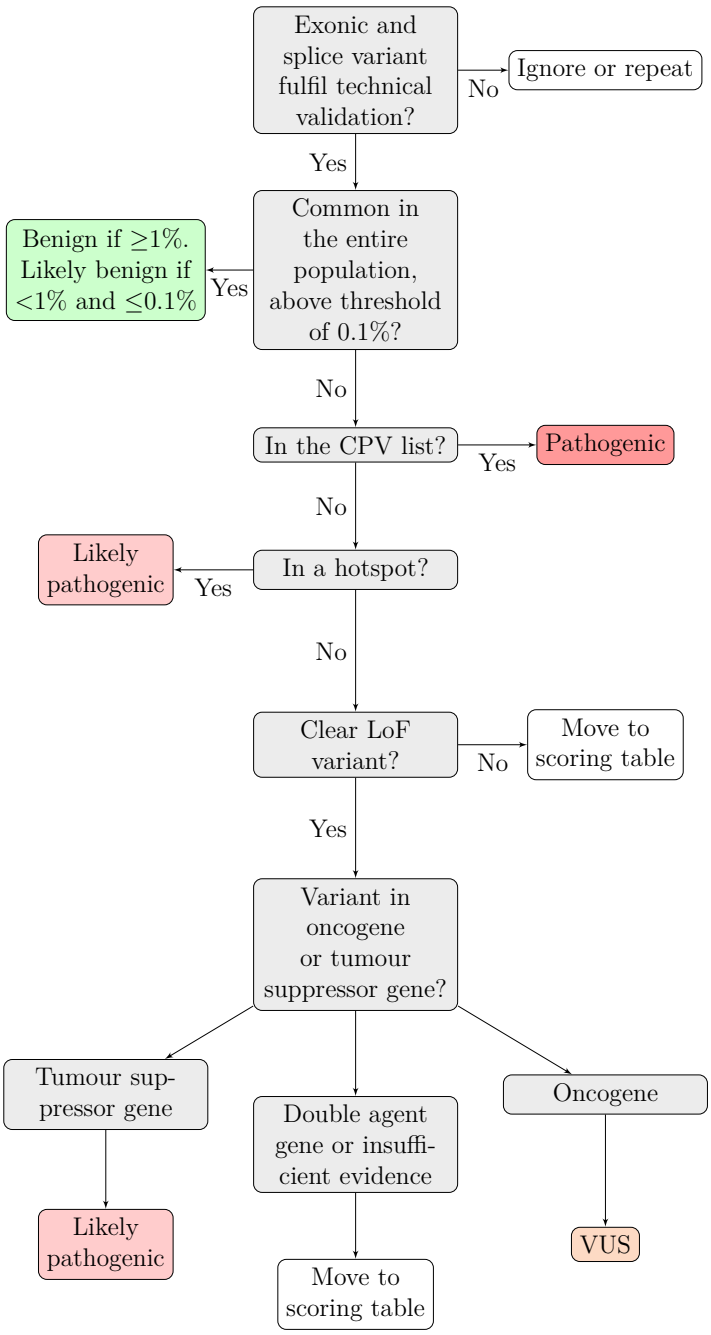


Figure 3.4: Workflow for classification of somatic variants.

The workflow presented in figure 3.4 is essentially as the standardisation workflow described by Froyen et al. [14]. Starting with technical validation, where three criteria needs fulfilment (VAF, in splice site or exon and not technically complex mapping). Usualness of variants are checked in gnomAD. The Consensus Pathogenic

Variant (CPV) list for solid tumours can be found in appendix, A.1. Hotspots are found in the CPV list and cancer-hotspots.com. Non-LoF variants are moved to scoring table. If a variant is in a TS- or oncogene is determined.

A clear LoF variant refer to variants with nonsense, frameshift or splice site variants. Non-LoF variants and LoF in the last exon of genes *BRCA1* and *BRCA2*, exonuclease domain variants in *POLE* and *POLD1* and all *TP53* variants are exceptions and thereby excluded from original the workflow, exceptions are further described in 3.3.1.

In the first step of the workflow, exonic and splice variants were sorted out based on their fulfilment of technical criteria. The variants were manually checked in Integrative Genomics Viewer (IGV; v2.14.1(hg38)) [43] and variants were excluded as artifacts if present in at least one of the criteria presented below. To be classified as a true variant, it cannot be present in any criteria.

1. **VAF <5% in IGV.** Exceptions from criteria 1 were made, based on tumour/healthy tissue ratio, to instead lower the threshold to <3%.
2. **Splice site variant other than ± 2 or ± 1.** Exception from criteria 2 were made for splice variants surrounding *MET* exon 14, *BRCA1* and *BRCA2*; in coherence with recommendations from Froyen et al., where these variants are described as pathogenic or likely pathogenic [14].
3. **Variants with technical complexity.** Technically complex variants were sorted out as method- or panel-specific sequencing errors if present if VAF was >10% while simultaneously having some-what technical complexity of mapped reads. Also, technical complex variants include variants where reads are strand biased, unevenly distributed between strands, in an error prone region (i.e. towards the end of a read or surrounded by many errors), of insufficient read depth or of poor mapping quality (e.g. surrounded by soft-clipped reads).

The second step was to determine if a variant is common in the population according to Genome Aggregation Database (gnomAD; v2.1.1 (GRCh37) [44] and v3.1.2 (GRCh38)[45]). As mentioned, during processing in Alissa, a filter was added to remove variants with frequencies >1%, however when checking variants in gnomAD an even lower threshold of >0.1% was used. The variants with VAF >1% was classified as benign, while the variants with VAF between <1% and ≥0.1% were directly classified as likely benignn unless previously classified as pathogenic or likely pathogenic according to ClinVar [46].

If <0.1% in gnomAD, the variant was checked in a third step for its presence in the Consensus Pathogenic Variant (CPV) list for solid tumours provided by Froyen et al. [14] recreated in appendix/figure A.1. If present in a hotspot of a gene described in the CPV list, the variant was instantly classified as pathogenic. If not in the CPV list, but present in a hotspot according to the database Cancer Hotspots (https://www.cancerhotspots.org/#/home, q-value/FDR <0.1) [47, 48], the variant will be classified as likely pathogenic.

All genes containing variants were then checked in the following four databases: COSMIC's v.96 [33] project Cancer Gene Census (CGC) [49], a list provided by Vogelstein et al. [50], TSGene 2.0 [51] and OncoKB [52]. If a gene is in the mentioned databases, its characteristics in cancer was documented, either as a tumour suppressor gene, an oncogene or a double agent. An overall evaluation of the information

from the databases was used to determine what characterisation each gene exhibits. This information was then used in the next step of the analysis.

Regardless of the gene characteristics, variants with clear non-LoF variants, are handled via a scoring table to determine the variant's probable effect, see table 3.2. For clear LoF variants in a tumour suppressor gene was classified as likely pathogenic; in an oncogene, a clear LoF was described as VUS. Nevertheless, if a LoF variants was present in a gene with double agent characteristics or insufficient evidence, it was analysed in the to the scoring table for non-LoF variants, see table 3.2. Next, each parameter included in the scoring table, 3.2 is thoroughly described, followed by the scoring table below.

1. **# entries (COSMIC).** The number of entries in COSMIC are described for each variant. If the number equal to or exceed 50, 3 p was added, if 50>x>10 then 1.5 p was added. If $\leq$ 10, then no points was added.
2. **Driver gene (CGC).** If a gene is in CGC, it means that it is detected as a driver in cancer [49]. CGC is a literature-based database, where the minimum requirements for inclusion is that at least two separate research groups has shown functional evidence of driving in cancer, in what way (tumour suppressor gene or oncogene) and increased variant frequency in the gene. Divided into two tiers, with tier I has higher confidence than tier II. No difference was made between tier I and II genes. If present in CGC in a somatic cancer type, 1 p was added, otherwise no points.
3. **Predicted driver gene (intOGen or cancer-genes.org).** For genes not present in CGC, the genes were checked in theoretical prediction tools presented in databases intOGen [15] and cancer-genes.org (Memorial Sloan Kettering Cancer Center, New York Cite, USA; q-value/FDR <0.25). This parameter is only considered for genes not found in CGC. If a gene was detected as driver in somatic cancers in either intOGen [15] or cancer-genes.org 1 p was added, otherwise no points.
4. **Interpreted variant effect (CGI).** CancerGenomeInterpreter (CGI) [53, 54] was implemented to predict if a variant was a driver or a passenger. If a variant is interpreted to be a driver it result in an addition of 1 p to the total score. If the variant is interpreted to be a passenger, the score remains unchanged, the same goes for errors or if the interpretation of a variant is blank.
5. **Described in functional studies (VarSome, LitVar, dbSNP, Mastermind, CIViC, DoCM or UniProt).** If a variant has been evaluated in functional studies in VarSome [55], LitVar [56], dbSNP [57], Mastermind Genomic Search Engine [58], Clinical Interpretation of Variants in Cancer (CIViC) [59], Database of Curated Mutations (DoCM) [60] or UniProt [61], it is a strong evidence of its class. If pathogenic, disease causing or connected to drug resistance or response, a value of 1 p was added. If some evidence of being pathogenic was found in studies, e.g. the variant is classified as likely pathogenic, 0.5 p was added to the total score. Contrary, if a variant was described as benign based on functional evidence, -1 p was subtracted. When no functional evidence was found, the total score was unchanged.
6. **ACMG classification (VarSome).** In VarSome a build-in variant classifier was present, based on a point system where the variants were scored. The

scores used for the different classifications were as follows; $\geq$ 10 p = pathogenic, 9–6 p = likely pathogenic, 0–5 p = VUS (whereas 4–5 p are VUS+), -1 – -6 p = likely benign and $\leq$ -7 p = benign. The scoring implemented in this table was classified accordingly: pathogenic or likely pathogenic (9p) = 1 p, likely pathogenic or VUS+ = 0.5 p, VUS = 0 p, likely benign = -0.5 p and benign = -1 p. A few variants had various transcript version whereas the reference base differ between transcript versions, these variants were described as VUS in VarSome's ACMG classification.

7. **Bayesdel addAF algorithm (VarSome).** Bayesdel addAF [62] is a scoring parameter which was added in order to compensate for the strict ACMG criteria in VarSome, affecting mostly novel genes. In some cases the Bayesdel addAF score suggested that the variant was of pathogenicity, nevertheless, it was discarded due to the gene not having any pathogenic variants. In the cases described, the resulting ACMG classification was likely benign. Therefore, when the algorithm suggested moderate to strong pathogenic and was discarded for contradicting the ACMG classification, 0.5 p was added. Practically, the addition result in 0 p, since the ACMG classification of likely benign decreased the score by -0.5 p. When tested on a set of variants, the classifier Bayesdel addAF performed best (when tested on the IDUA gene) [53]. The algorithm combines multiple deleteriousness predictiors with ClinVar [46], and in addition addAF means that population VAFs are included [53].

8. **Conservation of amino acid, Grantham distance (SOPHiA GENETICS).** In case the total score from the table was $\geq$ 1.5 p, the variant was checked in SOPHiA GENETICS (Alamut Visual Plus v.1.4, Saint-Sulpice, Switzerland). More precisely the conservation of the amino acid was checked between orthologues (Ensembl), the Grantham distance, i.e. a measurement of physiochemical difference between amino acids [63], was checked and the predicted effect on splicing, which was a combined percentage from MaxEnt, NNSPLICE and SSF. If highly conserved amino acid and moderate to high Graham distance, 1 p was added to the total score. If the variant was predicted to affect splicing more than 10%, then 0.5 p was added.

Table 3.2: Scoring table for non-LoF somatic variants. The parameters seen in blue are only checked under certain circumstances.

		Score						
#	Parameter	+3	+1.5	+1	+0.5	0	-0.5	-1
1	# entries (COSMIC)	≥50	50>x>10			≤10		
2	Driver gene (CGC)			Yes		No		
3	Predicted driver gene (cancer-genes.org or IntOGen)*			Yes		No		
4	Interpreted mutation effect (CGI)			Yes		Passenger, blank or error		
5	Described in functional studies (VarSome, LitVar, dbSNP, Mastermind, CIViC, DoCM or UniProt).			Pathogenic, disease causing or drug resistant	Likely pathogenic	No data available		Benign
6	ACMG classification (VarSome)			Pathogenic or likely pathogenic (9p)	Likely pathogenic or VUS+	VUS	Likely benign	Benign
7	Bayeldel addAF algorithm (VarSome)				Strong to moderate pathogenic			
8	Conservation of amino acid, Grantham distance (SOPHiA GENETICS)			Highly conserved amino acid and moderate to high Grantham distance	Predicted to affect splicing >10%			

Note, in the workflow described by Froyen et al. [14] a variant must have ≥50 entries in COSMIC to be classified as likely pathogenic. Being in COSMIC, does not mean that the gene is in CGC, i.e. the variant is in cancer samples but might not cancerous. However, the implemented approach have other parameters weighted higher. Further, the classes for classifying was also broadened regarding the VUS classification, where very strong (++), strong (+) and weak (-) VUS:es were determined from the original class. The following classifications are connected to the total scores from the table; likely benign: ≤ -1, VUS-: -0.5 p, VUS: 0–2 p, VUS+: 2.5–3 p, VUS++: ≥ 3.5 p.

3.3.1 Exceptions from workflow

Rather than fitting every variant into the workflow, exceptions were made. The excepted variants were to undergo a procedure in separate pipelines. The variants excluded from the workflow described in figure 3.4 were found in the genes *TP53*, *POLE*, *POLD1*, *BRCA1* and *BRCA2* [14].

TP53 variants were separately checked due to its complexness [14]. Databases specified for use of *TP53* variants were assessed, these were The *TP53* Database ((R20, July 2019): https://tp53.isb-cgc.org) [64] and the predictive database Seshat [65]. The special databases were combined with information from ClinVar [46] and OncoKB [52]. For each variant an individual evaluation was performed wrighing in the available information.

All non-LoF variants for *BRCA1* and *BRCA2* should be treated as exceptions,

as should LoF variants in the last exons [14], exon 23 for *BRCA1* and exon 27 for *BRCA2*. The excluded variants were similarly to *TP53* evaluated in specific databases, the data bases were ARUP (ARUP Laboratories, Salt Lake City, UT, USA), InterVar [66], ClinVar [46], *BRCA* Exchange [67] and LOVD [68].

Variants in *POLE* and *POLD1* were classified as VUS in case the variant was outside of the exonuclease domains [20]. If the variant was in the exonuclease domains, they were to undergo the original classification system.

3.4 Mutational signatures

To prepare the samples for mutational signature analysis, Mutect2 (GATK, v. 4.1.3.0, Broad Institute of MIT and Harvard, Cambridge, MA, USA) were ran on the high-quality tumour and normal blood BAM-files to extract somatic variants. FilterMutectCalls (filter within Mutect2) was applied for filtering with default settings. Further, samples were grouped by clinical response (RECIST 1.1 criteria [5]) and ran in SigProfilerExtractor [37], with and without FFPEsig [39], and in SigProfilerAssignment. The mutational signatures were produced by bioinformatician Katarina Truvé (University of Gothenburg). Each patient's signatures was achieved, and the mutational signatures were analysed. Patients were grouped based on mutational signatures alone and in combination with genetic variants for statistical analyses.

3.5 Genetic landscape and waterfall plot

To visualise the genetic landscape of the genes with variants, waterfall plots were constructed. The plots were constructed in RStudio (v.4.2.2, R Core Team, Vienna, Austria), essentially using the package GenVisR [69]. Separate plots were made for all patients, then split into histological diagnoses (LUAD and LUSC). Clinical data, as sex, histological diagnosis and response to ICB was also provided, together with a list of patients and its belonging mutated genes and variant type. In addition, a TMB plot was constructed with information from each sample. The TMB was calculated by Eurofins Genomics, using only non-synonymous SNVs. When calculating the TMB, some criteria were setup for exclusion; the following variants were excluded: non-coding variants, known germline variants in dbSNP or gnomAD, predicted germline variants by algorithm, known somatic variants in COSMIC or ClinVar [46], variants with read depth <50 reads, variants with VAF <5%, variants in tumour suppressor genes. The code used for the construction of the waterfall plot can be seen in appendix B.2.

3.6 Pathway enrichment analysis

The pathway enrichment analysis was performed in RStudio using the ReactomePA package by Bioconductor [70]. ReactomePA uses Reactome's database of pathways to map the genes onto pathways, as one gene can participate in more than one

pathway. A list of with only gene names in Gene Symbol-format was provided and converted to ENTREZID. The code for the pathway enrichment analysis used can be seen in appendix B.

