FACULTY OF
BIOSCIENCE ENGINEERING

A PERSONALISED THERAPEUTIC VACCINE FOR CANCER IMMUNOTHERAPY: TOWARDS AN OPTIMISATION OF NEOANTIGEN DETECTION IN LUNG CANCER USING RNA SEQUENCING

Number of words: 27.798

Lore Van Oudenhove
Student number: 01403359

Promotors:
Prof. dr. ir. B. Menten
dr. ir. G. Menschaert

Tutor:
ir. Laurenz De Cock

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master Of Science in Bioscience Engineering: Cell and Gene Biotechnology

Academic year: 2018 – 2019

GHENT
UNIVERSITY

WOORD VOORAF

Met veel motivatie en ijver begon ik vijf jaar geleden aan de opleiding Bio-ingenieurswetenschappen. Gedurende deze periode kreeg ik de kans om met verschillende onderzoeksdomeinen in aanraking te komen. Niettemin bleek kankeronderzoek al snel mijn grote passie te zijn. In het derde jaar koos ik dan ook voor de afstudeerrichting Cel- en genbiotechnologie. Naast *Immunologie* en *Medical Biotechnology* was het vooral het vak *Bio-informatica* dat mijn interesse in gepersonaliseerde kankertherapieën heeft aangesterkt. Ik ben dan ook zeer dankbaar dat ik tijdens mijn laatste jaar aan een thesisonderwerp heb kunnen werken dat zo nauw aanleunt bij mijn grootste interesse.

Ik zou dan ook mijn promotors Prof. dr. ir. Björn Menten en dr. ir. Gerben Menschaert willen bedanken om mij de kans te geven mee te werken aan onderzoek dat mij nauw aan het hart ligt. Daarnaast zou ik in het bijzonder mijn begeleider Laurenz De Cock willen bedanken om mij wegwijs te maken in de wereld van bio-informatica. Bedankt om mij te helpen bij de talrijke *errors* die mijn pad kruisten en steeds mee te zoeken naar een oplossing. Je stond steeds paraat om mij bij te staan met raad en daad. Verder wil ik je ook bedanken voor de tijd en energie die je gestoken hebt in het nalezen en verbeteren van deze masterproef.

Tenslotte wil ik ook mijn vrienden, mijn familie, mijn ouders, mijn broer en mijn vriendin bedanken voor de onvoorwaardelijke steun die ik van hen kreeg zowel tijdens het schrijven van deze masterproef, alsook tijdens mijn volledige opleiding tot bio-ingenieur. Mijn vrienden zou ik willen bedanken om mijn studententijd zo aangenaam te maken. De momenten die we samen hebben meegemaakt staan voor altijd in mijn geheugen gegrift. Daarnaast wil ook mijn vriendin, Laurien, bedanken om er op elk moment voor mij te zijn en elk alledaags moment speciaal te maken. Tenslotte wil ik in het bijzonder mijn ouders bedanken om mij de kans te geven te studeren aan de UGent en mij altijd te steunen, zowel op studievlak als daarbuiten. Woorden schieten mij te kort om te omschrijven hoe belangrijk zo’n warm gezin voor mij is. Duidend maal dank allemaal!

Lore Van Oudenhove
Gent, juni 2019

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A	Adenosine
ACT	Adoptive cell transfer
ADAR	Adenosine deaminase
APC	Antigen-presenting cell
bcbio	Blue Collar Bioinformatics
BWA	Burrows-Wheeler Aligner
C	Cytosine
CAR	Chimeric antigen receptor
cDNA	Complementary DNA
CLIP	Class II-associated invariant chain peptide
CMGG	Center for Medical Genetics Ghent
CTL	Cytotoxic T lymphocyte
CTLA-4	Cytotoxic T-lymphocyte-associated antigen 4
DC	Dendritic cell
DNA	Deoxyribonucleic acid
DP	Read depth
EGFR	Epidermal growth factor receptor
ER	Endoplasmic reticulum
FF	Fresh frozen
FFPE	Formalin-fixed paraffin-embedded
FN	False negative
FP	False positive
G	Guanine
GATK	Genome Analysis Toolkit
GMP	Good manufacturing practice
GSNAP	Genomic Short-read Nucleotide Alignment Program
HLA	Human leukocyte antigen
I	Inosine
IC_{50}	Half maximal inhibitory concentration
ICI	Immune checkpoint inhibition
IFN-ɣ	Interferon-gamma
IGV	Integrative Genomics Viewer
IL-12	Interleukin-12
Indel	Insertion and deletion
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MDSC	Myeloid-derived suppressor cell
MHC	Major histocompatibility complex
MMP	Maximal Mappable Prefix
mRNA	Messenger RNA
NGS	Next-generation sequencing
NSCLC	Non-small cell lung carcinoma
PCR	Polymerase chain reaction
PD-1	Programmed Death 1

PON	Panel-Of-Normals
pRb	Protein retinoblastoma
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
rRNA	Ribosomal RNA
RT	Reverse transcriptase
SCLC	Small cell lung carcinoma
snoRNA	Small nucleolar RNA
SNP	Single-nucleotide polymorphisms
SNV	Single-nucleotide variant
SSH	Secure Shell
STAR	Splice Transcripts Alignment to a Reference
T	Thymidine
TAA	Tumour-associated antigens
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor
TIL	Tumour infiltrating lymphocytes
TN	True negative
TP	True positive
TPM	Transcripts per kilobase million
Treg	Regulatory T cell
tRNA	Transfer RNA
TSA	Tumour-specific antigens
U	Uracil
VAF	Variant allele frequency
VCF	Variant call format
VEP	Variant Effect Predictor
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

ABSTRACT

Cancer is characterized by an accumulation of genetic alternations resulting in the development of abnormal cells that divide uncontrollably and have the potential to invade or spread to other parts of the body. Somatic mutations can generate tumour-specific neoantigens that can be presented as neoepitopes on malignant cancer cells. Subsequently, recognition of neoepitopes by immune cells can induce an anticancer immune response. Therefore, these neoepitopes are considered ideal cancer vaccine targets. Recent advances in next-generation sequencing technology and novel bioinformatics tools have enabled the careful identification of cancer-specific somatic mutations and subsequent neoantigens, therefore, paving the way for the development of a personalised therapeutic vaccine. The bioinformatics pipeline involved in the identification of neoantigens entails the alignment of sequencing data, processing of alignment files, variant calling, HLA-typing, neoepitope prediction and selection of immunogenic neoantigens.

Current approaches often rely on whole-exome sequencing (WES) for the detection of somatic variants due to its reliability and relatively low cost. Nevertheless, RNA sequencing (RNA-Seq), while mainly used for gene expression analysis, can also be used for the detection of genetic variants. Therefore, in this master thesis alignment, pre-processing and variant calling were analysed and evaluated to allow the accurate identification of somatic variants from transcriptome analysis. The proposed pipeline involves GSNAP alignment and subsequent pre-processing using GATK's SplitNCigarReads and BaseRecalibrator. For accurate variant calling a new method called MuVaSt, combining variant calling algorithms MuTect2, VarDict and Strelka2, was evaluated. The combination of these three variant callers revealed a higher precision for the detection of SNVs than any single variant caller. Nevertheless, the precision for the detection of indels remained considerably low.

MuVaSt was applied to RNA-Seq data obtained from a formalin-fixed paraffin-embedded (FFPE) sample and a fresh frozen (FF) sample. Subsequently, identified variants were compared to a Gold Standard set containing somatic variants identified using WES data. Only a small overlap was found between somatic variants called in RNA-Seq and in WES. The DNA unique variants were mainly attributed to a low expression. The RNA unique variants, on the other hand, were mainly due to their location outside the WES capture regions. Furthermore, it was observed that a large fraction of the discordant variants had a low variant allele frequency (VAF) potentially caused by tumour heterogeneity and allele-specific expression. In addition, variants with a low read count may also be the result of artefacts originating from sequencing errors, misalignments, library preparation artefacts or sample preservation damage. Finally, comparison of both fresh frozen (FF) and formalin-fixed paraffin-embedded (FFPE) sample types revealed inferior performances for the FFPE sample primarily due to artefacts originating from the formalin fixation process. Moreover, only a limited overlap was observed between FF and FFPE variants potentially caused by the geographical tumour heterogeneity.