3.7 ctDNA analyses

Briefly, the analysis of ctDNA was performed as follows, first extraction, then amplifications, detection and lastly data analysis. Two different workflows were used, superRCA [27] and SiMSen-seq [26]. Since extraction and data analysis are equal within the superRCA and SiMSen-seq, those paragraphs are generally described, while the methods are separated where they differ.

3.7.1 Extraction

The extraction of cfDNA, which include ctDNA, was performed with various methods, whereas all were optimised for extracting cfDNA from plasma. The extraction done in-house was performed using magnetic bead-based kits, manually using QIAamp MinElute ccfDNA Mini Kit (Qiagen, Germany) or automated using EZ1&2 ccfDNA Kit (Qiagen, Germany). If extracted in-house then the quantity was measured using Qubit™ dsDNA High Sensitivity Assay Kit (Invitrogen™, Waltham, MA, USA). Extraction of some samples were done at Simsen Diagnostics and Rarity BioScience, using similar techniques.

3.7.2 superRCA

Some samples where only the clinical NGS panel had been used for tumour-sequencing were analysed using the superRCA method [27]. The analysis was performed by Rarity BioScience (Uppsala, Sweden), using their assay for *KRAS* and *EGFR*. The main steps of superRCA can be seen in figure 3.5. For more in-dept method description, see Chen et al, [27].

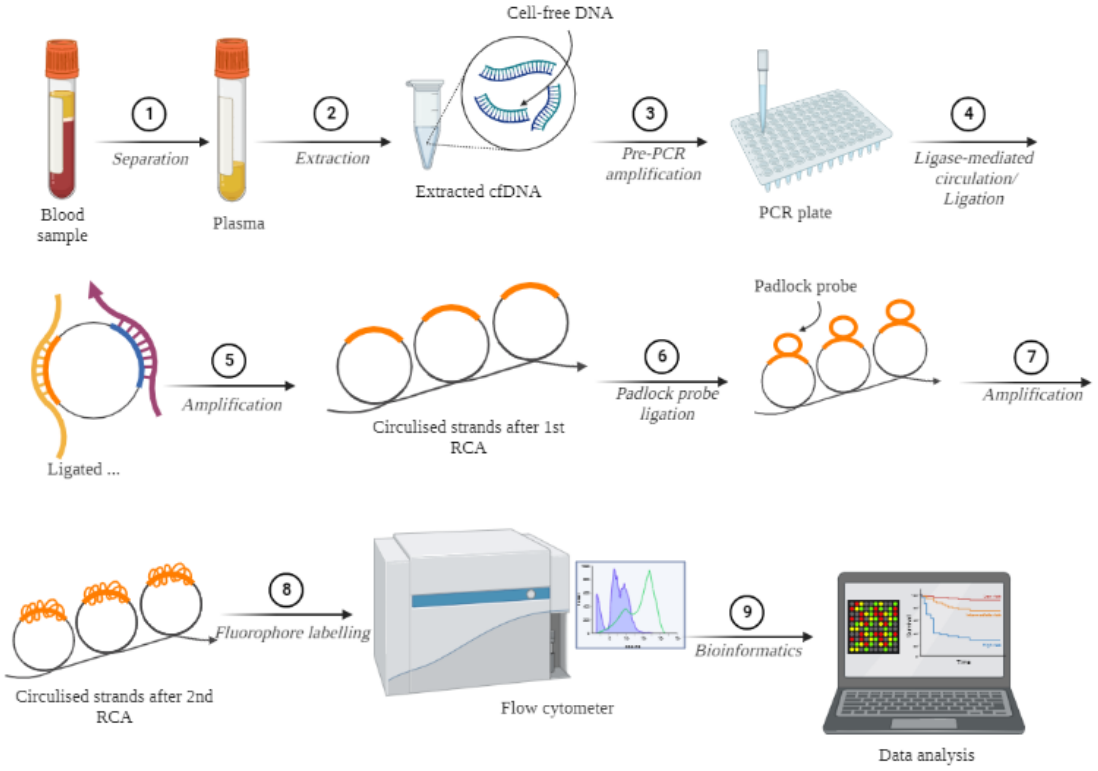


Figure 3.5: Schematic workflow overview of superRCA. Inspired from Chen et al [27], created with BioRender.com.

3.7.3 SiMSen-seq

The main steps of SiMSen-seq workflow can be seen in figure 3.6 [26]. For more detailed method and protocol, see Ståhlberg et al. [71, 72].

SiMSen-seq was ran by Simsen Diagnostics, and also performed in-house. Several different variants were followed in the analysis by Simsen Diagnostics, and assays were designed accordingly. In-house the method was set-up for testing out and following *KRAS* p.G12D variants.

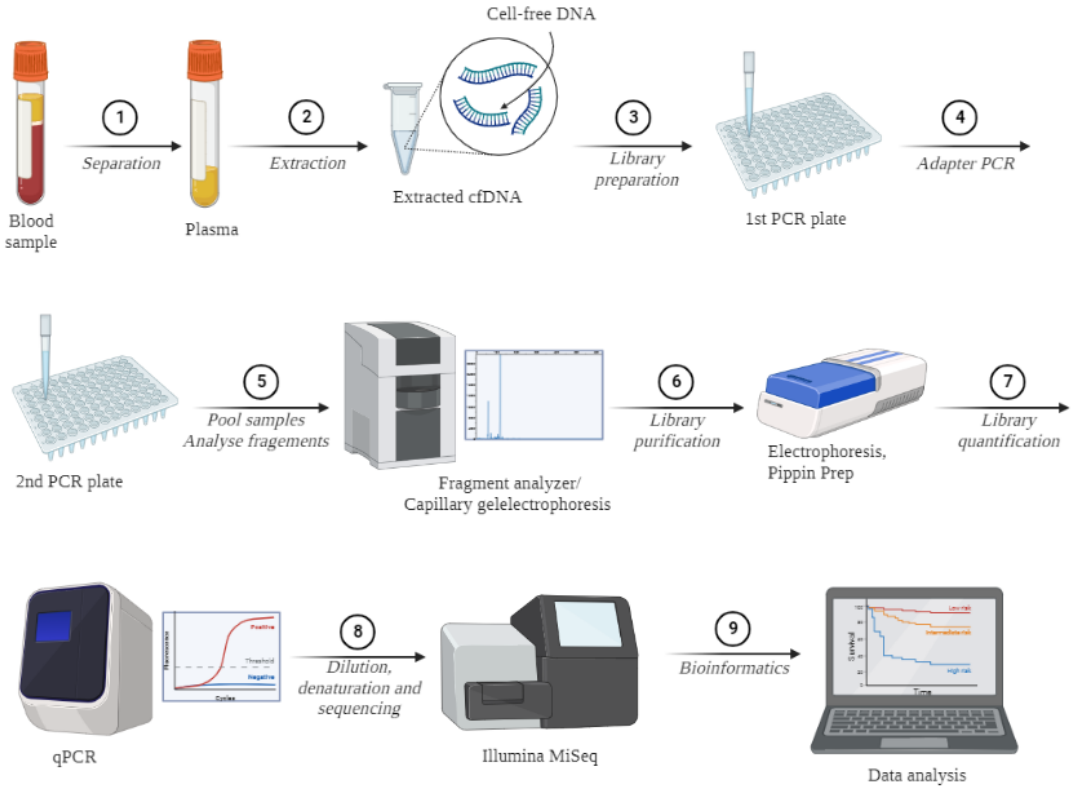


Figure 3.6: Schematic workflow overview of SiMSen-seq. Inspired from Ståhlberg et al. [71, 72] and Andersson et al. [26], created with BioRender.com.

3.7.4 Data analysis

For the data analysis different plots were constructed together with charts over all variants in patients. ctDNA load in percentage was calculated for all patients, variants and time points and grouped based on response.

3.8 Statistical analyses

Statistical analyses were performed in IBM SPSS Statistics (v. 29, Armonk, NY, USA) these were Kaplan-Meier survival analysis for OS, Cox-regression for confounding factor analysis.

4

Results and discussion

Results for the main topics, mutational signatures, variant classification, pathway analysis and ctDNA is statistically presented mostly in terms of OS. Discussion of results are interlaced with the results, then ethical and societal aspects are pointed out, together with delimitations of the thesis.

4.1 Mutational signatures

Firstly, SigProfilerAssignment was used as it can be applied to analyse each sample separately (unpublished). The cohort was seen as quite diverse from earlier studies, and SigProfilerAssignment was thought to better fit with these samples. However, since the samples were panel-sequenced not enough variants were found which resulted in overfitting. The overfitting was identified with support from creator Marcos Díaz Gay, and avoided by using FFPEsig in combination with SigProfilerExtractor. Even though different variations of mapping were tried to enhance quality, it was not possible to reduce overfitting using SigProfilerAssignment.

4.1.1 Signatures and SBS's for all patients

The number of *de novo* signatures that are extracted is based on the optimal number of solutions, which depend on stability in combination with mean sample cosine distance. Aiming towards solutions with low mean sample cosine distance while reassuring high stability [35]. The groups' optimal numbers of solutions was determined in SigProfiler to be 2; SBS96A and SBS96B was then constructed for both responders and non-responders, see figure 4.1. The y-axes represent percentage of SBS's and the x-axes are the 96 mutational types.

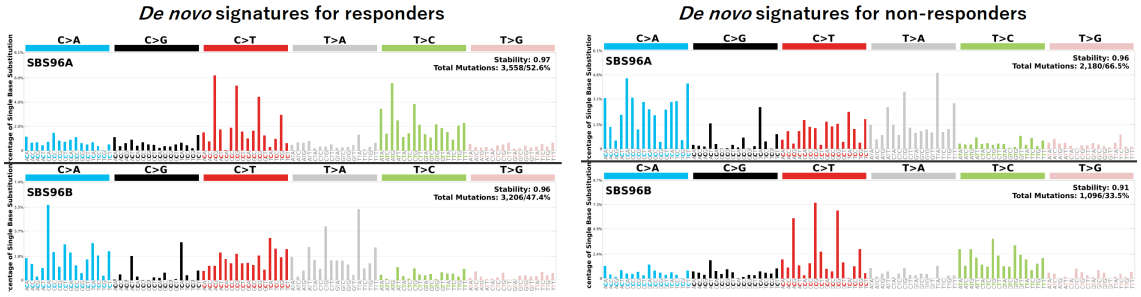


Figure 4.1: *De novo* SBS's for all patients. Responders to the left and non-responders to the right. Proportions of SBS's are represented by a percentage.

The *de novo* signatures were then reconstructed and then decomposed to COSMIC signatures, an examplary reconstruction and decompositon can be seen in figure 4.2. The example is for SBS96A for non-responders and is representative for the rest of the *de novo* signatures as well.

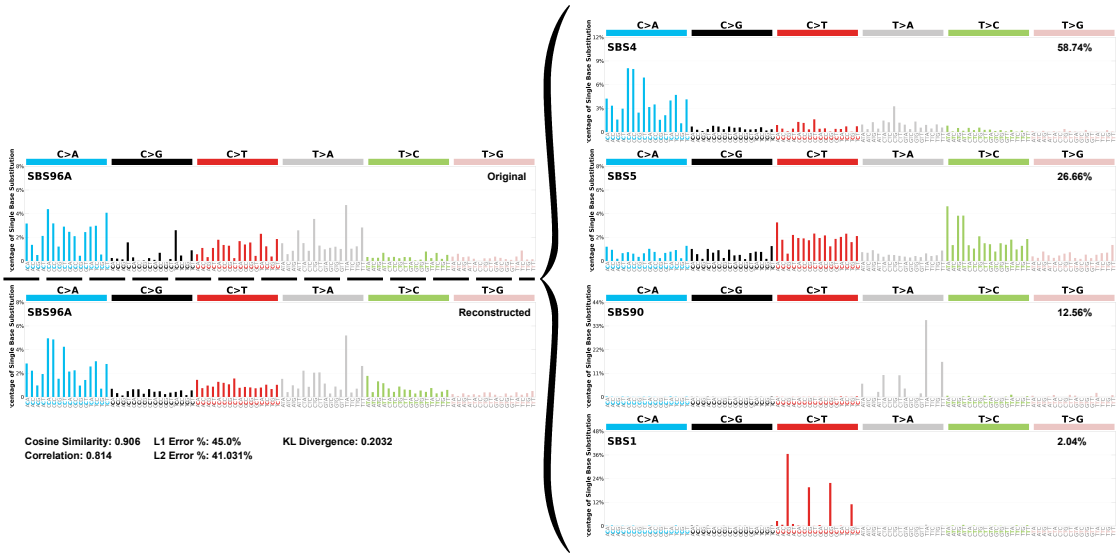


Figure 4.2: *De novo* reconstruction and decomposition. Here, SBS96A for non-responders is shown as example of the process.

For the decomposed samples of the patients, a combined plot for proportionate signatures from activities for all patients was made, see figure 4.3. Notably, the same signatures were seen in both the non-responders and the responders group, the groups were not diverse in terms of mutational signatures. The activity plot in figure 4.3 show proportions of all signatures for all patients in the cohort, without the outlier ID-31. The samples are composed of the signatures SBS1, SBS4 and SBS5 which are found in the majority of patients, also SBS90 and SBS54 were found in both groups but in fewer samples.

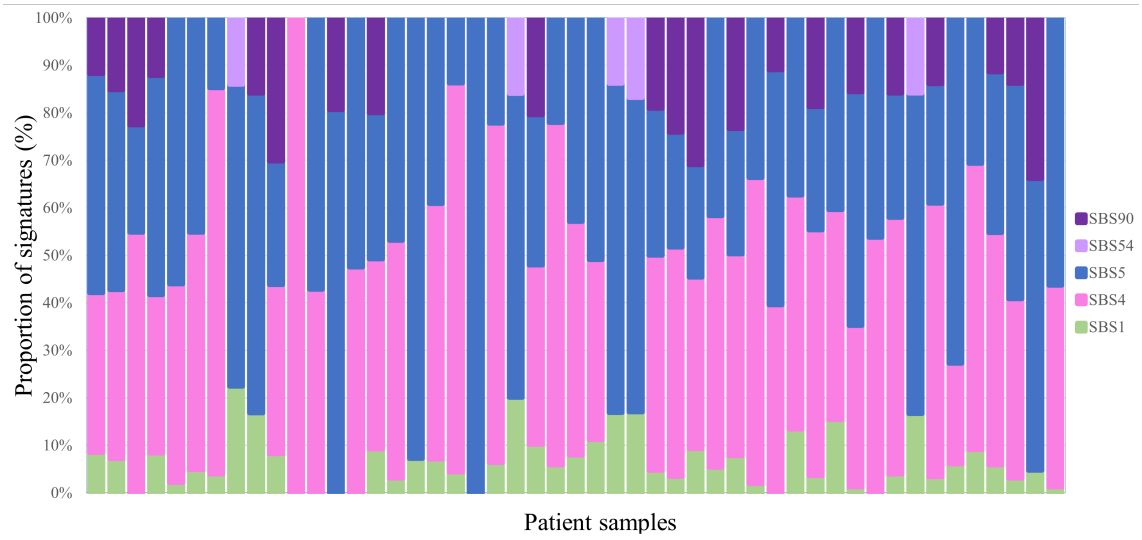


Figure 4.3: Activities for all patients, responders and non-responders. Proportions of signatures are represented by SBS1, SBS4, SBS5, SBS54 and SBS90.

SBS1 is a clock-wise signature which is correlated with age at cancer diagnosis [73]. Suggestively, SBS1 is generated at a rather constant rate from egg to tumour cell and appear due to cell divisions. The signature can therefore be connected to DNA replication substitutions occurring during mitosis. As most patients in the BioLung cohort with NGS data are at a median of 72 years at diagnosis, SBS1 is thereby expected to be seen in a lot of BioLung patients.

SBS4 is one of the signatures related to tobacco smoking [74, 24]. As expected, many patients in the BioLung cohort possess the SBS4 signature, nevertheless no distinct extinction between previous smokers, current smokers and non-smokers have been seen in this study. Non-smokers also possess smoking signatures which was expected, as already when discovering signature 4 it was seen in non-smokers but elevated in smokers [74]. SBS4 in non-smokers most probably originated from passive smoking, apart from possible misreported.

SBS5 is a signature connected to both age and smoking. However, the true aetiology of SBS5 remain unknown [75].

SBS54 is a part of the subgroup of possible sequencing artefact signatures (COSMIC; Mutational Signatures v.3.3). As it is known that samples are of varying quality, no further results will be presented with SBS54. Artefact signatures could have been excluded from analysis, however, that could have led to biased results.