To validate the suggested bioinformatics workflow, an FFPE sample from a second patient was analysed. Nevertheless, a low tumour purity and a high fraction of subclonal mutations complicated the detection of somatic variants. Therefore, the bioinformatics pipeline should be optimised in order to allow more precise variant calling of heterogeneous tumour samples.

Keywords: RNA sequencing, Lung cancer, Neoantigen, Alignment, Somatic variant calling, FFPE

SAMENVATTING

Kanker wordt gekenmerkt door een opeenstapeling van genetische veranderingen die leiden tot de ontwikkeling van abnormale cellen die op een chaotische manier delen en zich uiteindelijk verspreiden naar andere delen van het lichaam. Somatische mutaties kunnen aanleiding geven tot tumor-specifieke neoantigenen die als neoepitopen kunnen worden gepresenteerd op kankercellen. De herkenning van deze neoepitopen door cellen van het immuunsysteem kan vervolgens een anti-kanker immuun respons induceren. Om deze reden worden neoepitopen beschouwd als ideale *targets* voor een kanker vaccin. De recente vooruitgang in de *Next Generation Sequencing* technologie en de nieuwe bio-informatica algoritmen die vandaag de dag beschikbaar zijn hebben het mogelijk gemaakt om kanker-specifieke somatische mutaties en de daaruit resulterende neoantigenen te identificeren. Deze nieuwe mogelijkheden dragen bij tot de verdere ontwikkeling van een gepersonaliseerd therapeutisch vaccin. De bio-informatica *pipeline* die wordt gebruikt voor de identificatie van neoantigenen bestaat uit de *alignment* van *sequencing data*, het verwerken van deze *alignment* bestanden, de *variant calling,* HLA-typering, neoepitoop predictie en de selectie van immunogene neoantigenen.

De aanpak die momenteel vaak wordt toegepast om somatische varianten te identificeren bestaat uit *whole-exome sequencing* (WES), vanwege de betrouwbaarheid en relatief beperkte kosten. Hoewel *RNA sequencing* (RNA-Seq) voornamelijk gehanteerd wordt voor de kwantificatie van de genexpressie, is het ook mogelijk om deze RNA-Seq gegevens te gebruiken om genetische varianten te detecteren. Met dit voor ogen werd in deze masterproef de *alignment*, de verwerking en de *variant calling* van het transcriptoom, afkomstig van een long tumor, geanalyseerd en geëvalueerd om een zo accuraat mogelijke identificatie van somatische varianten mogelijk te maken. De pipeline die hiervoor werd geselecteerd bestaat uit *alignment* met GSNAP en verdere verwerking met behulp van GATK's SplitNCigarReads en BaseRecalibrator. Voor de *variant calling* werd een nieuwe methode MuVaSt ontwikkeld, deze combineert de *variant callers* MuTect2, VarDict en Strelka2. De combinatie van deze drie *variant callers* zorgde ervoor dat SNVs met een hogere precisie konden worden gedetecteerd dan wanneer slechts één *variant caller* afzonderlijk werd gebruikt. De precisie voor de detectie van indels bleef daarentegen aanzienlijk laag.

MuVaSt werd toegepast op de RNA-Seq gegevens die werden bekomen van een 'formaline gefixeerd en paraffine ingebed' (FFPE) staal en van een vriescoupe (FF). Vervolgens werden de geïdentificeerde varianten vergeleken met een Gouden Standaard set van somatische varianten die werden gedetecteerd in WES. Slecht een beperkt aantal van deze somatische varianten werd zowel in RNA-Seq als in WES geïdentificeerd. De DNA unieke varianten konden deels worden toegeschreven aan hun beperkte expressie in het tumor weefsel. De RNA unieke varianten konden daarentegen onder andere worden verklaard doordat ze buiten de genomische regio's lagen die worden *getarget* door WES. Daarnaast werd ook opgemerkt dat een groot deel van de unieke varianten een zeer lage allelfrequentie (VAF) had. Het is mogelijk dat dit werd veroorzaakt door de aanwezige heterogeniteit binnen de tumor en door allel specifieke expressie. Varianten met een lage VAF kunnen ook vals positieve varianten zijn afkomstig van artefacten die ontstonden tijdens de *sequencing*, de *alignment,* de *library preparation* of door schade veroorzaakt door de staalname techniek. Ten slotte bleek uit de vergelijking van beide staalname technieken dat het FFPE staal aanleiding gaf tot een lagere precisie. Dit was voornamelijk het gevolg van artefacten afkomstig van het formaline fixatieproces. Daarnaast was de overlap tussen de varianten geïdentificeerd in het FF staal en het FFPE staal beperkt door onder andere de geografische heterogeniteit binnen het tumorweefsel.

Om deze bio-informatica pipeline te valideren werd een FFPE staal van een tweede patiënt geanalyseerd. De accurate detectie van somatische varianten werd hier echter bemoeilijkt doordat dit staal slechts een beperkt percentage tumor cellen bevatte en er bovendien meer subklonale mutaties aanwezig waren. Bijgevolg vergt deze pipeline nog verder onderzoek om ervoor te zorgen dat de *variant calling* van heterogene tumor stalen preciezer kan worden uitgevoerd.

Trefwoorden: RNA sequencing, Longkanker, Neoantigenen, Alignment, Somatic variant calling, FFPE

INTRODUCTION

Lung cancer remains the most prevalent cancer worldwide, both in term of new cases (1.8 million cases, 12.9% of total) and deaths (1.6 million deaths, 19.4% of total) because of the high mortality rate [1]. As a genetic disease, somatic mutations accumulate in cancer cells during cancer progression. These mutations can alter protein functions, ultimately disrupting cellular control of pathways, resulting in the hallmarks of cancer [2].

Conventional chemotherapies are not considered to be the ultimate therapy since not only tumours are targeted but also dividing cells in healthy organs. For more than a century, immunotherapy has been postulated as a promising alternative since it focusses on deploying a patient's own immune system to specifically target malignant cancer cells. A possible approach is to target immune checkpoints, these are molecules that modulate the immune system. For example, Programmed Death 1 (PD-1) and cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) are immune checkpoint receptors that can be targeted using specific antibodies, so-called immune checkpoint blockers. When an antibody binds the receptor or its ligands, the cancer cell is no longer able to suppress the immune response. However, only a limited number of patients experience complete treatment response, and even if they do so, resistance is often acquired. Owing to these limitations of current immunotherapies, there is considerable scope for novel therapies. One promising new approach is a personalised cancer immunotherapy that targets patient-specific tumour antigens. In recent years, technological advances such as next-generation sequencing (NGS) and machine-learning approaches for epitope prediction have paved the way for the development of these personalised anti-tumour therapies directly targeting a patient's somatic mutations, so-called neoantigens. Furthermore, it was already widely recognised that somatic mutations within a cancer cell give rise to neoepitopes that are recognised by the adaptive immune system as 'mutated-self' and in this way differentiate cancer cells from normal cells [3]. However, tumours develop immune evasion strategies in order to avoid the recognition and degradation by the immune system. Combining computational advances and knowledge on the interactions between the immune system and cancer cells allows the development of a personalised therapeutic cancer vaccine that specifically targets neoepitopes on cancer cells and in this way activates and enriches the immune response. Nevertheless, a standardized protocol for the detection, selection and targeting of these neoantigens hasn't been adapted yet since this is a complex and computationally demanding process that requires specialised bioinformatics tools and optimised selection criteria.

This master thesis is part of a research project funded by *'Kom op tegen Kanker'* that aims to improve the outcome of immune checkpoint blockade in lung cancer by combination with a tumour mutanome-targeted dendritic cell vaccine. The focus of this thesis is to optimise the alignment and the variant calling on RNA sequencing (RNA-Seq) data in order to be able to detect and select the most promising neoantigens in further steps of the bioinformatics pipeline. The therapeutic pipeline that will be evaluated has some features not adapted in previously described pipelines. First of all, the ultimate goal is to use formalin-fixed paraffin-embedded (FFPE) samples to extract genomic material. This is challenging because accurate detection of mutations is often problematic in FFPE material due to fragmentation and modification of DNA and RNA. Second, the transcriptome will be sequenced and analysed to detect somatic variants. RNA-Seq has some major advantages over whole-exome sequencing (WES) but at the same time brings some computational challenges. Third, both single-nucleotide variants (SNV) and insertions and deletions (indels) will be targeted. Nevertheless, alignment and detection of indels is challenging and thus requires careful selection and optimisation of bioinformatics tools.

1 LITERATURE STUDY

1.1 Cancer

1.1.1 Mutational Process

Cancer is a disease characterised by the development of abnormal cells that divide uncontrollably and have the potential to invade or spread to other parts of the body. The progression of cancer comprises three major phenotypes: growth, invasion and metastasis. When a normal somatic cell acquires resistance to apoptosis and undergoes uncontrolled cell proliferation the number of neoplastic cancer cells, and hence the tumour tissue mass, will increase which results in a primary, benign tumour. Following this growth phase, the cancer cells will start to invade healthy tissue. Eventually, some cancer cells will enter the blood circulation and cellular dissemination will result in the formation of a secondary metastatic tumour (carcinoma).

Genetic and epigenetic alterations caused by both exogenous carcinogens (e.g. radiation, chemical carcinogens) and endogenous carcinogens (e.g. tumour promoting inflammation) underlie this cancer progression. Mutated genes contributing to tumour development are called driver genes and can be classified into oncogenes and tumour suppressor genes. A normal cellular gene that codes for a protein involved in normal cell division, invasion and/or migration can be overexpressed or mutated into an oncogene which, in its turn, contributes to the formation of a cancer cell [4]. An example of an oncogene is *EGFR* (Epidermal Growth Factor Receptor) which in its oncogenic form results in an elevated responsiveness to growth factors and a reduced sensitivity to apoptosis and malignant cell growth. A tumour suppressor gene, on the other hand, is a gene normally acting to inhibit cell proliferation and tumour development. In many tumours, these genes are lost or inactivated, thereby removing negative regulators of cell proliferation and contributing to cancer progression [4]. For example, *pRb* (protein retinoblastoma) is a tumour suppressor gene that is mutated in several major cancers. One key function of pRb is to prevent excessive cell growth by actively inhibiting cell cycle progression into the S phase.

Mutations do not occur at predefined positions in a genome but they rather happen at random. This means that not all mutations in a cancer cell drive cancer progression. So-called passenger mutations do not confer a selective growth advantage for the cancer cell in contrast to driver mutations. As a result, driver mutations can be similar between different patients, conversely, passenger mutations are all different. It is important to note that there is a fundamental difference between a driver gene and a driver mutation. A driver gene is a gene that contains driver mutations, and in this way contributes to cancer progression, but on the other hand, this driver gene may also contain passenger mutations. Although it is relatively straightforward to define a driver mutation as one conferring a selective growth advantage, it is more difficult to identify which somatic mutations are drivers and which are passengers [5].

The presence of multiple genetic alterations in cancer cells strongly indicates that those alterations accumulate in the cells in a stepwise manner during tumour progression. Analyses of genetic alterations in different tumours have shown that the number of genetic alterations in late-stage tumours is usually higher than those in early-stage tumours in various types of cancers [6]. This concept of multistage carcinogenesis explains why not all cancer cells in a tumour have a similar mutational profile. At first, normal cells carefully control the production and release of growth-promoting signals regulating the cell growth-and-division cycle, thereby ensuring a homeostasis of cell number and thus preservation of normal tissue morphology and function. When a normal cell acquires a driver mutation in a gene regulating this process it will sustain proliferation enabling replicative immortality, the most fundamental trait of cancer cells [2]. This first gatekeeping mutation provides a selective growth advantage to a normal cell, allowing it to outgrow the surrounding cells [5]. The small adenoma that results from this mutation slowly grows and the cancer cells continue to proliferate which eventually results in the accumulation of both driver and passenger mutations. This mutational process followed by clonal expansion continues, thereby dividing the tumour tissue into different tumour populations with different mutational profiles. Mutations that are present in the majority of the neoplastic cells in the tumour are called clonal. Subclonal mutations, on the other hand, can only be found in a small subpopulation of tumour cells [5]. Beside replicative immortality, seven other biological capabilities are acquired during the multistep development of a metastatic tumour: the hallmarks of cancer. These include sustained proliferative signalling, evading growth suppressors, resisting cell death, inducing angiogenesis, activating invasion and metastasis, evading immune destruction and reprogramming of energy metabolism [2]. Eventually, this mutational process results in malignant cancer cells that can invade through the underlying basement membrane and ultimately metastasize to lymph nodes and other parts of the body [7].

Over the past decade, comprehensive sequencing efforts have facilitated the research of mutational processes in tumours. The cost of high-throughput sequencing has been reduced to a $1000 per genome [8], thereby enabling the production of a massive amount of sequencing data that can help to unravel the genomic landscape of common types of human cancer, the so-called mutanome [5]. It has been shown that in a solid tumour on average 33 to 66 genes have acquired somatic mutations that would be expected to alter their protein products, so-called nonsynonymous mutations [5]. However, lung tumours together with melanoma tumours have an exceptionally high prevalence of somatic mutations as can be seen in Figure 1 [9]. Since the detection of somatic mutations is computationally challenging, a higher mutational frequency results in a higher chance of detection. As a result, lung tumours are a prime candidate to evaluate somatic mutations. About 95% of the nonsynonymous mutations are SNVs, whereas the remainders are indels of one or a few bases. On average 90.7% of these nonsynonymous SNVs result in missense mutations, 7.6% result in nonsense mutations, and 1.7% result in alterations of splice sites or untranslated regions [5]. Other types of mutations found in solid tumours are changes in chromosome number (aneuploidy), gene amplifications, and translocations that may result in the fusion of two genes thereby creating an oncogene. As already described before, the majority (>99.9%) of these genetic alterations appear to be passenger mutations rather than driver mutations [5].