SBS90 is a signature correlated with duocarmycin exposure [76]. Duocarmycin is an anti-cancer drug, and the signature in exposure is connected with duocarmycin itself and its antibody-conjugate drugs. There is no clear connection or similarity between patients in the BioLung cohort possessing SBS90. One hypothesis might be that at least one patient in each group has received treatment for previous other cancers using any of these duocarmycin-conjugates. The fact that it is seen in multiple patients can be caused by spill over to other samples, as the number of activities is quite few. Nonetheless, before investigating SBS90 further, it cannot be removed or ignored. Patient records will be checked by clinicians to try to determine

if any patient has gotten antibody-conjugate drugs, however results were not finished in time to be included here.

DBS's and ID's were not possible to analyse when FFPEsig had been applied. In the future, studying ID's and DBS's would be of interest as well, even CNV's can be analysed with mutational signature analysis. It was determined more important to add FFPEsig than analysing the raw data and receiving potentially artefact variants.

One outlying sample was identified which lead to further investigations of patient ID-31. The signature profile for ID-31 can be seen in figure 4.4.

4.1.2 ID-31 and SBS7a/b

Based on the results from mutational signatures in the group of responding patients, ID-31 stood out compared to the rest of the responders, possessing SBS7a and SBS7b.

Interestingly, SBS7 consist of activities strongly connected to UV light exposure and are most frequently found in skin cancers [73]. Since SBS7 is so rarely found in lung cancer, the signatures had to be further investigated. SBS7 has however been found in two previous NSCLC cohorts, where those samples were re-reviewed by pathologists [24, 77]. In one of the studies the tumours had originated from skin squamous cell carcinomas [77] and melanoma in the other [24] In both studies, the findings lead to exclusion since it originated from skin cancers in both cohorts [24, 77].

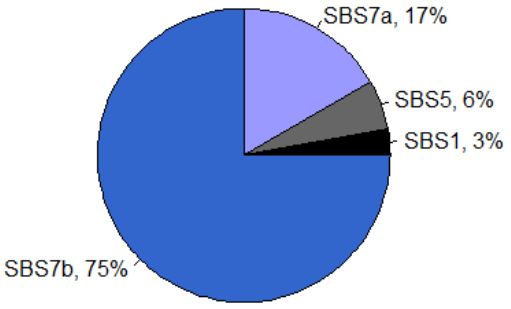


Figure 4.4: Pie chart representing of mutational signatures for outlying patient ID-31.

Moreover, the findings were brought to attention of clinicians and pathologists, whom from patient records found that ID-31 have had surgery for skin squamous cell carcinoma a few years before the lung tumours were found and hence included in the BioLung study. ID-31 was diagnosed with LUSC. Pathologists has not yet been able to confirm if the primary source of the tumour was skin squamous cell carcinoma. From a genomic or genetic point of view, as the SBS7 signatures hardly exist in lung cancer, patient ID-31 must be excluded as an outlier in accordance with previous studies [24, 77]. The mutational signatures leads to hypothesising that the tumour seen in the lungs was a metastasis. Clinically, ongoing investigations are taking place to determine the origin of cancer, which unfortunately could not be completed in time to include in this thesis. The cosine similarity was 0.984 for ID-31 which indicate that the true signatures were found for the sample. Detecting SBS7 in ID-31 highlights the robustness of mutational signature analysis, before the analysis no suspicions were made for ID-31 not being lung cancer.

A few number of the other responders had SBS7a/b signatures too. Nevertheless, after excluding ID-31 none of the remaining samples showed SBS7 signatures. This could be explained by the algorithm of SigProfilerExtractor, since it extract what

signatures are in each group of patients and uses only them to assign to the samples, the algorithm aims to find all possible solutions. Resulting in SBS7a/b was used to explain ID-31 and the signatures spilling over to other samples as well. After exclusion, and re-analysis no SBS7 signatures were found in any of the remaining samples, as hypothesised.

4.2 Classification of somatic variants

In total, 5184 variants were found in the 49 patients after the processed in both bioinformatic pipelines and analysed in Alissa. Out of the 5184, 2700 were only present in the first pipeline, 1112 only in the second and 686 were found in both pipelines, visualised in figure 4.2. From here, only the total number of variants will be referred to, not split into pipelines. 4323 variants were classified as artefact due to not fulfilling at least one of the technical criteria, resulting in 859 true variants classified according to the workflow described in figure 3.4.

.

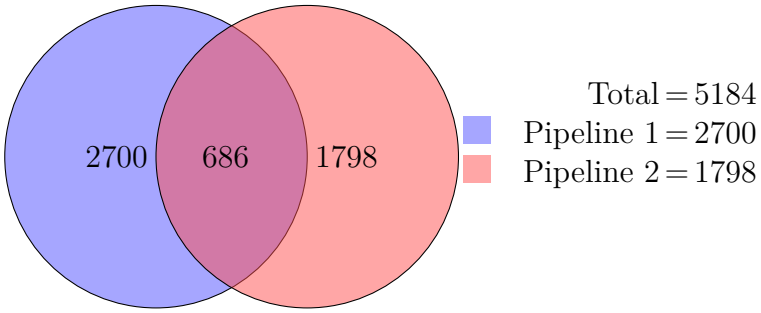


Figure 4.5: Venn diagram of variants from the two different bioinformatical pipelines.

The number of true variants within the samples varied from 0 to 70 per sample. In total the classification was as follows, 40 (4.7%) pathogenic, 96 (11.2%) likely pathogenic, 34 (4.0%) VUS++, 75 (8.7%) VUS+, 501 (58.3%) VUS, 75 (8.7%) VUS-, 35 (4.1%) likely benign and 3 (0.3%) benign, see table 4.2. No true variants were found in ID-10, ID-32 nor ID-37; therefore, these samples together with ID-9 were exceptions in terms of lowered VAF cutoff to 3% in the technical validation. The decision was based on not finding any variants and also low tumour/healthy tissue ratio of 2%, 5%, 5% and 20% corresponding to ID-9, ID-10, ID-32 and ID-37, respectively.

Table 4.1: Result from classification of somatic variants. The true variants have been classified as pathogenic, likely pathogenic, VUS++, VUS+, VUS, VUS-, likely benign or benign. Both number of variants and the percentage it represents are presented.

Pathogenic	Likely pathogenic	VUS++	VUS+	VUS	VUS-	Likely benign	Benign
40 (4.7%)	96 (11.2%)	34 (4.0%)	75 (8.7%)	501 (58.3%)	75 (8.7%)	35 (4.1%)	3 (0.3%)

Variants classified as pathogenic from the CPV list were found in genes *BRCA1* (LoF frameshift variant), *BRCA2* (LoF frameshift variant), *IDH1* (p.R132C), *KRAS*

(p.G12C/A/V, p.G13C and p.Q61L/H), *MET* (exon 14 skipping). In addition, 17 LoF or non-LoF variants in *TP53* were classified as pathogenic.

The majority of variants are VUS. However, we have managed to extract information to further refine the VUS category. Here, we manage to differentiate between VUS with more pathogenic influences and VUS that appears to be more benign. Upon introduction of VUS++, VUS+ and VUS-, the original VUS category was thereby reduced from 685 to 501 variants. A classification like this with new VUS categories has to our knowledge in this context not been reported before.

There are very few benign variants overall because these are already sorted out with filters in Alissa interpret. Even though, not all variants are driving and as involved in cancer, the variants cannot yet be classified as benign without functional evidence. Both versions of gnomAD were used since v2.1.1 with GRCh37 included a lot more data compared to the later version having fewer sequenced genomes available. Even though information differs between genomes, it is reasonable to use the available information from both.

Originally, it was planned that LoF variants with insufficient evidence of being a tumour suppressor gene or an oncogene or non-LoF variants with evidence of being a double agent gene, also should have been excluded from the workflow in figure 3.4. Froyen et al. [14] suggested to use the characteristic that the gene exhibited in the specific tumour type investigated. However, this was not practically doable due to for example lack of information on how double agents act in different cancer types [17]. A decision was made to treat these variants in the same scoring table as non-LoF variants, table 3.2. However, in the future when there is available databases, the workflow for variants mentioned should be modified to fit double agent genes according to functionality in tumour types.

VarSome's ACMG classification can potentially be biased by the fact that predictive data is discarded, when not in conclusion with other combined evidence, this can cause issues when looking at potential harmful variants in an unknown gene. Clinically, this is the way to go, however, not when trying to find novel genes in research. Resulting in a circular argument, for example, if the gene *LRP1B* has no pathogenic variants, the evidence suggesting this variant's pathogenicity is discarded due it not having any pathogenic variant. To try and solve this issue, the predictive meta score from BayesDel addAF algoritm was added, in cases where it was discarded due to it not being in conclusion with the other evidence.

4.3 Genetic landscape

The following section gives an overview of BioLung cohort and the most common variants throughout the NGS subcohort. The genetic landscape of mutated genes for the NSCLC cohort is presented in figure 4.6. [21, 78].

Figure 4.6 shows the 8 most mutated genes through out the NGS cohort. The genes are ranked from the most mutated, as followed *TP53*, *CSMD3*, *LRP1B*, *KRAS*, *FAT3*, *SPTA1*, *TRRAP*, *ERBB4*. The number of patients seen in the waterfall plot are 49.

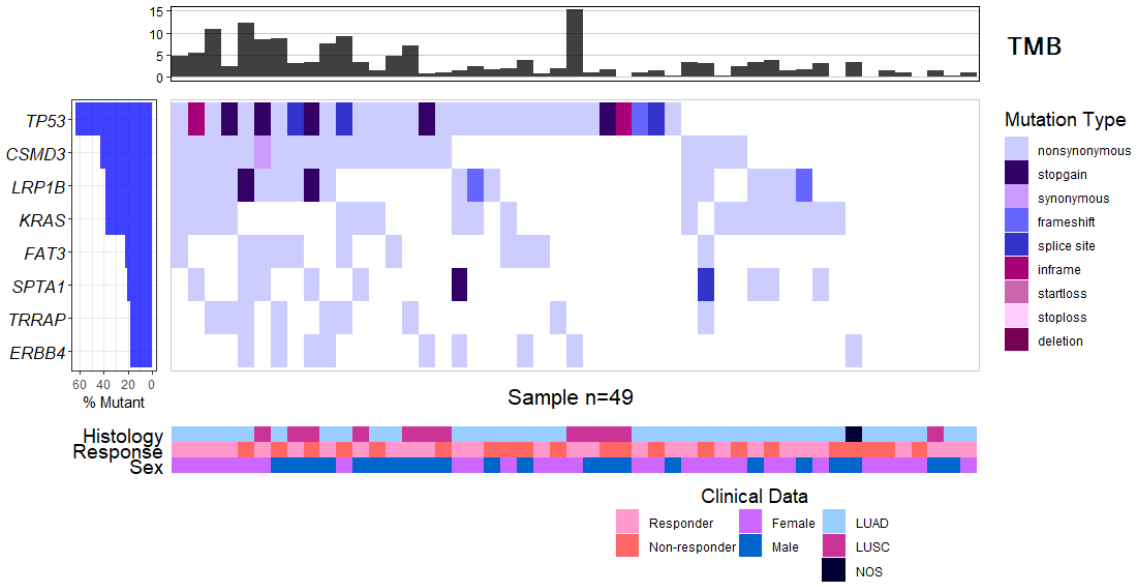


Figure 4.6: Waterfall plot showing the most frequently mutated genes in the NSCLC patients. The top plot shows the TMB for each sample and the bottom show clinical data.

Based on histodiagnosis, waterfall plots were also constructed for LUSC and LUAD patients separately, see figures 4.7 and 4.8. Apart from *TP53* and *CSMD3*, LUSC patients possess variants in other genes compared to LUAD. Recent studies has suggested that LUAD and LUSC are genetically diverse and should be separated from each other [79]. LUAD and LUSC are treated as different diagnoses, nevertheless, in terms of treatments, no differences are made. Moreover, the results shown in figures 4.7 and 4.8 identifies differences in the genetic landscape of the two diagnoses.However, in the BioLung cohort the vast majority are LUAD and the number of patients with LUSC is rather limited. For future considerations, histodiagnosis could also be a factor affecting treatment decisions.

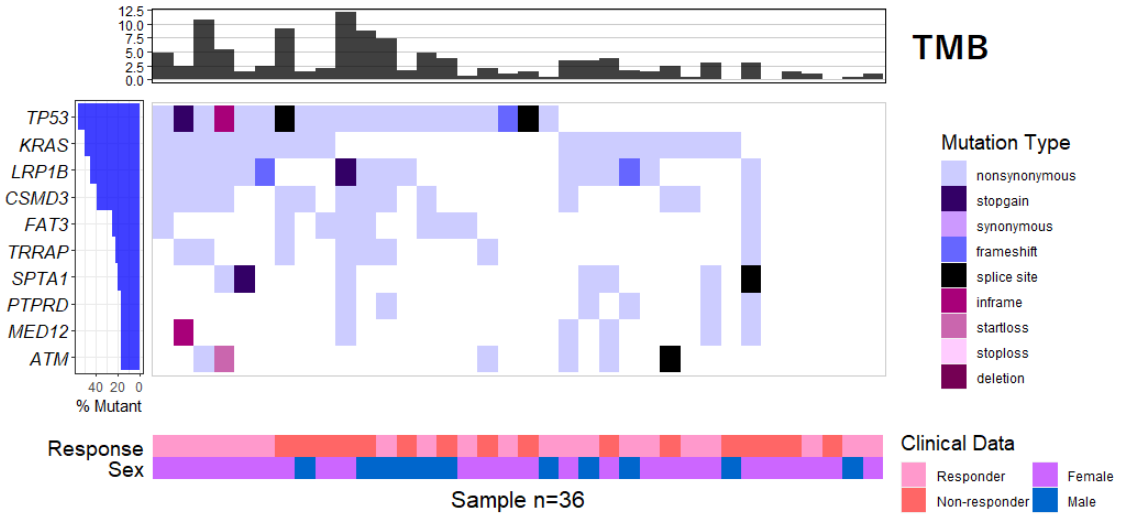


Figure 4.7: Waterfall plot showing the most frequently mutated genes in the LUAD patients. The top plot shows the TMB for each sample and the bottom show clinical data.

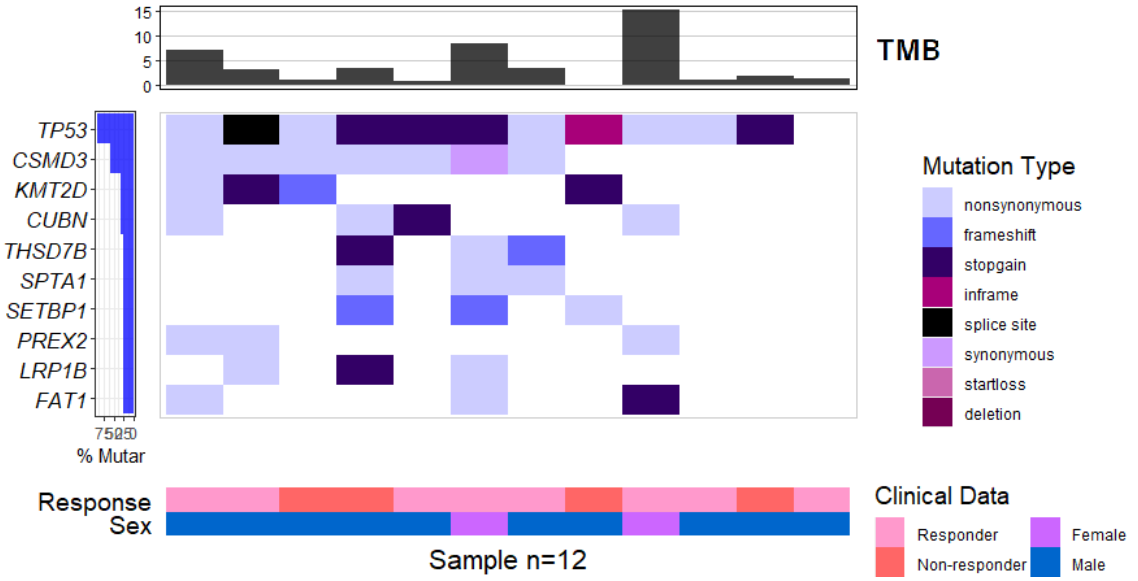


Figure 4.8: Waterfall plot showing the most frequently mutated genes in the LUSC patients. The top plot shows the TMB for each sample and the bottom show clinical data.

Statistically in an NSCLC cohort, more variants should be expected in driver genes *ALK*, *EGFR*, *BRAF*, *MET*, *RET* and *ROS1* [78]. In this study, we will not see the usual distribution of NSCLC drivers. As there are well-functioning treatments for many of the known alterations, those patients will not get ICB. What is present here are the patients with other drivers, or unknown drivers.

4.3.1 Co-occuring variants

Patients with co-occuring variants in various genes were also generated. The variants presented in table 4.2 represent the eight genes with most variants. A series of programmed Excel-sheet calculations was applied to minimise human errors when combining big sets of data.

Most co-occuring variants was seen between *TP53* and *CSMD3*, but also combinations of *LRP1B*, *TP53*, *CSMD3*, *FAT3* was observed among several patients. Co-variants has been suggested as possible biomarkers and can affect response to treatment, but also be a way for investigation of mutational exclusivity. For example, in this cohort, only one co-variant is seen in *ERBB4* and *KRAS*, possibly suggesting mutual exclusivity between them. More cases of *KRAS/ERBB4* would need to be studied and the pathways they are part of. Also, functional studies in model organisms is needed to confirm mutual exclusivity.

Table 4.2: Co-variants within the most mutated genes in the cohort.

	TP53	*CSMD3*	*KRAS*	*LRP1B*	*FAT3*	*SPTA1*	*ERBB4*	*TRRAP*
TP53	31	17	10	13	10	6	8	8
CSMD3	17	21	10	12	8	6	5	8
KRAS	10	10	19	11	3	6	1	3
LRP1B	13	12	11	19	6	8	5	6
FAT3	10	8	3	6	11	3	3	4
SPTA1	6	6	6	8	3	10	3	2
ERBB4	8	5	1	5	3	3	9	3
TRRAP	8	8	3	6	4	2	3	9

4.4 ctDNA

An overview of ctDNA found in tumour and also in blood plasma is shown in figure 4.9. In figure 4.9 genes with variants tested in ctDNA, divided into clinical responders to the right and clinical non-responders to the left, is shown. Artefacts discovered post-analysis when re-reviewing tumour variants are coloured in blue, this also in accordance with no ctDNA alteration found for the artefact variants presented. Variants detected in both tumour and plasma are red and those tested in ctDNA but only found in the tumour is presented with yellow boxes. Detection of ctDNA is determined if present in at least one time point within the first five cycles (baseline, B, C, D and E). The variant must be present in >1 MM to be considered

found in ctDNA, as 1 MM is too uncertain to confidently be determined as true. The white box indicate that the variants were not tested in this personalised ctDNA analysis. Two patients, ID-53 and ID-59, were excluded from the ctDNA analysis due to lack of follow-up data to determine response.

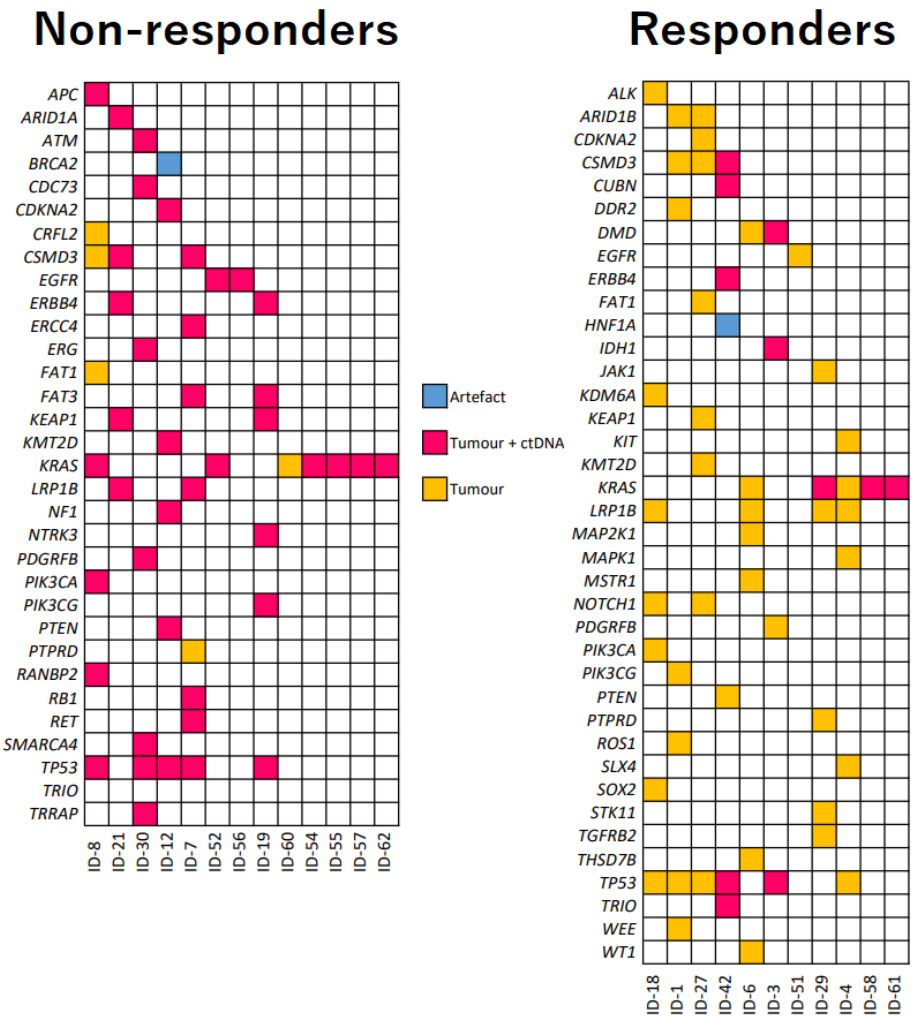


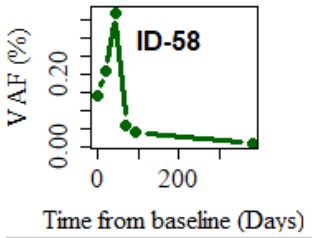
Figure 4.9: Genes tested in ctDNA in responders (right) and non-responders (left). Artefacts are shown in blue, variants found in both tumour and blood plasma in red and variants found in only the tumour are in yellow.

As presented in figure 4.9, there is more yellow in responders, i.e. only detected in tumour, while the non-responding group contain more red boxes, i.e detected in both tumour and ctDNA. The results indicate that detection of ctDNA is correlated with worse response to treatment.

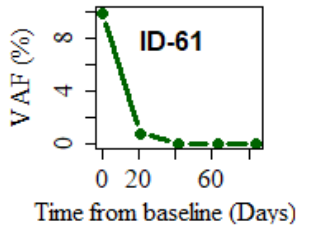
Non detectable (ND) levels of ctDNA during all time points was observed in seven patients. The patients with ND levels were ID-1, ID-4, ID-6, ID-18 and ID-27 analysed with SiMSen-seq all of which were responders. ID-60 and ID-51 also had ND levels and was analysed by Rarity, where only ID-51 was classified as a clinical responder. For ID-60 and ID-56 it was harder to determine response since these patients are stage III, receiving durvalumab (PD-L1) to reduce risk of relapse

after chemo- and radiation therapy. ID-60 had SD for 6 months and had PD at 9 months, to properly assess ctDNA for these patients, more samples or variants to follow would have been needed, and especially a follow-up at every three months or more often to be able to detect PD earlier. ID-56 also had SD when the samples were taken and had only 2 MM at baseline, then clearance.

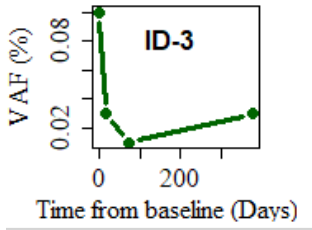
In figure 4.10, plots for responding patients with at least 1 detectable MM in at least one time point can be seen. The patients presented are ID-58, ID-61, ID-3, ID-29 and ID-42, with the corresponding MM/mL also is presented in tables, all of which has ND levels in at least one time point. VAF's shown are for the ctDNA load (%) meaning that the sum of the MM's were divided by the sum of coverage. ID-51 and ID-57 only had two samples each and are therefore not presented in figures 4.11 nor 4.10



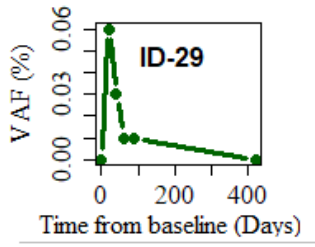
Time point	MM/mL
A	7
B	27
C	1
D	1
E	0
1 y	4



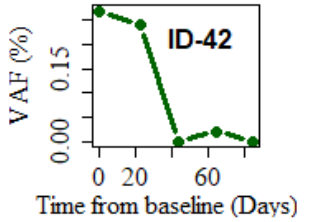
Time point	MM/mL
A	238
B	9
C	0
D	0
E	0



Time point	MM/mL
A	14
B	7
D	0
1y	2



Time point	MM/mL
A	0
B	9
C	18
D	0
E	1
1y	0



Time point	MM/mL
A	48
B	31
C	0
D	1
E	0

Figure 4.10: Responding patients with detectable values with corresponding VAF's in the plot and MM/mL in the table next to it.

Non-responding patients are presented in figure 4.11, ctDNA load (%) is plotted against time from inclusion in days. All non-responders except ID-7 have ctDNA in detectable levels already at baseline. And generally, 50-fold higher VAF's were detected in non-responders compared to responders already at baseline for this cohort.

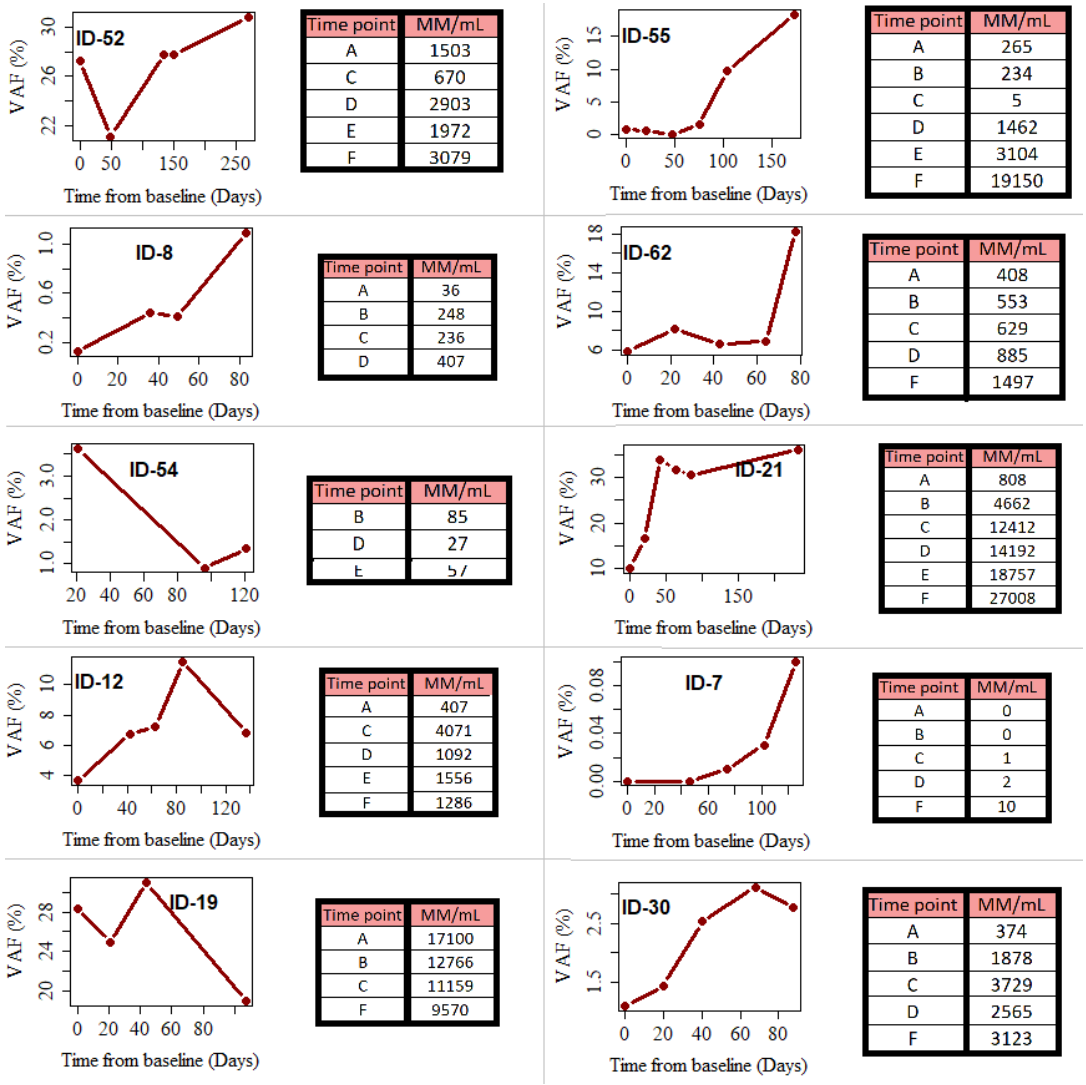


Figure 4.11: Non-responding patients with with corresponding VAF's in the plot and MM/mL in the table to its right.

Correlations between tumour size and ctDNA levels in plasma has been found in previous studies [80]. For example ID-7 would be a patient interesting to study further to understand the levels detected. Investigating the tumour proliferation rate and also the tumour size for ID-7 would might explain the ctDNA MM/mL and VAF. PD was detected clinically at 3 months for ID-7, and as the baseline MM/ml and VAF was ND suggests ID-7 might have a resistance variant to ICB or had developed one.

Showing only SiMSen-seq patients, the variations in VAF's for responders and non-responders respectively can be seen in figure 4.12. Also, notice the difference in x-axes here, responding patients all have below 0.3 % while responders have much more. For all responders a decrease is seen while increase in ctDNA is seen for non-responders.

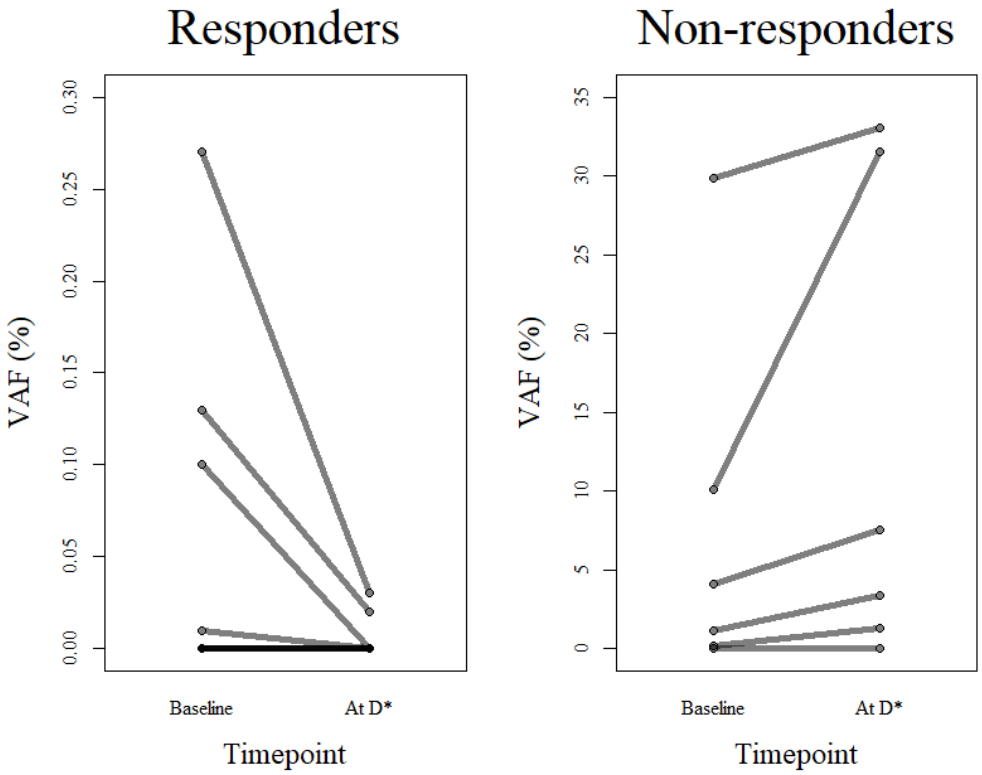


Figure 4.12: Responding and non-responding patients from baseline to D *or C if D is missing.

For both responding and non-responding patients, the response to treatment seems to be able to be detected early and followed during treatment in blood. As mentioned, some patients had different patterns, ID-7 for instance had ND in A, B and C. Usually with biomarkers it does not work for all patients and some cases will be more difficult and tricky to interpret. To follow-up this research it would be interesting to take samples every month when patients have appointments to see in a more long-term perspective how ctDNA can be followed to detect PD earlier. Most patients included here are early non-responders, i.e. they have PD at 3 months. It would be interesting to follow more non-responders with initial response and later PD to see trends there as well.