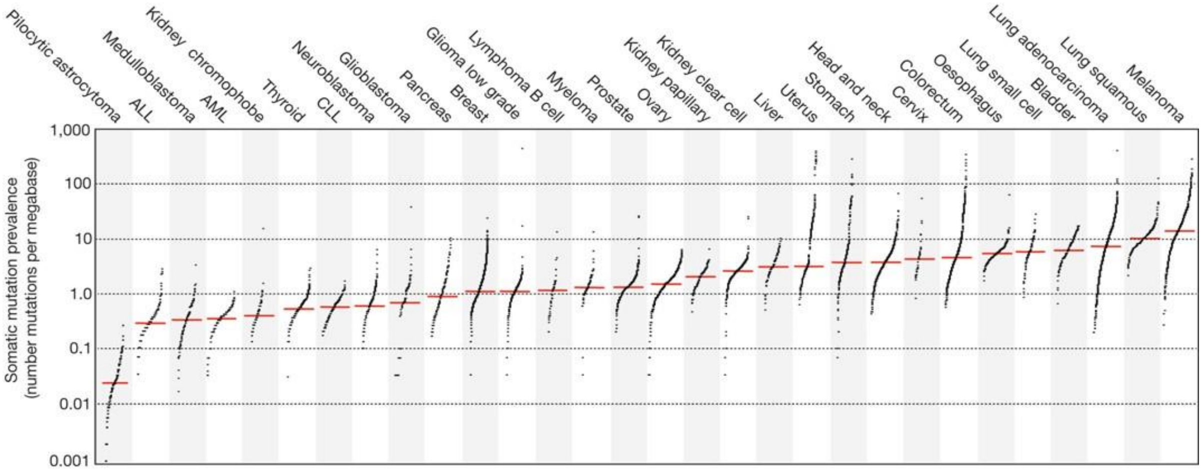


Figure 1. The prevalence of somatic mutations in human cancer types [9]*. Every black dot represents a sample and the red horizontal lines indicate the median number of somatic mutations per megabase in the respective cancer types. The vertical axis is log scaled and depicts the number of somatic mutations per megabase.*

1.1.2 Lung Cancer

Lung cancer accounts worldwide for more deaths than any other cancer type in both men and women. In 2012, 12.9% of new cancer cases and 19.4% of cancer deaths were caused by lung cancer [1]. Lung cancer can be categorized into two main histological groups [10]: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). NSCLCs comprises the largest group (85% of all lung cancers) and can be subcategorized into lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and large cell carcinoma. Squamous cell carcinomas generally develop in the central bronchi, while adenocarcinomas often emerge more at the edges of the lungs [11]. Although NSCLCs are associated with cigarette smoke, LUAD may occur in patients who have never smoked. LUSC, on the other hand, appears most often in patients who have smoked. In contrast to SCLC, NSCLCs are relatively insensitive to standard chemotherapy and radiation therapy.

Lung adenocarcinoma accounts for more than 50% of all cases in lung cancer and arises from the glandular cells located in the epithelium of the bronchi. The comprehensive molecular profiling of 230 adenocarcinomas by The Cancer Genome Atlas (TCGA) was published in 2014 [12]. Whole-exome sequencing of tumour and germline DNA revealed a mutation rate of 8.87 somatic mutations per megabase of DNA, the nonsynonymous mutation rate was 6.86 per Mb. Consistent with previous studies [13], a significantly higher exonic mutation rate was observed in tumours from smokers. The authors also identified 18 statistically significant mutated genes: *TP53* (46%), *KRAS* (33%), *KEAP1* (17%), *STK11* (17%), *EGFR* (14%), *NF1* (11%), *BRAF* (10%), *SETD2* (9%), *RBM10* (8%), *MGA* (8%), *MET* (7%), *ARID1A* (7%), *PIK3CA* (7%), *SMARCA4* (6%), *RB1* (4%), *CDKN2A* (4%), *U2AF1* (3%), and *RIT1* (2%) [14]. The percentages reflect the proportion of LUAD containing a mutation in the aforementioned gene. These genetic mutations were also found in previous studies [13], [15]. The receptor tyrosine kinase (RTK)/RAS/RAF pathway is frequently mutated in LUAD [12]. Most important genetic alternations that promote this pathway are *KRAS, EGFR* and *BRAF* mutations. In addition to these alterations, also MET exon 14 skipping, ERBB2 (or HER2) mutation and/or amplification, *ROS1* fusion, *ALK* fusion, *MAP2KA* mutation, *RET* fusion, *NRAS* mutation, *HRAS* mutation, *MET* amplification, *NF1* mutation and *RIT1* mutation promote the (RTK)/RAS/RAF pathway [14]. All these genetic alterations are identified as driver mutations since they promote the (RTK)/RAS/RAF pathway which may lead to increased or uncontrolled cell proliferation and resistance to apoptosis. This pathway plays a key role in oncogenesis since 76% of LUAD driver mutations can be identified within the (RTK)/RAS/RAF pathway [12].

Lung squamous cell carcinoma is defined as a lung carcinoma that begins in the squamous cells (the flat cells lining the inside of the airways in the lungs) and is more strongly associated with smoking than any other type of NSCLC. The spectrum of mutations in LUSC is very different from LUAD, which explains why LUSC has not been responsive to drugs that work for LUAD. For example, mutation of *EGFR* and *KRAS*, the two most abundant oncogenic alterations in LUAD, are extremely rare in LUSC. LUSC is characterised by complex genomic alternations, as can be expected from the history of heavy smoking in LUSC patients [9]. A recent comprehensive molecular profiling by TCGA identified 11 statistically significant genetic mutations: *TP53, CDKN2A, PTEN, PIK3CA, KEAP1, MLL2, HLA-A, NFE2L2, NOTCH1, RB1* and *PDYN* [14], [16]. Mutation in *TP53* has been identified in 90% of the cases. The authors also identified novel loss-of-function mutations in the *HLA-A* class I major histocompatibility gene which can be linked to the cancer hallmark of avoiding immune destruction [14]. Frequent alterations were also identified in the following pathways: *CDKN2A/RB1, NFE2L2/KEAP1/ CUL3, PI3K/AKT* and *SOX2/TP63/NOTCH1* pathways, providing evidence of common dysfunction in cell cycle control, response to oxidative stress and apoptotic signalling [9].

The most frequent mutational signatures found in lung cancer tumours are C<T transitions and C<A transversions [9], [13]. This mutational signature is associated with tobacco smoking and is probably an imprint of the bulky DNA adducts generated by polycyclic hydrocarbons originating from tobacco smoke and their subsequent removal by transcription-coupled nucleotide excision repair [16]. As a result, driver genetic alterations in LUAD differ between smokers and non-smokers. For example, mutations in the *KRAS* gene are frequently detected in LUAD in smokers, while oncogenic aberrations in *EGFR*, *ALK*, *ROS1* and *RET* are more abundant in tumours of non-smokers [14].

1.1.3 Cancer Immunology

Oncogenes not only contribute to tumour development and progression, but they can also be the source of tumour-associated antigens (TAA). Several types of TAA exist: differentiation antigens (e.g. melanocyte differentiation antigens), mutational antigens (e.g. p53), overexpressed cellular antigens (e.g. HER2), viral antigens (e.g. human papillomavirus proteins) and cancer/testis antigens that are expressed in germ cells of testis and ovary but silent in normal somatic cells (e.g. MAGE and NY-ESO-1) [17]. TAA are relatively restricted to tumour cells, and, to a limited degree, to normal tissues (differentiation antigens). Tumour-specific antigens (TSA), on the other hand, are solely expressed in tumour cells. These TSA, also called neoantigens, can be presented as neoepitopes on malignant cancer cells. Recognition of neoepitopes by immune cells may induce an anticancer immune response [18].

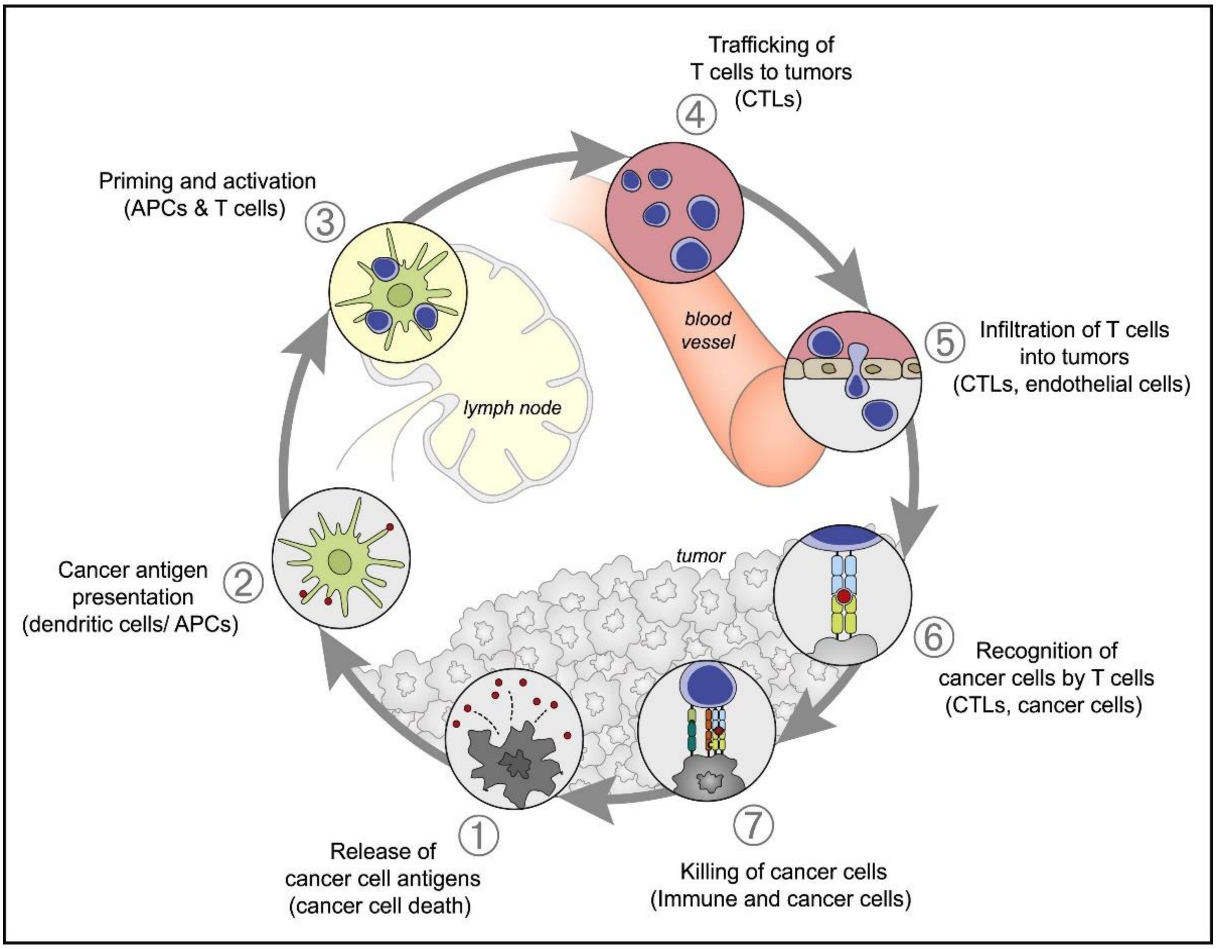


Figure 2. The Cancer-Immunity Cycle. This cycle can be divided into seven major steps, starting with the release of cancer cell antigens and ending with the elimination of respective cancer cells [18].

The mutational landscape and the resulting TSA described above imply that there is a clear opportunity for the immune system to differentiate tumour cells from healthy tissue. However, in order for an immune response to induce the effective elimination of cancer cells, a series of successive events must be initiated and expanded iteratively, the so-called Cancer-Immunity Cycle [18] (Figure 2). In a first step, TSAs are released, for example, by cancer cell death. Dendritic cells (DCs) capture and process these neoantigens thereby presenting them in association with major histocompatibility complex (MHC) class I or class II molecules on their cell surface. The DCs migrate to the lymph nodes were T lymphocytes, also called T cells, recognise the presented TSA. This results in the priming and activation of CD8$^+$ and/or CD4$^+$ T cells. These T cells can now recognise the specific TSA and induce an immune response. The activated CD8$^+$ T cell, also called the cytotoxic T cell (CTL), migrates and infiltrates the tumour tissue. Next, the CTL binds a cancer cell through interaction between its T-cell receptor (TCR) and its cognate neoantigen presented on an MHC I molecule on the cancer cell surface. Finally, perforins and granzymes are released by the CTL and the cancer cell is killed which results in the release of additional TSAs. Each step of this Cancer-Immunity Cycle is controlled by both stimulatory and inhibitory factors such as interleukins, chemokines and growth factors.

MHC class I and class II molecules play a key role in the Cancer-Immunity Cycle. These two classes of MHC molecules differ in structure, associated peptides and the type of activated T cells. MHC I molecules are present on all nucleated cells in the body and have a heterodimeric structure that consists of one membrane-spanning α chain (heavy chain) and one β chain (light chain). The α chain consists of 3 domains: α1-microglobulin, α2-microglobulin and α3-microglobulin. The latter binds non-covalently with β2-microglobulin. The α chain is encoded by three genes: HLA-A, HLA-B and HLA-C in humans [19]. These are highly polymorphic genes which explains the unique character of each specific MHC I molecule. Polymorphisms of the MHC proteins result in different peptide-binding grooves that recognise unique peptides due to variations in the anchor residues to which peptides dock [20]. MHC class I molecules generally process and present endogenous antigens through a specific pathway (Figure 3a) [21]. First, an endogenous protein is cut into small peptides by the proteasome in the cytosol. Next, these peptides are transported into the endoplasmic reticulum (ER) by the TAP1/TAP2 transporter where they are trimmed to appropriate length by ER aminopeptidases. With the help of chaperone proteins, such as tapasin, these trimmed peptides, generally 8 to 9 amino acids long, are loaded onto the peptide-binding groove (α1- and α2-microglobulin) of the MHC I molecule. Finally, the peptide-MHC I complex is transported to the cell surface where it can bind with a TCR of a CD8$^+$ cell [20].

MHC class II molecules are only present on specific antigen-presenting cells (APC) like DCs, B cells and macrophages. MHC II molecules consist of two membrane-spanning chains, α and β, both produced by three polymorphic genes: HLA-DR, HLA-DQ and HLA-DP in humans [19]. The α1 and β1 microglobulins come together to make a membrane-distal peptide-binding groove. In contrast to MHC I molecules, this groove is open and therefore allows to bind longer peptides, generally between 15 and 24 amino acid residues. Since MHC II molecules present exogenous peptides, the processing and presenting pathway is different from the MHC I pathway (Figure 3b) [21]. Exogenous proteins are captured in an endosome where they are degraded into smaller peptides. Subsequently, the endosome fuses with a vesicle containing MHC II molecules and the class II-associated invariant chain peptide (CLIP) region that blocks the peptide-binding groove is replaced by an exogenous peptide. Finally, the peptide-MHC II complex is transported to the APC's cell surface where it can bind with the TCR of CD4$^+$ T cells [20].

This T cell reactivity against cancer cells has been widely studied in the past decades [3], [22]–[25]. Accumulating evidence suggests that the T-cell-based immune system reacts to both MHC class I-restricted and MHC class II-restricted neoantigens [7], [24]. Nevertheless, only a small fraction of the nonsynonymous mutations leads to the production of neoantigens that can be detected by CD4$^+$ or CD8$^+$ T cells. For this reason, tumours with a high mutational load (e.g. melanoma and lung adenocarcinoma) are more likely to provoke T cell reactivity against neoantigens [26], [27].