For patients with ND ctDNA levels, we cannot be sure that the variants were true throughout all treatment points. However, when following multiple tumour-specific variants the risks of all disappearing with tumour evolution or being artefacts is low. Especially when following driving variants, they should not disappear due to clonal evolution. To ensure that the tumour variants are still present, we would also like to have another tissue biopsy some months into treatment to ensure the tumour still has the variants followed. Also, the importance of analysing both MM/mL and VAF is shown in figures 4.11 and 4.10. Variations in VAF's for equal concentrations of MM/mL is due to the amount of cfDNA is higher in some patients resulting in lower and perhaps underestimated VAF. Therefore, both VAF and ctDNA should be reported.

4.4.1 Special cases

ID-52 in figure 4.9 possesses variants in *KRAS* and *EGFR* and the variants were found in both tumour and ctDNA; even though variants within these genes are mutually exclusive. Nevertheless, as presented by Lee et al. [23], a few NSCLC patients possess both *EGFR* and *KRAS* variants. It should not be beneficially for the tumour cell as the pathway already is activated by one of the variants. For ID-52, it is of interest to determine which of the variants are more driving in tumourigenesis to discover the best treatment regime. The biopsy was taken from the primary tumour, where both variants were found.

In figure 4.11 the ctDNA load was shown, but in figure 4.13 the contribution of each variant is seen more clearly. In ID-52, *EGFR* is found in higher VAF and MM/mL compared to *KRAS*, which could be a slight indication of *EGFR* being more driving compared to *KRAS*. If present in the same cells, *KRAS* is downstream of *EGFR* suggesting abnormal activation of *KRAS* [81]. ID-52 is a stage IV patient having spread disease, leading to another explanation for difference in VAF's being that *EGFR* is present in more metastases while *KRAS* is present in only a subclonal population of the cells. Differences in VAF and MM/mL can also be due to tumour heterogeneity. It would need further investigations to determine if variants are found in the same tumour cells. No conclusion can be drawn from only one patient.

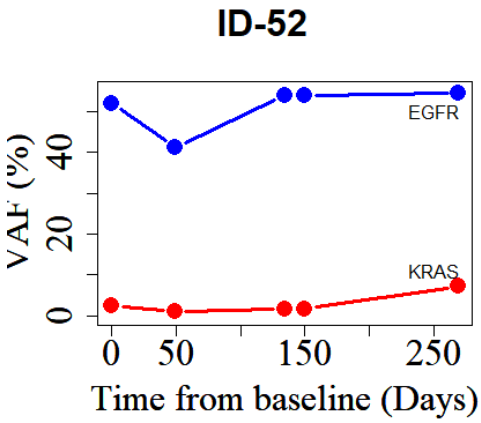


Figure 4.13: VAF curve for ID-52 for ctDNA for *KRAS* and *EGFR* separately. *KRAS* is in red and *EGFR* is in blue.

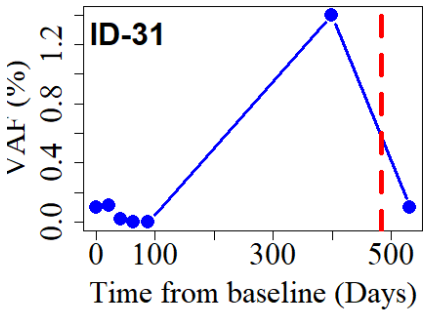


Figure 4.14: VAF curve for ID-31 ctDNA analyses. The red line show the time for clinical PD.

ID-31 is as mentioned excluded due to mutational signatures, but shown as an example of ability to detecting progress earlier than in clinical settings, see figure 4.14. As an increase in ctDNA indicating lesser response to treatment is seen after 1 year but clinical progress (PD) is not detected until three months later, at the red dashed line. The time point with the highest VAF is in fact not the progress sample F, the last sample in figure 4.14. When clinical progress was seen another treatment was started, so the results in F shows decreased levels. The patient responded well (PR) to the new treatment, as confirmed by lowered ctDNA levels. Showing how ctDNA can potentially be used to monitor treatment and detect progress earlier than with clinical measures.

4.4.2 In-house method comparison

For two patients, ID-60 and ID-62 with samples from five available time points analysed with superRCA, and also with SiMSen-seq in-house. The patients both had *KRAS* p.G12D variants. In figure 4.15, for ID-60 samples A, B, C, D and E is presented while ID-62 samples are A, B, C, D and F, as clinical progress was seen before E.

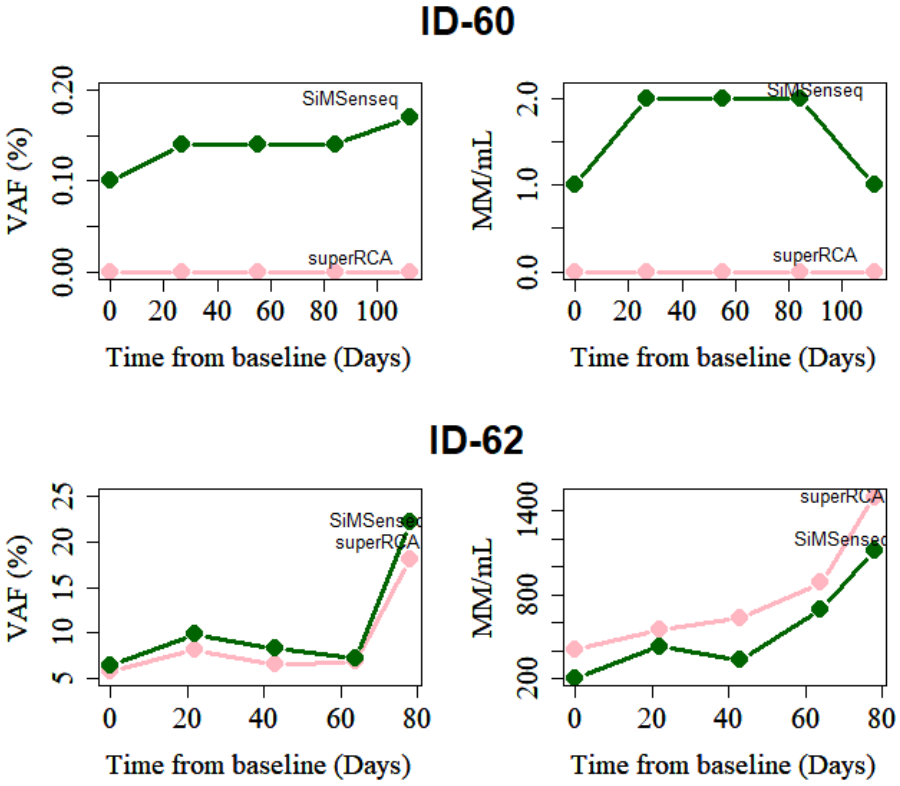


Figure 4.15: ID-60 is shown at the top while ID-62 is shown below.

Interestingly, plasma volumes for ID-62-B and ID-62-D were only 0.5 mL, still a sufficient amount of cfDNA was extracted and gave comparable results. This was an important discovery, since it shows such small amounts still can give similar results. Small differences in VAF's and MM/mL are expected to see, as aliquots from same tube of blood results in different cfDNA amounts. Through amplification steps, small errors between aliquots also grow larger. More samples ran with both methods are needed to properly assess the differences.

4.5 Pathway analysis

A pathway analysis was performed using ReactomePA, some pathways were significantly more mutated than others. However, the pathway analysis was aimed to use for discovering genes with variants connected by being in the same signalling pathways. Nevertheless, ReactomePA was not designed for genomics data, and required

RNA-seq data, i.e transcriptomics. Despite the unsatisfactory results from the pathway analysis this time, further, we would have liked to develop this analysis by doing transcriptomics data also. The analysis could have been done manually for all genes, however that would be time-consuming and another programming approach using other pathway databases focused on signalling pathways would be preferable.

Using expression data could differentiate more between pathways and the genes in them. The differentially expressed genes is determined by a fold-change and p-value, requiring transcriptomics. The results we achieved did not take into consideration the number of mutated genes, i.e. if *TP53* had 35 variants, it was only counted as one. Also, if the groups of responders and non-responders that were more genetically diverse, clearer results could also have been observed when splitting groups.

4.6 Statistical analysis of biomarkers for ICB response

In the following section statistical survival analyses are presented, to achieve statistical significance, a p-value <0.05 was required.

4.6.1 Established biomarkers

There are two established predictive biomarkers for ICB response in NSCLC, TMB and PD-L1 expression. TMB was tested and results for OS is presented in figure 4.16 (logrank p-value: 0.698). The cut-off for high TMB was $\geq$10 mut/Mb, since that is the FDA approved threshold.

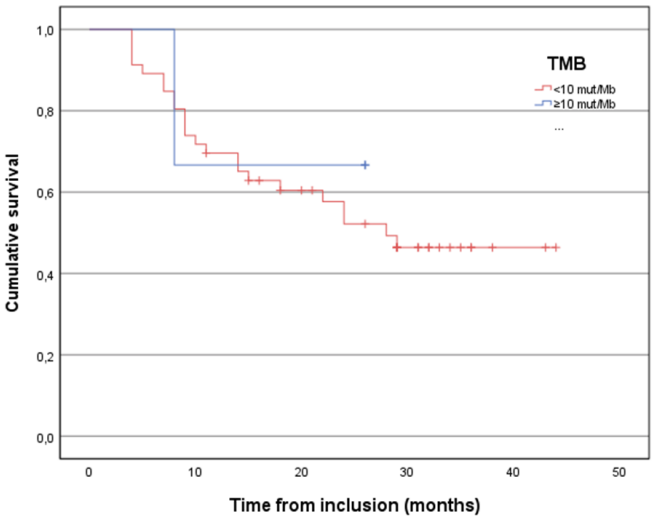


Figure 4.16: KM survival curve for TMB with cut-off $\geq$10 mut/Mb.

Figure 4.17 show KM OS curves for PD-L1 expression calculated in two ways. In figure 4.17 A the expression cut-off is $\geq$50% (logrank p-value: 0.288), in figure 4.17 B, <1%, 1-49% and $\geq$50% as cut-offs (logrank p-value: 0.372). No significant

difference can be seen in the KM survival curves for PD-L1 expression, for neither of the calculations.

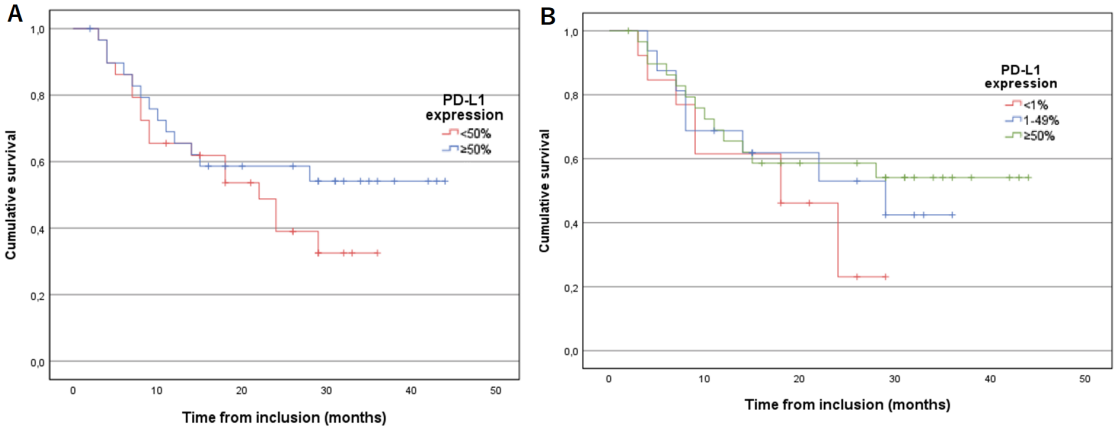


Figure 4.17: A: KM survival curve for PD-L1 expression for cut-off of ≥50%. B: KM survival curve for PD-L1 expression for cut-off of <1%, 1-49% and ≥50%.

TMB and PD-L1 expressions are far from statistically significant in prediction of ICB response in terms of OS. As PD-L1 expression is affected by surrounding factors such as previous treatments and TMB calculations has no standardisation, there is an urge for other biomarkers. In this cohort, the biomarkers approved for use were unable to predict the response to ICB in terms of OS. The results presented in figures 4.16 and 4.17 emphasise the bias in established biomarkers.

Further, the OS was investigated. The sections below shows the most promising and significant data from the analyses.

4.6.2 Individual genes

For individual genes, *TP53* (logrank p-value: 0.316), *CSMD3* (logrank p-value: 0.920), *KRAS* (logrank p-value: 0.041, Cox regression p-value: not significant (ns)), *LRP1B* (logrank p-value: 0.041, Cox regression p-value: ns), *FAT3* (logrank p-value: 0.506), *SPTA1* (logrank test p-value: 0.843), *TRRAP* (logrank p-value: 0.865) and *ERBB4* (logrank p-value: 0.575) were investigated for OS. Only *KRAS* and *LRP1B* is presented in figures 4.22 and 4.18 as the other individual genes did not gave significant difference between wild-type and mutated gene. Variants in other genes appeared in too few of the patients to be properly assessed. Even when dividing *LRP1B* into its classes patients with variants classified as VUS were connected to longer OS (logrank p-value: 0.043, Cox regression p-value: ns).

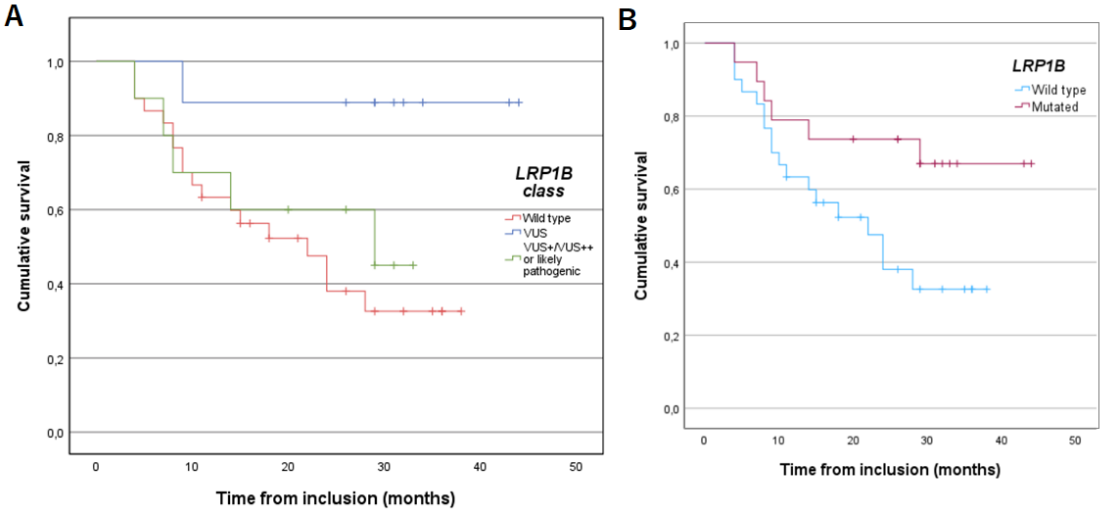


Figure 4.18: A: KM survival curve for *LRP1B*, also divided into classification, VUS or of pathogenic influence (denoted by LP/VUS++/VUS+). B: KM survival curve for *LRP1B*

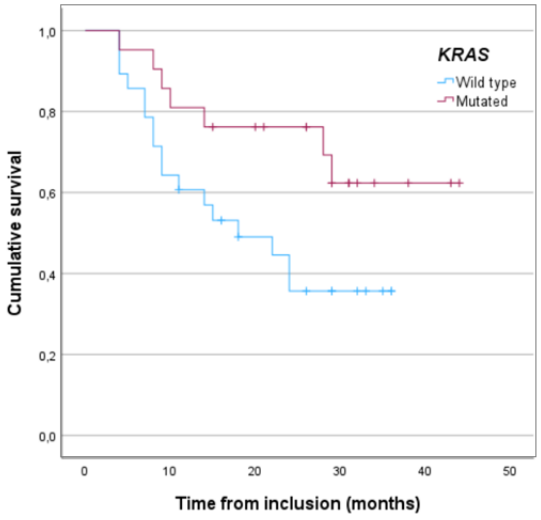


Figure 4.19: KM survival curve for *KRAS*.

Most studies find *KRAS* correlated with longer OS, however there are contradicting evidence on the efficiency of *KRAS* as a predictive biomarkers [82]. Different studies have came to different conclusions. In this study, *KRAS* seems to be connected to longer OS with immunotherapy treatment.