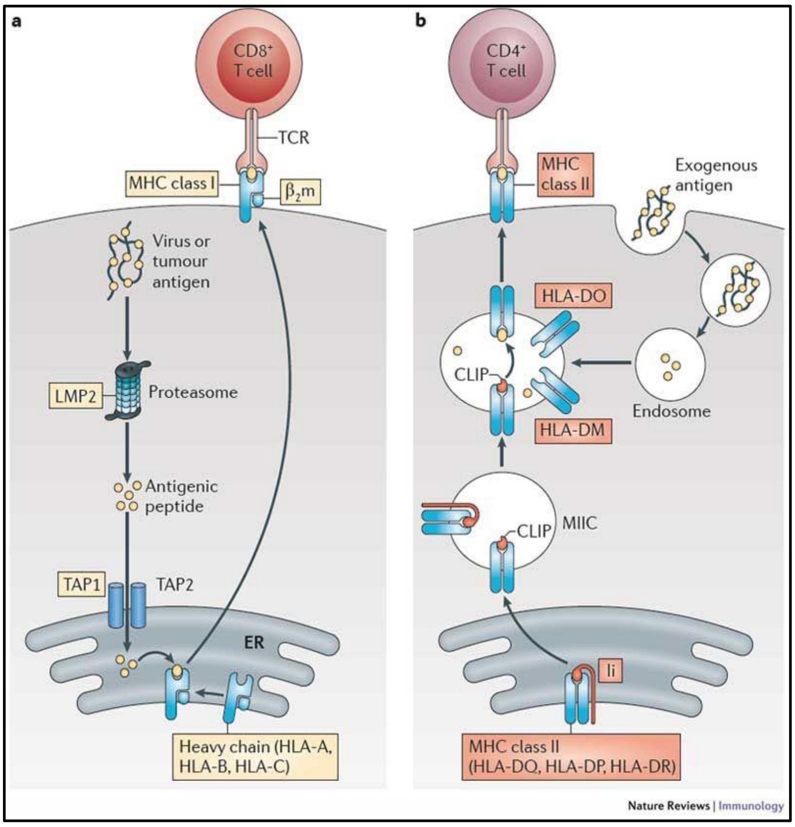


Figure 3. (a) MHC class I antigen-presentation pathway. Intracellular antigens, such as virus or tumour antigens, are processed into peptides by the immunoproteasome. The subsequent peptides are transported into the ER, where they are loaded into the groove of the MHC class I complex. MHC class I complexes present antigens on the cell surface to CD8$^+$ T cells. (b) MHC class II antigen-presentation pathway. Exogenous antigens, e.g. bacterial antigens, are cleaved by endolysosomal enzymes into peptides. These peptides bind to the groove of the MHC class II complex by displacing the CLIP. The resulting MHC class II complex presents the antigens to CD4$^+$ T cells [21].

Still, in most cancer patients this Cancer-Immunity Cycle no longer performs optimally. The selective pressure of our immune system results in the acquisition of cancer cells that have developed immune evasion strategies. Thus, the immune system not only protects against tumour formation but at the same time also shapes tumour immunogenicity. This so-called immunoediting hypothesis emphasises the dual host-protective and tumour-promoting actions of the immune system on developing tumours [17], [28]. Three distinct phases that proceed sequentially can be distinguished in this cancer immunoediting process: elimination, equilibrium and escape [17].

The elimination phase is characterised by cancer immunosurveillance, in which both innate and adaptive immune systems cooperate in order to detect the presence of a developing tumour long before they become clinically apparent [17]. If a cancer cell variant is not destroyed in this phase, it may enter the equilibrium phase, in which the immune system prevents further tumour outgrowth and also shapes the immunogenicity of the respective tumour cells. This is the longest phase of the immunoediting process since tumour cells are maintained in a functional state of dormancy in which only the adaptive immune system, particularly interleukin-12 (IL-12) and interferon-gamma (IFN-ɣ) producing CD4$^+$ and CD8$^+$ T cells, prevents tumour outgrowth. However, as a consequence of constant immune selection pressure placed on genetically unstable tumour cells in equilibrium, tumour cell variants may escape immunosurveillance through different mechanisms. Alterations in tumour cells can lead to reduced immune recognition (e.g. due to loss of antigen presentation) or to increased resistance to the cytotoxic effect of immunity (e.g. due to induction of anti-apoptotic mechanisms) [17]. The most important and best-studied escape mechanism is the loss of TSA presentation which can occur in three different ways: through emergence of tumour cells lacking expression of strong TSA, through loss of MHC I expression on the tumour cell surface, or through loss of antigen processing function that is required for the production of the epitope and the subsequent loading onto the MHC molecule [17]. These alterations are probably caused by a combination of the genetic instability of the cancer genomes and the T cell-dependent process of immunoselection [29]. This Darwinian selection process results in poorly immunogenic tumour cells that escape the immune response and acquire the ability to grow progressively [17]. Another known immune evasion mechanism is the establishment of an immunosuppressive state within the tumour microenvironment. This tumour microenvironment is formed during tumour progression and consists of a number of different cell types that support immunosuppression, e.g. regulatory T cells (Tregs) and myeloid-derived suppressor cells (MDSCs) [30], [31]. These MDSCs can suppress T cell activity employing different suppressive mechanisms and in certain circumstances differentiate into highly immunosuppressive macrophages.

1.2 Detection and Evaluation of Neoantigens

1.2.1 Personalised Therapeutic Pipeline

The previous sections explained the mutational process in cancer cells and how this results in cancer progression and the production of TSA that can be targeted by the immune system. However, as already described before, cancer immunoediting allows tumours to escape the immune response.

In order to overcome cancer immune escape several new therapeutic strategies are being developed to expand and broaden the T cell responses against cancer cells, e.g. adoptive cell transfer (ACT), immune checkpoint inhibition (ICI) and CAR-T cell therapy. Another possible strategy is to (re)activate and expand the antitumour immune response using cancer-specific neoepitopes to target the immune system to the malignant cancer cells. In order to do so, potential neoantigens are selected solely based on genomic information and RNA expression data followed by the prediction of subsequent neoepitopes. These detected neoepitopes are used to prime T cells to specifically recognise them and in this way enhance the immune response against cancer cells presenting these specific neoepitopes on their cell surface. Since a large fraction of the genetic alterations in human tumours is not shared between patients, this therapeutic approach will be patient-specific: personalised cancer immunotherapy. Several studies have already provided evidence that such a cancer mutanome-based approach can be used to identify neoantigens that can be recognised by T cells [7], [32]–[38]. For example, Kreiter *et al.* [38] evaluated this approach and provided evidence that a considerable fraction of nonsynonymous mutations is immunogenic and that vaccination with both CD8$^+$ and CD4$^+$ immunogenic mutations confers strong antitumour activity. Moreover, the clinical feasibility of this approach has already been assessed in melanoma first-in-human clinical trials [35], [36], [39]. These studies confirmed a high overall immunogenicity rate of 60% against individual mutations. Each patient developed strong T cell responses against several of their tumour-specific mutations. Pre-existing T cells were expanded and more important the majority of vaccine-induced T cell responses were newly primed and not detected prior to vaccination [7].

Nonetheless, no standardized protocol has been instated. As a result, different strategies and different bioinformatics tools have been used to detect mutations in cancer cells that will result in neoantigens. Most common pipelines focus on the prediction and selection of SNVs based on the comparison of WES data of tumour and normal samples and expression data of RNA-Seq [7], [35], [36], [39], [40]. However, since RNA-Seq data also contains information on variants, it would be possible to use only RNA-Seq data to call somatic variants. Besides, it was also proven that indels causing a frameshift mutation are highly immunogenic and thus induce a strong immune response [41]. Therefore, the goal of this master thesis is to evaluate the alignment and variant calling process of a therapeutic pipeline that compares RNA-Seq data of tumour tissue and WES data of healthy tissue in order to detect both somatic SNVs and indels (Figure 4) in the cancer genome. In the following sections, this personalised therapeutic pipeline will be described starting from a lung biopsy and finally resulting in the evaluation of neoepitopes and the production of a dendritic cell vaccine.

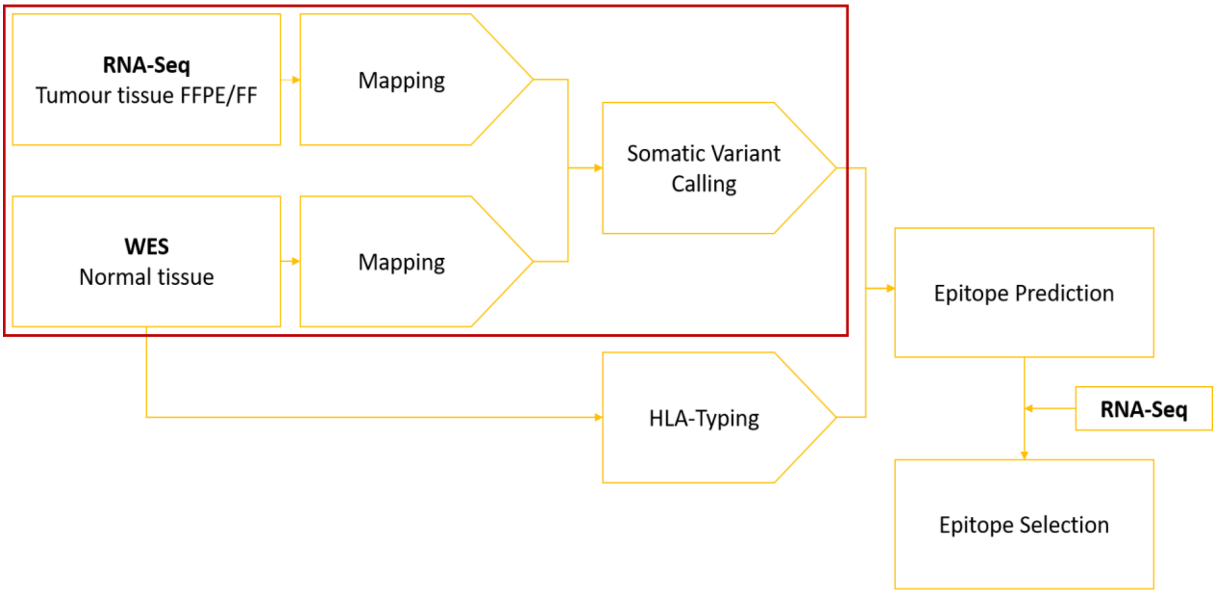


Figure 4. Personalised therapeutic pipeline for the detection and evaluation of neoepitopes. The red frame represents the focus of this master thesis.

1.2.2 Biological Material

The first step in this personalised therapeutic pipeline is the collection of a lung tumour biopsy. This biopsy tissue needs to be specially prepared for long-term preservation. Two major types of sample preparation exist: FFPE samples and fresh frozen (FF) samples. For FF samples, biopsied tissue is immediately dipped in liquid nitrogen ("flash freezing") and stored in a freezer at less than -80 degrees Celsius [42]. One major drawback of FF samples is the fact that enormous efforts and financial support are required to keep FF samples stable for a longer period of time. For the preparation of FFPE samples, tumour tissue is first fixed in formaldehyde, also known as formalin, in order to preserve tissue morphology and proteins. In a second step, the FFPE tissue is embedded in a paraffin wax block that allows easy cutting of slices for microscopic examination [43]. Once prepared, FFPE tissue is very hardy and does not require special equipment to preserve for decades [42]. FFPE samples are most commonly used in a clinical setting because pathologists are accustomed to examining FFPE tissue biopsies. For this reason, FFPE is mostly considered as a standard preservation technique. However, FFPE samples can be difficult to process in many molecular biology assays because the fixation process and subsequent tissue storage often cause nucleic acid degradation, resulting in fragmented DNA and RNA transcripts [44]. Therefore, FFPE samples are generally inferior for genetic analysis, such as polymerase chain reaction (PCR), quantitative PCR (qPCR), or NGS [45]–[50]. FF samples, however, preserve the DNA, RNA, and native proteins [42]. As a result, most of the studies of transcriptome analyses have traditionally been conducted using fresh or FF tumour samples. To date, most clinical samples collected are not fresh or FF but FFPE and often only limited amounts of material are available. Therefore, specialised extraction kits have been developed to extract and capture RNA from low-quality FFPE samples.

Using FFPE samples for NGS may be challenging due to the fixation process, and the storage time and conditions [45]–[50]. On molecular basis, the common issues encountered when using FFPE tissue are nucleic acid fragmentation and modification by chemical reactions between formaldehyde and nucleic acids, including crosslinking with proteins and other biomolecules [47]. Nucleic acid degradation and modification can significantly reduce the quantity and quality of DNA and RNA extracted from FFPE samples which can lead to sequencing artefacts ultimately resulting in an increased risk of false positive mutation calls. Besides, nucleic acid fragmentation can reduce the library fragment size and uniformity [51]. In addition, the fragmentation and modification of RNA transcripts may affect the poly(A) tail of mRNA molecules which make them no longer suitable for traditional RNA-Seq library preparation protocols [47], [51]. Several studies have been conducted to assess the variability of RNA between FFPE and FF samples [48], [49]. These studies concluded that, in contrast to gene expression analysis, the identification of somatic mutations in FFPE samples can be challenging due to artefacts appearing in RNA material. Graw *et al.* [49] identified mutational artefacts G>A and C>T specific to FFPE samples. Since these artefacts only occurred at low variant allele frequencies (VAF), they applied a VAF filter to remove them. Contrarily, Esteve-Codina *et al.* [48] found some of these artefacts at very high variant allele frequencies. It can be concluded that the use of FFPE samples brings additional challenges in the detection of somatic mutations compared to FF samples. One should be aware that the quality of RNA extracted from FFPE tissue, and as a result the possibility to correctly identify somatic mutations, is mainly dependent on the fixation process, storage time and conditions, and the conditions of the paraffin blocks [50].

In order to overcome these challenges, a robust DNA and RNA extraction protocol is required for accurate identification and quantification of somatic mutations in FFPE tissue. Some commonly used extraction protocols are the AllPrep DNA/RNA FFPE Kit and the RNeasy FFPE Kit (QIAGEN, Inc., Hilden, Germany) [48], [49]. These protocols use a Deparaffinization Solution and Proteinase K for the digestion of cross-linked proteins for the purification of nucleic acids. However, other protocols like the truXTRAC FFPE Kit from Covaris and FormaPure Kit from Beckman Coulter have shown to perform equally good or even better than the QIAGEN kits [52].

1.2.3 RNA Sequencing

The development of NGS technology has facilitated the comprehensive analysis of the full genome and transcriptome of human tumours. Moreover, a comparison of the genomic data of cancer tissue with genomic data of healthy tissue from the same patient can be used to reveal the full range of somatic mutations within a tumour [53]. Nowadays different sequencing techniques are available, each with its own advantages and disadvantages. Currently, a frequently used strategy to identify neoantigen candidates is based on a comparative analysis of WES data from both tumour and normal tissue combined with gene expression analysis by RNA-Seq [7], [35], [36], [39], [40]. However, the goal of this master thesis is to detect neoantigens solely based on somatic variant calling using RNA-Seq data of tumour tissue and WES data of healthy tissue [32]–[34]. As a result, WES of tumour tissue is no longer necessary. The feasibility of this approach has already been assessed in several studies [32]–[34], [54]–[56].

WES targets the exome, this is the protein-coding region of the human genome which represents less than 2% of the genome but contains about 85% of known disease-related variants making WES a cost-effective alternative to whole-genome sequencing (WGS) [57]. Besides, WES achieves a higher sequence coverage than WGS since only the exonic region of the human genome is sequenced. WES of tumour tissue can be examined to detect somatic variants. Nevertheless, for our pipeline only WES of healthy tissue is performed in order to detect germline variants thereby preventing the incorrect classification of these germline variants as somatic variants. For the tumour tissue, on the other hand, RNA-Seq is employed.