4.6.3 Co-occuring variants and groups of variants

KRAS/STK11 and *KEAP1* represent genes where variants within them lead to worse outcome with ICB treatment in comparison with no treatment [7]. The number of patients possessing *KEAP1* or *KRAS/STK11* in this cohort was insufficient for OS analysis.

Co-occurring variants in genes *LRP1B* and *KRAS* correlated to prolonged survival compared to wild-type of both genes (logrank p-value: 0.003, Cox regression p-value: 0.025), see figure 4.20 B. *KRAS/LRP1B* is also correlated with longer PFS, hence better response to treatment (figure will not by shown here, unpublished). Our data also suggests double mutants being better responders and longer survival than only *KRAS* and *LRP1B* individually see figure 4.20 A.

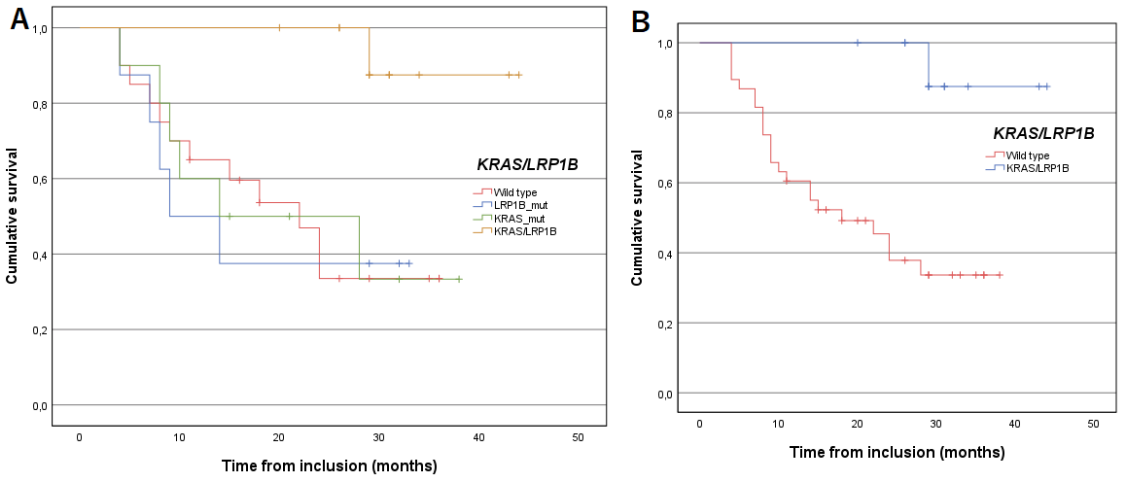


Figure 4.20: A: KM survival curve for *KRAS/LRP1B*, also showing only *LRP1B* variants and *KRAS* variants individually. B: KM survival curve *KRAS/LRP1B*.

The significant OS seen in *KRAS* and *LRP1B* genes individually is probably due to the subgroup of patients having the co-occurring variants are all responding well and have long OS. Purely speculative, *LRP1B* might enhance immunotherapy and tumour infiltration by being involved in cellular processes involving exocytosis. Furthermore, *LRP1B* would be interesting to test more, alone and in combination with *KRAS* and correlate to immune-related factors, conduct functional experiments and also sequence more *KRAS* patients with a larger NGS panel to find more patients with the combination of *LRP1B*.

As previously mentioned, different studies came to different conclusion regarding *KRAS* individually [82]. The difference might have been due to *KRAS* co-occurrence with *LRP1B* variants not investigated in those cohorts. It has might been differing between wild-type or mutated *LRP1B*, with wild-type *LRP1B* and mutated KRAS indicate no difference from wild-type of both *KRAS* and *LRP1B*; suggesting the co-occuring variant to drive the response to ICB and longer OS.

Large variations in actual numbers of variants were seen when comparing the patients to each other, apart from pathogenic variants, which were either zero, one or two. For other classes and combinations of classes, different cut-offs were tried but without successfully addressing issues the intra-patient variations. Instead of using a number as cut-off, percentage of variants within classes were calculated and used. As seen in figure 4.21 A, there is a clear trend, suggesting that having less than 25% of the variants within the categories pathogenic, likely pathogenic, VUS++ has shorter survival, however not significantly correlated with survival (logrank p-value:

0.12 (ns)). Combining classes propose as a refinement of TMB. Nevertheless, refined TMB has lower p-value compared to TMB and PD-L1 expression.

For variants classified as pathogenic, having at least one pathogenic variant is significantly correlated with longer OS (logrank p-value: 0.037, Cox regression p-value: ns), see figure 4.21 B. The pathogenic class however depends more on the variants and genes within it, e.g., *KRAS* and/or *TP53* for most patients, knowing also *KRAS* is significant by itself. Although, the variant classification also showed a handful of patients with other pathogenic driving variants apart from *KRAS*.

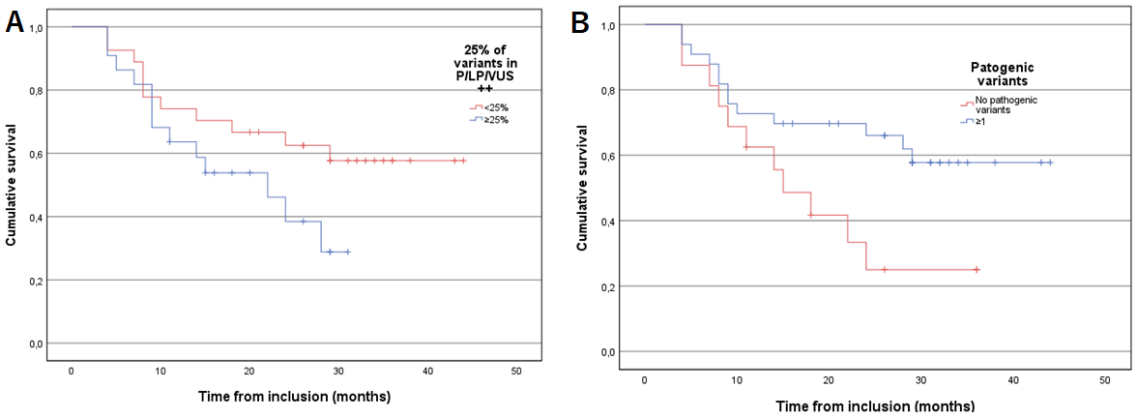


Figure 4.21: A: KM survival curve for percentage of variants within classes pathogenic, likely pathogenic or VUS++ (denoted P/LP/VUS++ in figure). Cut-off at 25%. B: KM survival curve for pathogenic variants with cut-off of at least 1 variant.

When mutational signatures were evaluated for OS, no threshold for cosine similarity was made; thresholds were instead set to 20% of activity frequency of the signature, and the signature being composed of at least 10 activities. Suggestively, patients with an oncogenic driver in combination with SBS4 signature show significantly longer OS as well (logrank p-value: 0.022; Cox regression p-value: 0.044). Both *KRAS* and *EGFR* is common in smokers and it is known that smokers with *KRAS* respond well to immunotherapy, however, what really respond might be the patients with smoking signatures, i.e also some of the non-smokers.

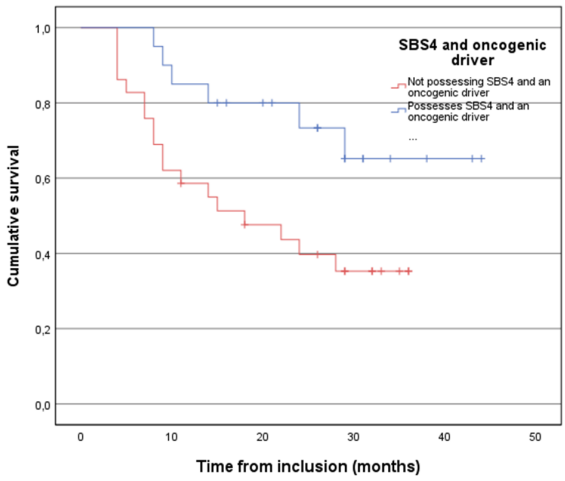


Figure 4.22: KM survival curve for SBS4 and an oncogenic driver.

None of the other classes tested were found to be significantly correlated with OS, hence will not be presented.

4.6.4 Biomarkers in plasma

Having ND ctDNA levels in at least one time point throughout the first five treatment cycles is even more strongly correlated to prolonged OS, see figure 4.23 A (logrank p-value: <0.001, Cox regression p-value: <0.001). In addition, Having ND ctDNA levels at baseline i.e. ≥1 MM, is significantly correlated to longer OS, see figure 4.23 B (logrank p-value: 0.049, Cox regression p-value: ns). As only 24 samples are considered, there is a need for more samples, and then perhaps ctDNA levels already at baseline can be a predictive biomarker. In one previous study 16 patient-specific variants were followed, the variants were selected based on among other factors, tumour VAF [29]. The results showed significant differences between increased or decreased ctDNA levels from baseline to third treatment cycle, in terms of OS and PFS. Increased ctDNA levels were associated with shorter OS and PFS.

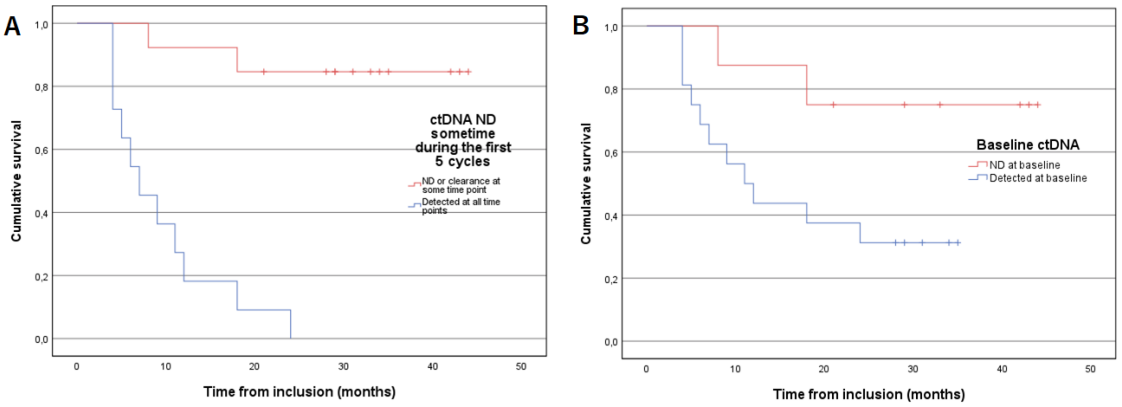


Figure 4.23: A: KM survival curve for ctDNA detection in at least one time point. B: KM survival curve for ctDNA detection at baseline.

The cohort is rather limited, nonetheless monitoring ctDNA is essential, already within the first treatment cycles, absence of ctDNA is a strong indicator of response to treatment. Clinical treatment monitoring today is based on CT scans performed every third month. CtDNA is thought to function as a complement to CT scans and depending on future studies might lead to fewer CT scans and perhaps only if indicated PD by ctDNA levels. CtDNA can by itself might be used as a prognostic and predicative biomarker. Before implementing clinically, the importance of using ctDNA monitoring must be established further with more research. Further, it would be of interest to have samples at more timepoints to further investigate the monitoring opportunities, perhaps one sample per month. The ctDNA project is ongoing and more patients are waiting to be included in the analysis, hopefully strengthening our current results.

4.7 Ethical and societal aspects

Since lung cancer has the second most incidents of cancer worldwide, a lot of people will benefit with more research in this field [1]. It has been established that most patients with NSCLC will not respond to treatment with ICB [2]. Therefore, it is inconvenient, economically and for patient safety, to give treatment to a patient who is unlikely to respond. Both for society at large, and for the cancer community, it is of great interest to establish biomarkers with higher efficiency. Treatments for cancer are very expensive, and the total health care costs for cancer patients in Sweden were in 2013 around 36 billion SEK [83]. The costs are expected to be around 70 billion SEK in 2040, with lung cancer being one of the four most costly. The increase is expected due to for instance, an aging population. If treatments can be more personalised, the cost of treatment would be reduced since only the patients likely to respond will be treated.

The Regional Ethical Review Authority in Gothenburg has given their permission to perform the BioLung study. All patients participating in the study have given their informed consent and are aware of what their samples is going to be used for. The patients also have the right to withdraw their consent if desired. Due to regulations regarding disclosure of information between regions in Sweden, all patients in the cohort are residents of the Västra Götaland region.

Inclusion in the BioLung cohort study is still ongoing. Their samples will be stored at Sahlgrenska Biobank (nr. 890) marked with an anonymous code that is protected by the Swedish Biobank in Medical Care Act (SFS 2002:297). The patients' integrity is always a high priority, therefore solely the study supervisors are aware of the patients' personal details. The personal details are also protected by the General Data Protection Regulation (EU 2016/670).

The samples used for the BioLung cohort study are collected at the regular sampling appointments, then a few additional blood samples are taken. A risk assessment has been done, showing that the additional samples are suggested to perhaps cause minor discomfort, however, the risk for long-term discomfort is minimal. The Swedish pharmaceutical insurance and the Swedish patient insurance are protecting the participants. In the consent form it is also clearly stated that the research probably will not benefit the participants directly, but this research could

rather help others get more personalised treatment.

The vast majority of lung cancer cases seen today are caused by smoking, actually around 85% of the patients are smokers or have been [84]. Worldwide, the number of people that smokes have increased, even though the proportion of smokers is decreasing. Among the never-smokers with lung cancer, exposure to for example asbestos and radon are some probable causes for their cancer. As lung cancer patients have a worse expected outcome compared to many other cancer types [84], it is crucial to investigate treatments more. In a purely hypothetical case, that the number of smokers would decrease, and thereby also the lung cancer incidence. Then some might argue that it should not be as important to focus the resources on lung cancer research. However, it is not exclusively smokers that get lung cancer, about 10-20% of the cases are not directly connected to smoking [84], meaning that they will still benefit from this research. Nevertheless, even in this hypothetical case, lung cancer research would still benefit other cancer form; to exemplify, insight in lung cancer could give knowledge on what determines responders from non-responders to ICB.

4.8 Future perspectives

Further, it is of great interest to analyse other omics apart from genomics, both transcriptomics and proteomics can give valuable information [15]. For example, analyses can be performed on if the mutated proteins are functioning or even transcribed. Analysing multiomics is however a costly process. In addition, recent attention has been directed towards identifying DNA methylation events in cancer, and there are long-read sequencing methods able to detect these events as simultaneously sequencing. In addition, single-cell sequencing techniques is also a hot-topic in cancer research, with the ability to breakdown complexity and also investigate cell types in normal and diseased material.

Difficulties remain with extracting mutational signatures for each sample from a sequencing panel, as more input data is needed to achieve trustworthy results. If WES or WGS is used, single samples can be used for extracting signatures, however, aiming to implement in clinical setting, a single sample must be able to extract data from, without having to rely on other samples. Factors as treatment cost, has to be weighted against cost of WGS or WES and also against the benefit of mutational signatures as a possible biomarker. An option may be to only select the patient with an oncogenic driver for WES/WGS and in turn mutational signatures, to reduce cost. A combination between SBS4 signature and an oncogenic driver, here *KRAS* or *EGFR*, has been correlated to longer OS in one previous study, and our results are in line.

Personally, I think that the future is to find more subgroups within NSCLC responders and non-responders with specific characteristics in common and thereby understand they response more. The development and mechanism behind of cancer and tumourigenesis is widely unknown and require more research, where investigating mutational signatures might be a natural point to start. To date, whole genome or exome sequencing is not standard procedure in clinical practice, although, the importance of implementation and molecular diagnostics have been proved. Clinically implementing whole genome or exome sequencing together with liquid biopsies

and ctDNA analyses would be a first step towards better and more equal cancer care treatment.

4.9 Delimitations

When the tumour samples are extracted from patients, they are stored in FFPE for conservation until further analyses. FFPE may result in reduced quality of the tumour DNA after biopsy, leading to an increased number of artefacts and false positives in the data. Some artefact structures have been linked to FFPE storage [39]. Nevertheless, FFPE storage is needed for pathology and determination of histodiagnosis.

When the true variants are determined, the artefacts are manually sorted out by analysis in IGV. On one hand, since this is a manual process, it introduces the human reliability as a possible source of error in the analyses. To ensure high quality, uncertain variants should be checked with the supervisor, to minimise discardment of the true variants. On the other hand, it is favorable that the variants are manually checked to sort out false positives and sequencing artefacts that otherwise, would still be in the data. Here, if the variants were not to be manually checked, if would have been over 5000 variants instead of 850 that is true.