RNA sequencing is a highly sensitive and accurate sequencing technique for measuring gene expression across the transcriptome. This expression data can be used to analyse the expression of the genes containing somatic variants. Moreover, RNA-Seq data also provides sequence information. Therefore, it is possible to use RNA-Seq data both for expression analysis and for the detection of somatic variants. This technique has the advantage over WES that it is not limited to known genes. Besides, RNA-Seq has the potential to detect novel transcripts, gene fusions, SNVs, indels, alternative splicing events and other features without the limitation of required prior knowledge [32], [33], [54], [58]. For this reason, RNA-Seq is considered a powerful method for detecting mutations in cancer transcriptomes that would be missed by WES.

Nevertheless, employing RNA-Seq for the identification of somatic variants remains a challenge because of the transcriptome's intrinsic complexity (e.g. splicing), which leads to the technical difficulty of the computational analysis [32]. Besides, it is difficult to identify germline variants if only RNA-Seq is used because RNA expression profiles of tumour and normal tissue will not be identical. This makes it challenging to distinguish tumour-specific somatic mutations from germline variants [59]. When including germline variants called from WES data of healthy tissue this problem can be avoided. Another concern is the ability to reliably call somatic mutations within RNA-Seq that are only present at a low-level, either because of low-level gene expression, allele-specific expression or because of low mRNA stability (for instance due to non-sense-mediated RNA decay) will be limited [53]. Another important drawback is the library preparation which not only introduces some technical variation in RNA-Seq data but also is responsible for PCR duplicates [60]. During PCR amplification fragments are amplified in order to obtain sufficient reads to load onto the sequencer. However, it is possible that one fragment generates multiple exact copies and, in this way, introduces PCR bias, a common phenomenon in RNA-Seq. Furthermore, sequencing can introduce another type of duplicates: optical duplicates. These optical duplicates arise when during Illumina® sequencing one read cluster is falsely considered as two clusters. Duplicate reads are identified as sequence reads that align to the same genomic coordinates using reference-based alignment. However, it is not possible to differentiate between duplicates and fragments that overlap precisely in highly expressed genes. For this reason, it is not recommended to remove duplicate reads for gene expression analysis of RNA-Seq data as this will underestimate the abundance of highly expressed transcripts. On the other hand, it is recommended to mark these duplicates for appropriate detection of somatic mutations in order to avoid false positive called variants [60]. Another biological factor to consider is RNA editing. RNA editing is a process through which the nucleotide sequence specified in the genomic DNA is modified to produce a different nucleotide sequence in the transcript [61]. In humans, the most prevalent type of RNA editing converts adenosine (A) residues into inosine (I) in double-stranded RNAs through the deamination reaction conducted by members of the adenosine deaminase (ADAR) family of enzymes [9], [62]. Subsequently, inosine can be interpreted as guanosine (G) by the cellular machinery. This co- and post-translational process results in an A:T→G:C mutational signature in RNA. Another but less frequent appearing type of RNA editing is cytosine (C) to uracil (U) editing mediated by the APOBEC deaminase enzyme family [9].

An important factor to consider when identifying somatic variants is the fact that tumour tissue does not consist of identical copies of cancer cells, a phenomenon called tumour heterogeneity. Since the tumour sample will also contain a portion of healthy cells, it is more difficult to identify present mutations. Of course, the higher the healthy tissue content, the harder it is to detect somatic variants. For this reason, an appropriate depth of coverage should be attempted to ensure a sufficient representation of tumour-derived sequence reads. Another factor contributing to the tumour heterogeneity is the existence of both clonal and subclonal mutations. As already described before, clonal mutations can be found in the majority of cancer cells in the tumour tissue since these mutations were part of the original set of mutations present when a cell transformed into a neoplastic cancer cell. Subclonal mutations were acquired by daughter cancer cells during tumour growth [5]. Similarly, a higher coverage enables to distinguish between clonal and subclonal mutations. Nevertheless, this genetic variability is not the only dimension of tumour heterogeneity. The complex interplay between the tumour and a set of host and environmental factors which define the immunological status, shape each individual cancer. For example, HLA haplotype, the microbiome, age, comorbidity, the immune cell repertoire and the composition of the tumour microenvironment are important factors that contribute to tumour progression [7].

Both WES and RNA-Seq are performed using the same sequencing platform provided by Illumina. Illumina is currently the leader in the NGS industry and most library preparation protocols are compatible with this system. In addition, Illumina offers the highest throughput of all platforms, the lowest per-base cost and a low sequencing error rate [57], [63]. Both RNA-Seq and WES will be performed using Illumina® Hiseq 3000 technology. A fundamental step in each NGS workflow is the conversion of nucleic acid fragments into a sequencing library. As already described before, the extraction process of DNA or RNA depends on the type of biological material that is used. Next, a specific NGS library preparation protocol for DNA or RNA will be performed. A wide variety of NGS library preparation protocols exist, but they all have in common that DNA and/or RNA molecules are fused with platform-specific adapters [57]. The NGS library preparation of RNA differs somewhat from the library preparation of DNA since conversion from single-stranded RNA to double-stranded cDNA using reverse transcriptase is required for sequencing.

The goal of this project is to use RNA data obtained from low-quality FFPE tissue. As already described before, FFPE samples contain fragmented RNA transcripts due to the fixation process and storage conditions [44], [47], [51]. Therefore, traditional transcriptome capturing methods using oligo-dT primers that target the polyadenylation sequence of mRNA molecules are ineffective. In order to overcome this problem, specific library preparation protocols for RNA originating from FFPE tissue have been developed. By applying sequence-specific capture that does not rely on the presence of polyadenylated transcripts, it is possible to use low-quality FFPE tissue and samples with limited starting material for RNA-Seq. This strategy is being employed in Illumina's TruSeq® RNA Exome Kit, commonly used in library preparation protocols [48], [49]. Library preparation for RNA-Seq involves the conversion of cellular RNA into cDNA molecules and the addition of adapter oligonucleotides in order to make the RNA fragments suitable for sequencing. Paired-end sequencing will be employed since it provides additional information for alignment because the length of both the total cDNA fragment and the reads are known which increases the probability of mapping across splice junctions and indels [60]. Therefore, paired-end data can be advantageous for the detecting alternative splicing, identification of novel transcripts, identification of chromosomal rearrangements and for mapping to high-homology regions.

1.2.4 Alignment and Pre-Processing

As already described before, recent technological advancements in NGS have allowed fast sequencing of genomes at a low cost. Together with advanced bioinformatics tools, NGS allows comprehensive mapping of mutations in cancer, collectively called the mutanome [7]. However, one critical challenge of the development of a personalised cancer vaccine is the accurate alignment of the cancer mutanome in order to select the most relevant somatic mutations to induce an optimal immune responses [7], [60], [64]. No standardized bioinformatics protocol has been instated yet. As a result, different studies have applied a wide variety of protocols to align sequencing data and to call variants [32]–[36], [38], [65].

It is important to note that mapping of RNA-Seq data requires splice-aware alignment tools. Since not all exons and splice junctions are known, splice junction mapping is critical to map reads across unknown splice junctions and to understand alternative transcript usage. A lot of splice-aware alignment tools are available nowadays [66], [67]. A second concern is the gapped alignment for indels which still remains a significant bioinformatics challenge [55]. Gaps caused by insertion or deletion of nucleotides can slow down alignment speed because it is more difficult to identify the right mapping positions of these reads. The sensitivity of the detection of splice junction and the alignment of gaps is enhanced when longer reads are sequenced. Since the accuracy of downstream analyses heavily depends on the alignment, it is important to select the most accurate aligner for this pipeline. In order to do so, splice-aware alignment tools were selected based on previous comprehensive benchmarking [55], [66], [67] and tool performances were evaluated for our data.

The first splice-aware aligner that was selected is GSNAP (Genomic Short-read Nucleotide Alignment Program) [68]. GSNAP is a fast and memory-efficient method to align both single- and paired-end reads as short as 14 nucleotides and of arbitrary length. The alignment algorithm can detect short- and long-