The variant search is limited to SNVs and indels, leading to some important genes such as fusion genes not being considered. It would indeed be interesting to analyse other types of gene alterations as well, but this thesis have not considered those. It should also be noticed that there are a few limited cancer-related genes investigated, meaning that only some parts of the exons are sequenced. Only considering limited number of genes minimises the chance of investigating unknown genes and information present in the non-coding sequences and non-sequenced genes are not included.

Ethnicity is another factor that could effect the variant search and classification, as some variants are proved to be more common in some ethnic groups compared to other. The variant classification analysis takes ethnicities into account in the step to check if a variant is common in the general population. However, the ethnicity of the patients are unknown, hence this cannot be addressed properly. To exemplify, a non-Asian patient has a variant common in Asians, this variant is then removed as (likely) benign, when it in fact can be uncommon in non-Asians.

Perhaps it can be conflicting evidence since not all patients get immunotherapy as first-line treatment. Some have gotten for example radiation or chemotherapy, either before or in combination with immunotherapy. The observations of progression can possibly come from the other treatments as well. Especially concerning stage III patients, where immunotheraphy is given to ensure lower relapse rates. Nevertheless, effects of other treatments can be identified with for example mutational signatures, since some have a proposed aetiology of for instance chemotherapy, in addition is also known from the patient data.

5
Conclusion

Understanding the genetic landscape and identifying biomarkers of ICB are two key considerations for development of personalised treatment for patients. By applying mutational signature analysis, a deeper understanding of cancer origin and drivers was given, eventually leading to exclusion of one patient. Without analysing mutational signatures, the primary cancer would probably not have been noticed as an outlier, and the patient would have blurred results in finding biomarkers for NSCLC. The combination of mutational signatures and genetic variants further enhanced refinement of biomarkers of response to ICB, as hypothesised. Further, being in-line with the latest research suggesting SBS4 in combination with KRAS, or oncogenic driver, also poses as one of the best prognostic biomarkers from the analyses here.

The approach of a thorough classification including subclassification of the VUS:es led to identification of variants that can prognostically function as biomarkers, in combination with pathogenic and likely pathogenic variants. In particular the tumour suppressor gene *LRP1B* was found in multiple patients and were even more refined when split by classification. Also, suggesting that the classification is sufficient for less well-known potential cancer-causing genes. The sub-group of patients possessing co-occurring variants in *LRP1B/KRAS*, is a group responding well and correlate with long term response and significantly longer survival, which to our knowledge has not been shown priorly.

Bibliography

[1] Giustini N, Bazhenova L. Recognizing prognostic and predictive biomarkers in the treatment of non-small cell lung cancer (NSCLC) with immune checkpoint inhibitors (ICIs). Lung Cancer (Auckl). 2021;12:21–34. doi: 10.2147/LCTT.S235102.

[2] Hallqvist A, Rohlin A, Raghavan S. Immune checkpoint blockade and biomarkers of clinical response in non–small cell lung cancer. Scand J Immunol. 2020;92(6). doi: 10.1111/sji.12980.

[3] Wang L, Yue H, Wang S, Shen J, Wang X. Biomarkers of immunotherapy in non-small cell lung cancer (review). Oncol Lett. 2020;20(5). doi: 10.3892/OL.2020.11999.

[4] Cascone T, Fradette J, Pradhan M, Gibbons DL. Tumor immunology and immunotherapy of non-small-cell lung cancer. Cold Spring Harb Perspect Med. 2022;12(5). doi: 10.1101/cshperspect.a037895.

[5] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–47. doi: 10.1016/j.ejca.2008.10.026.

[6] Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. Am J Cancer Res. 2021;11(4):1121–31.

[7] Rossi G, Russo A, Tagliamento M, Tuzi A, Nigro O, Vallome G, et al. Precision medicine for NSCLC in the era of immunotherapy: new biomarkers to select the most suitable treatment or the most suitable patient. Cancers (Basel). 2020;12(5). doi: 10.3390/cancers12051125.

[8] Läkartidningen. 2021;118:21115.

[9] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. doi: 10.1038/gim.2015.30.

[10] Ogino S, Gulley ML, den Dunnen JT, Wilson RB. Standard Mutation Nomenclature in Molecular Diagnostics. J of Mol Diagn. 2007;9(1):1–6. doi: 10.2353/jmoldx.2007.060081.

[11] Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015;349(6255):1483–9. doi: 10.1126/science.aab4082.

[12] Muzzey D, Evans EA, Lieber C. Understanding the basics of NGS: from mechanism to variant calling. Curr Genet Med Rep. 2015;3(4):158–65. doi: 10.1007/s40142-015-0076-8.

[13] Chakravarty D, Solit DB. Clinical cancer genomic profiling. Nat Rev Genet. 2021;22(8):483–501. doi: 10.1038/s41576-021-00338-8.

[14] Froyen G, Mercier ML, Lierman E, Vandepoele K, Nollet F, Boone E, et al. Standardization of somatic variant classifications in solid and haematological tumours by a two-level approach of biological and clinical classes: an initiative of the belgian compermed expert panel. Cancers (Basel). 2019;11(12). doi: 10.3390/cancers11122030.

[15] Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72. doi: 10.1038/s41568-020-0290-x.

[16] Chial H. Proto-oncogenes to oncogenes to cancer. Nature educ. 2008;1(1):33.

[17] Shen L, Shi Q, Wang W. Double agents: genes with both oncogenic and tumor-suppressor functions. Oncogenesis. 2018;7(3):25. doi: 10.1038/s41389-018-0034-x.

[18] Juul RI, Nielsen MM, Juul M, Feuerbach L, Pedersen JS. The landscape and driver potential of site-specific hotspots across cancer genomes. NPJ Genom Med. 2021;6(1):33. doi: 10.1038/s41525-021-00197-6.

[19] Baeissa H, Benstead-Hume G, Richardson CJ, Pearl FMG. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget. 2017;8(13):21290–304. doi: 10.18632/oncotarget.15514.

[20] Rohlin A, Zagoras T, Nilsson S, Lundstam U, Wahlström J, Hulthén L, et al. A mutation in POLE predisposing to a multi-tumour phenotype. Int J Oncol. 2014;45(1):77–81. doi: 10.3892/ijo.2014.2410.

[21] Chevallier M, Borgeaud M, Addeo A, Friedlaender A. Oncogenic driver mutations in non-small cell lung cancer: Past, present and future. World J Clin Oncol. 2021;12(4):217–37. doi: 10.5306/wjco.v12.i4.217.

[22] Cisowski J, Bergo MO. What makes oncogenes mutually exclusive? Small GTPases. 2017;8(3):187–92. doi: 10.1080/21541248.2016.1212689.

[23] Lee T, Lee B, Choi YL, Han J, Ahn MJ, Um SW. Non-small Cell Lung Cancer with Concomitant EGFR, KRAS, and ALK Mutation: Clinicopathologic Features of 12 Cases. J Pathol Transl Med. 2016;50(3):197–203. doi: 10.4132/jptm.2016.03.09.

[24] van den Heuvel GRM, Kroeze LI, Ligtenberg MJL, Grünberg K, Jansen EAM, von Rhein D, et al. Mutational signature analysis in non-small cell lung cancer patients with a high tumor mutational burden. Respir Res. 2021;22(1). doi: 10.1186/s12931-021-01871-0.

[25] Rolfo C, Mack P, Scagliotti GV, Aggarwal C, Arcila ME, Barlesi F, et al. Liquid biopsy for advanced NSCLC: a consensus statement from the international association for the study of lung cancer. J Thorac Oncol. 2021;16(10):1647–62. doi: 10.1016/j.jtho.2021.06.017.

[26] Andersson D, Kristiansson H, Kubista M, Ståhlberg A. Ultrasensitive circulating tumor DNA analysis enables precision medicine: experimental workflow considerations. Expert Rev Mol Diagn. 2021;21(3):299–310. doi: 10.1080/14737159.2021.1889371.

[27] Chen L, Eriksson A, Weström S, Pandzic T, Lehmann S, Cavelier L, et al. Ultrasensitive monitoring of leukemia patients using superRCA mutation detection assays. Nat Commun. 2022;13(1):4033. doi: 10.1038/s41467-022-31397-y.

[28] Song Y, Hu C, Xie Z, Wu L, Zhu Z, Rao C, et al. Circulating tumor DNA clearance predicts prognosis across treatment regimen in a large real-world longitudinally monitored advanced non-small cell lung cancer cohort. Transl Lung Cancer Res. 2020;9(2):269–79. doi: 10.21037/tlcr.2020.03.17.

[29] Bratman SV, Yang SYC, Iafolla MAJ, Liu Z, Hansen AR, Bedard PL, et al. Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. Nat Cancer. 2020;1(9):873–81. doi: 10.1038/s43018-020-0096-5.

[30] Chow YP, Zainul Abidin N, Kow KS, Tho LM, Wong CL. Analytical and clinical validation of a custom 15-gene next-generation sequencing panel for the evaluation of circulating tumor DNA mutations in patients with advanced non-small-cell lung cancer. PLOS ONE. 2022;17(10):e0276161. doi: 10.1371/journal.pone.0276161.

[31] Bos MK, Nasserinejad K, Jansen MPHM, Angus L, Atmodimedjo PN, de Jonge E, et al. Comparison of variant allele frequency and number of mutant molecules as units of measurement for circulating tumor DNA. Mol Oncol. 2021;15(1):57–66. doi: 10.1002/1878-0261.12827.

[32] 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(7463):415–21. doi: 10.1038/nature12477.

[33] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):941–7. doi: 10.1093/nar/gky1015.

[34] Brady SW, Gout AM, Zhang J. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends Genet. 2022;38(2):194–208. doi: 10.1016/j.tig.2021.08.007.

[35] Alexandrov L, Nik-Zainal S, Wedge D, Campbell P, Stratton M. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3(1):246–59. doi: 10.1016/j.celrep.2012.12.008.

[36] Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 2019;20(1):685. doi: 10.1186/s12864-019-6041-2.

[37] Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genom. 2022 11;2(11):100179. doi: 10.1016/j.xgen.2022.100179.

[38] Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10(1):33. doi: 10.1186/s13073-018-0539-0.

[39] Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. The mutational signatures of formalin fixation on the human genome. Nat Commun. 2022;13(1):4487. doi: 10.1038/s41467-022-32041-5.

[40] Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. J Clin Med. 2020;9(1):132. doi: 10.3390/jcm9010132.

[41] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 2003 7;89(2):232–8. doi: 10.1038/sj.bjc.6601118.

[42] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. Br J Cancer. 2003 8;89(3):431–6. doi: 10.1038/sj.bjc.6601119.

[43] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6. doi: 10.1038/nbt.1754.

[44] 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(7809):434–43. doi: 10.1038/s41586-020-2308-7.

[45] Chen S, Francioli L, Goodrich J, Collins R, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv [Preprint]. 2022. doi: 10.1101/2022.03.20.485034.

[46] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):1062–7. doi: 10.1093/nar/gkx1153.

[47] 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(2):155–63. doi: 10.1038/nbt.3391.

[48] Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue MTA, Jonsson P, et al. Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discov. 2018;8(2):174–83. doi: 10.1158/2159-8290.CD-17-0321.

[49] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705. doi: 10.1038/s41568-018-0060-1.

[50] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546–58. Table S2A, Driver genes affected by subtle mutations. doi: 10.1126/science.1235122.

[51] Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 2016 1;44(D1):1023–31. doi: 10.1093/nar/gkv1268.

[52] Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;(1):1–16. doi: 10.1200/PO.17.00011.

[53] Borges P, Pasqualim G, Matte U. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants. Front Mol Biosci. 2021;8. doi: 10.3389/fmolb.2021.752797.

[54] Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical rele-

vance of tumor alterations. Genome Med. 2018;10(1):25. doi: 10.1186/s13073-018-0531-8.

[55] Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80. doi: 10.1093/bioinformatics/bty897.

[56] Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res. 2018;46(W1):530–6. doi: 10.1093/nar/gky355.

[57] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007;35:5–12. doi: 10.1093/nar/gkl1031.

[58] Chunn LM, Nefcy DC, Scouten RW, Tarpey RP, Chauhan G, Lim MS, et al. Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation. Front Genet. 2020;11. doi: 10.3389/fgene.2020.577152.

[59] Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170–4. doi: 10.1038/ng.3774.

[60] Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods. 2016;13(10):806–7. doi: 10.1038/nmeth.4000.

[61] Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):523–31. doi: 10.1093/nar/gkac1052.

[62] Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163–77. doi: 10.1016/j.ajhg.2022.10.013.

[63] Yampolsky LY, Stoltzfus A. The exchangeability of amino acids in proteins. Genetics. 2005;170(4):1459–72. doi: 10.1534/genetics.104.039107.

[64] de Andrade KC, Lee EE, Tookmanian EM, Kesserwan CA, Manfredi JJ, Hatton JN, et al. The TP53 Database: transition from the International Agency for Research on Cancer to the US National Cancer Institute. Cell Death Differ. 2022 5;29(5):1071–3. doi: 10.1038/s41418-022-00976-3.

[65] Soussi T, Leroy B, Taschner PEM. Recommendations for Analyzing and Reporting TP53 Gene Variants in the High-Throughput Sequencing Era. Hum Mutat. 2014;35(6):766–78. doi: 10.1002/humu.22561.

[66] Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017;100(2):267–80. doi: 10.1016/j.ajhg.2017.01.004.

[67] Cline MS, Liao RG, Parsons MT, Paten B, Alquaddoomi F, Antoniou A, et al. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. PLOS Genet. 2018;14(12):e1007752. doi: 10.1371/journal.pgen.1007752.

[68] Fokkema IFAC, Kroon M, López Hernández JA, Asscheman D, Lugtenburg I, Hoogenboom J, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet. 2021;29(12):1796–1803. doi: 10.1038/s41431-021-00959-x.

[69] Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, et al. GenVisR: Genomic Visualizations in R. Bioinformatics. 2016;32(19):3012–4. doi: 10.1093/bioinformatics/btw325.

[70] Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12(2):477–9. doi: 10.1039/C5MB00663E.

[71] Ståhlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc. 2017 4;12(4):664–82. doi: 10.1038/nprot.2017.006.

[72] Ståhlberg A, Krzyzanowski PM, Jackson JB, Egyud M, Stein L, Godfrey TE. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res. 2016 6;44(11):e105. doi: 10.1093/nar/gkw224.

[73] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101. doi: 10.1038/s41586-020-1943-3.

[74] Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354(6312):618–22. doi: 10.1126/science.aag0299.

[75] Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019;176(6):1282–94. doi: 10.1016/j.cell.2019.02.012.

[76] Boot A, Ng AWT, Chong FT, Ho SC, Yu W, Tan DSW, et al. Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. Genome Res. 2020;30(6):803–13. doi: 10.1101/gr.255620.119.

[77] Landi MT, Synnott NC, Rosenbaum J, Zhang T, Zhu B, Shi J, et al. Tracing Lung Cancer Risk Factors Through Mutational Signatures in Never-Smokers. Am J Epidemiol. 2021;190(6):962–76. doi: 10.1093/aje/kwaa234.

[78] Sequist LV, Neal JW. Personalized, genotype-directed therapy for advanced non-small cell lung cancer; 2016. Available from: `https://www.uptodate.com/contents/personalized-genotype-directed-therapy-for-advanced-non-small-cell-lung-cancer`.

[79] Chen JW, Dhahbi J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. Sci Rep. 2021;11(1):13323. doi: 10.1038/s41598-021-92725-8.

[80] Strijker M, Soer EC, Pastena M, Creemers A, Balduzzi A, Beagan JJ, et al. Circulating tumor DNA quantity is related to tumor volume and both predict

survival in metastatic pancreatic ductal adenocarcinoma. International Journal of Cancer. 2020;146(5):1445–56. doi: 10.1002/ijc.32586.

[81] Zhang SM, Zhu QG, Ding XX, Lin S, Zhao J, Guan L, et al. Prognostic value of EGFR and KRAS in resected non-small cell lung cancer: a systematic review and meta-analysis. Cancer Manag Res. 2018;10:3393–404. doi: 10.2147/CMAR.S167578.

[82] Eklund EA, Wiel C, Fagman H, Akyürek LM, Raghavan S, Nyman J, et al. KRAS Mutations Impact Clinical Outcome in Metastatic Non-Small Cell Lung Cancer. Cancers. 2022;14(9):2063. doi: 10.3390/cancers14092063.

[83] Lundqvist A, Andersson E, Steen Carlsson K. Kostnader för cancer i Sverige idag och år 2040 [Internet]. Lund: The Swedish Insitutet for Health Economics (IHE); 2016. IHE Report; 2016:1. [cited 2022 Sep 20]. Available from: `https://ihe.se/publicering/kostnad-for-cancer/`.

[84] The Public Health Agency of Sweden. Lungcancer, död [Internet]. Stockholm: The Public Health Agency of Sweden; 2022 [cited 2022 Sep 21]. Available from: `https://www.folkhalsomyndigheten.se/fu-lungcancer-dodlighet`.

A

Appendix 1

Table A.1 show the Consensus Pathogenic Variants list for solid tumours, the list is an essential part of the workflow for variant classification.

Table A.1: Consensus Pathogenic Variant (CPV) list of the ComPerMed genes selected for screening in solid tumours. Hs: Hotspot; if-del: inframe deletion; if-ins: inframe insertion; _: denotes the exact positions of that change; -: denotes a region in which the change has to be located; LoF: Loss of Function.

Gene	Transcript-ID	Hs1	Hs2	Hs3	Hs4	Hs5	Hs6	Hs7	Hs8	Hs9	Hs10	Hs11	Hs12
ALK	NM_004304.4	F1174L	R1275Q										
BRAF	NM_004333.5	G469A/E/R/V	D594G/M	T599-K601 if-del/ins	V600E/K/M/R	K601E							
BRCA1	NM_007294.3	all clear LoF variants (nonsense, frameshift, splice site)											
BRCA2	NM_000059.3	all clear LoF variants (nonsense, frameshift, splice site)											
EGFR	NM_005228.4	G719A/C/S	ex19if-del/ins	ex20 if-ins	T790M	C797S	L858R	L861Q					
ESR1	NM_000125.3	K303R	E380Q	V392I	S463P	V533M	V534E	P535H	L536H/P/Q/R	Y537C/N/S	D538G		
GNAS	NM_000516.5	R201C/H											
H3F3A	NM_002107.4	K28M	G35R/W										
HRAS	NM_005343.3	G12C/D/S/V	G13C/D/R/S/V	Q61H/K/L/R									
IDH1	NM_005896.3	R132C/G/H/L/S											
IDH2	NM_002168.3	R140L/Q/W	R172K/M/S										
KIT	NM_000222.2	ex8 D419 if-del	ex9 S501-F504 if-ins	ex11 K550-V560 if-indel	ex11 W557G/R	ex11 V559A/D	ex11 V560D	ex11 L576P	ex13 K642E	ex13 V654A	ex14 T670I	ex17 D816H/V/Y	ex17 N822K
KRAS	NM_004985.4	G12A/C/D/F/R/S/V	G13C/D/R/S/V	A59T	Q61H/K/L/R	K117N	A146T						
MET	NM_001127500.3	ex14 skipping											
NRAS	NM_002524.4	G12A/C/D/R/S/V	G13C/D/R/S/V	A59T	Q61H/K/L/R	K117N	A146T						
PDGFRA	NM_006206.5	S566_E577 if-del	D842V	D842_I843 if-del	V561D								

B

Appendix 2 - RStudio script

B.1 ReactomePA

```
# ReactomePA, 230102 Johanna Svensson

b=table(genelist)

#Required packages
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
 BiocManager::install(version = "3.16")
#BiocManager::install(version = "3.15")
 BiocManager::install("DOSE")
 BiocManager::install("ReactomePA")
 BiocManager::install("clusterProfiler")
 BiocManager::install("org.Hs.eg.db")
 BiocManager::install("ggnewscale")
 BiocManager::install("enrichplot")
 BiocManager::install("enrichMap")
 BiocManager::install("bitr")
 BiocManager::install("BiocParallel")
 BiocManager::install("cli")
 BiocManager::install("purrr")
 BiocManager::install("RSQLite")
 BiocManager::install("ggupset")
 BiocManager::install("europepmc")
 BiocManager::install("forcats")


# Start from here
library("ReactomePA")
library("DOSE")
library("clusterProfiler")
library("org.Hs.eg.db")
library("ggnewscale")
library("enrichplot")
library("ggupset")
```

```
library("europepmc")
library("ggplot2")
library("forcats")
library("graphite")


#Load data and prepare data sets (all genes)
genelist <- read.table(file = "~/Chalmers/MPBIO-2/
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Exjobb/Project/Pathway␣Analysis/
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣genelist.txt", sep = "\t")

colnames(genelist) <- c("sample", "gene")
clinical <- read.table(file = "~/Chalmers/MPBIO-2/
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Exjobb/Project/Pathway␣Analysis/
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣biolung.txt", sep = "\t")
clinical <- clinical[-1,]
clinical <- as.data.frame(clinical[,c(1,31)])
colnames(clinical) <- c("sample","response")

non_list <- clinical[clinical$response %in%
                       c("Responder"),]
colnames(non_list) <- c("sample","response")
non_list <- non_list[,c("sample")]

resp_list <- clinical[clinical$response %in%
                        c("Non-responder"),]
colnames(resp_list) <- c("sample","response")
resp_list<- resp_list[,c("sample")]

non_responders <- genelist[genelist$sample
                             %in% non_list,]
responders <- genelist[genelist$sample
                         %in% resp_list,]


#All
x_all <- unlist(c(genelist$gene))
all <- (mapIds(org.Hs.eg.db, x_all, 'ENTREZID', 'SYMBOL'))

pw_all <- enrichPathway(gene=all,
                        pvalueCutoff=0.05,
                        readable=T)
pw2_all <- pairwise_termsim(pw_all)
barplot(pw_all, showCategory=8, title=
          "Enriched␣pathways␣in␣all␣patients")
dotplot(pw_all, showCategory=20, title=
```

```
          "Enriched␣pathways,␣in␣all␣patients")
emapplot(pw2_all, title=
           "Enriched␣pathways,␣in␣all␣patients")
cnetplot(pw_all, categorySize="pvalue",
         title="Enriched␣pathways,␣in␣all␣patients")
cnetplot(pw_all, circular = T,
         title="Enriched␣pathways,␣in␣all␣patients")


#Responders vs non-responders
x_non <- unlist(c(non_responders$gene))
non <- (mapIds(org.Hs.eg.db, x_non,
                'ENTREZID', 'SYMBOL'))

x_resp <- unlist(c(responders$gene))
resp <- (mapIds(org.Hs.eg.db, x_resp,
                 'ENTREZID', 'SYMBOL'))


pw_non <- enrichPathway(gene=non,
                        pvalueCutoff=0.05,
                        readable=T)
pw2_non<- pairwise_termsim(pw_non)
#head(as.data.frame(pw_non))

selected_pathways <- c("DNA␣Repair",
                       "Diseases␣of␣DNA␣repair")

barplot(pw_non, showCategory=8, title=
          "Enriched␣pathways,␣in␣non-responding
␣␣␣␣␣␣␣␣patients")
dotplot(pw_non, showCategory=
          selected_pathways, title="Enriched␣pathways,
␣␣␣␣␣␣␣␣in␣non-responding␣patients")
emapplot(pw2_non, title="Enriched␣pathways,
␣␣␣␣␣␣␣␣␣␣in␣non-responding␣patients")
cnetplot(pw_non, categorySize="pvalue",
         title="Enriched␣pathways,
␣␣␣␣␣␣␣␣␣␣in␣non-responding␣patients")
cnetplot(pw_non, circular = T, title="Enriched␣pathways,
␣␣␣␣␣␣␣␣␣␣in␣non-responding␣patients")

pw_resp <- enrichPathway(gene=resp,
                         pvalueCutoff=0.05,
                         readable=T)
pw2_resp <- pairwise_termsim(pw_resp)
```

```
#head(as.data.frame(pw_resp))
barplot(pw_resp, showCategory=8,
        title="Enriched␣pathways,␣in␣responding␣patients")
dotplot(pw_resp, showCategory=20,
        title="Enriched␣pathways,␣in␣responding␣patients")
emapplot(pw2_resp, title="Enriched␣pathways,
␣␣␣␣␣␣␣␣␣␣in␣responding␣patients")
cnetplot(pw_resp, categorySize="pvalue",
         title="Enriched␣pathways,␣in␣responding␣patients")
cnetplot(pw_resp, circular = T,
         title="Enriched␣pathways,␣in␣responding␣patients")

heatmap(pw$result)

heatplot(pw_all, showCategory = 5)

#Treeplot
treeplot(pw2_all, cluster.params =
           list(method = "average"))
treeplot(pw2_non, cluster.params =
           list(method = "average"))
treeplot(pw2_resp, cluster.params =
           list(method = "average"))


#Comparison, responders vs non-responders
compare <- list(non = non, resp = resp)

require(clusterProfiler)
comparison <- compareCluster(compare,
                             fun = "enrichPathway",
                             organism = "human",
                             pvalueCutoff=0.05)

comparison2 <- pairwise_termsim(comparison)
dotplot(comparison, showCategory =
          selected_pathways, includeAll = F)
emapplot(comparison2, showCategory =
           selected_pathways, pie.params =
           list(pie = "count"))
cnetplot(comparison, showCategory =
           selected_pathways,
         categorySize="pvalue",
         name)
upsetplot(pw_all)
pmcplot(pw_all$Description[1:5], 2010:2020,
```

```
        proportion = F)
```

B.2 Waterfall plot

```
#GenVisR, 230511 Johanna Svensson

#Required packages
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install(version = "devel")
BiocManager::install("remotes")
remotes::install_github("cran/FField", force = T)
BiocManager::install("DOSE", force = T)
BiocManager::install("dbplyr", force = T)
BiocManager::install("GenVisR", force = T)
BiocManager::install("ggtext")
BiocManager::install("gplots")
BiocManager::install("TxDb.Hsapiens.UCSC.hg38.knownGene")
BiocManager::install("BSgenome.Hsapiens.UCSC.hg38")
BiocManager::install("GenomeInfoDb", force = T)
BiocManager::install("GenomicFeatures", force = T)

dev.off(dev.list()["RStudioGD"])

#Start from here
library("ReactomePA")
library("DOSE")
library("clusterProfiler")
library("org.Hs.eg.db")
library("ggnewscale")
library("enrichplot")
library("DOSE")
library("GenVisR")
library("gridExtra")
library("ggplot2")
library("reshape2")
library("gplots")
library("GenomeInfoDb")


#Load data from text files
genes <- read.table(file = "~/to_waterfall_plot.txt",
    sep = "\t")
genes = genes[-1,]
tmb <- read.table(file = "~/tmb3.txt", sep = "\t")
```

```
tmb = tmb[-1,]
clin <- read.table(file = "~/biolung.txt", sep = "\t")
clin = clin[-1,]

set.seed(426)


#Import all data
mutationData <- as.data.frame(c(genes))
colnames(mutationData) <- c("sample","variant_class","gene")
mutation_priority <-
        as.character(unique(mutationData$variant_class))
mutationColours <- c("nonsynonymous"='#CCCCFF',
                     "synonymous"='#cc99ff',
                      "frameshift"='#6666ff',
                      "splice site"='#000000',
                      "stopgain"='#330066',
                      "deletion"='#750054',
                      "inframe"='#A80079',
                      "startloss"='#ca66ae',
                      "stoploss"='#FFCCFF')

#Customised TMB
mutationBurden <- as.data.frame(c(tmb))
colnames(mutationBurden) <- c("sample","mut_burden")
mutationBurden$sample <- gsub("^WU(0)+", "",
                     mutationBurden$sample)


#Clinical information
clinData <- as.data.frame(clin[,c(1,3,8,4)])
colnames(clinData) <- c("sample", "Sex", "Response",
                     "Histology")
clinData <- reshape2::melt(data=clinData,
            id.vars=c("sample"))
clinicalColours <- c("Male"="#0066cc","Female"="#cc66ff",
                      "Responder"="#ff99cc",
                      "Non-responder"="#ff6666",
                      "LUAD"="#99ccff",
                      "LUSC"="#cc3399",
                      "NOS"="#000033")
clinicalOrder <- c("Responder","Non-responder","Female",
                    "Male","LUAD","LUSC","NOS")


#Waterfall plot for all
```

```
waterfall(mutationData,
          fileType = "Custom",
          variant_class_order = mutation_priority,
          mainRecurCutoff = 0.05,
          maxGenes = 8,
          mainXlabel = F,
          mainGrid = F,
          clinDat = clinData,
          mainPalette = mutationColours,
          mutBurden = mutationBurden,
          clinVarCol = clinicalColours,
          clinVarOrder = clinicalOrder,
          clinLegCol = 4,
          section_heights=c(2, 7, 2),
          mainDropMut = F,
          main_geneLabSize = 12)


#Waterfall plot for LUAD.
#Load data from text files
genes <- read.table(file = "~/to_waterfall_plot_LUAD.txt",
        sep = "\t")
genes = genes[-1,]
tmb <- read.table(file = "~/tmb3_LUAD.txt", sep = "\t")
tmb = tmb[-1,]
clin <- read.table(file = "~/biolung_LUAD.txt", sep = "\t")
clin = clin[-1,]

#Import all data
mutationData <- as.data.frame(c(genes))
colnames(mutationData) <- c("sample","variant_class","gene")
mutation_priority <-
        as.character(unique(mutationData$variant_class))
mutationColours <- c("nonsynonymous"='#CCCCFF',
                     "synonymous"='#cc99ff',
                     "frameshift"='#6666ff',
                     "splice site"='#000000',
                     "stopgain"='#330066',
                     "deletion"='#750054',
                     "inframe"='#A80079',
                     "startloss"='#ca66ae',
                     "stoploss"='#FFCCFF')

#Customised TMB
mutationBurden <- as.data.frame(c(tmb))
colnames(mutationBurden) <- c("sample","mut_burden")
```

```
mutationBurden$sample <- gsub("^WU(0)+", "",
                              mutationBurden$sample)


#Clinical information
clinData <- as.data.frame(clin[,c(1,3,8)])
colnames(clinData) <- c("sample", "Sex", "Response")
clinData <- reshape2::melt(data=clinData,
            id.vars=c("sample"))
clinicalColours <- c("Male"="#0066cc",
                     "Female"="#cc66ff",
                     "Responder"="#ff99cc",
                     "Non-responder"="#ff6666")
clinicalOrder <- c("Responder","Non-responder",
                   "Female","Male")


#Waterfall plot for LUAD
waterfall(mutationData,
          fileType = "Custom",
          variant_class_order = mutation_priority,
          mainRecurCutoff = 0.05,
          maxGenes = 10,
          mainXlabel = F,
          mainGrid = F,
          clinDat = clinData,
          mainPalette = mutationColours,
          mutBurden = mutationBurden,
          clinVarCol = clinicalColours,
          clinVarOrder = clinicalOrder,
          clinLegCol = 2,
          section_heights=c(2, 7, 2),
          mainDropMut = F,
          main_geneLabSize = 12)




#Waterfall plot for LUSC.
#Load data from text files
genes <- read.table(file = "~/to_waterfall_plot_LUSC.txt",
        sep = "\t")
genes = genes[-1,]
tmb <- read.table(file = "~/tmb3_LUSC.txt", sep = "\t")
tmb = tmb[-1,]
clin <- read.table(file = "~/biolung_LUSC.txt", sep = "\t")
```

```
clin = clin[-1,]

#Import all data
mutationData <- as.data.frame(c(genes))
colnames(mutationData) <- c("sample","variant_class","gene")
mutation_priority <-
        as.character(unique(mutationData$variant_class))
mutationColours <- c("nonsynonymous"='#CCCCFF',
                     "synonymous"='#cc99ff',
                     "frameshift"='#6666ff',
                     "splice site"='#000000',
                     "stopgain"='#330066',
                     "deletion"='#750054',
                     "inframe"='#A80079',
                     "startloss"='#ca66ae')

#Customised TMB
mutationBurden <- as.data.frame(c(tmb))
colnames(mutationBurden) <- c("sample","mut_burden")
mutationBurden$sample <- gsub("^WU(0)+", "",
                  mutationBurden$sample)


#Clinical information
clinData <- as.data.frame(clin[,c(1,3,8)])
colnames(clinData) <- c("sample", "Sex", "Response")
clinData <- reshape2::melt(data=clinData,
            id.vars=c("sample"))
clinicalColours <- c("Male"="#0066cc",
                     "Female"="#cc66ff",
                     "Responder"="#ff99cc",
                     "Non-responder"="#ff6666")
clinicalOrder <- c("Responder","Non-responder",
                  "Female", "Male")


#Waterfall plot for LUSC
waterfall(mutationData,
          fileType = "Custom",
          variant_class_order = mutation_priority,
          mainRecurCutoff = 0.05,
          maxGenes = 10,
          mainXlabel = F,
          mainGrid = F,
          clinDat = clinData,
          mainPalette = mutationColours,
```

```
mutBurden = mutationBurden,
clinVarCol = clinicalColours,
clinVarOrder = clinicalOrder,
clinLegCol = 2,
section_heights=c(2, 7, 2),
mainDropMut = F,
main_geneLabSize = 12)
```

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY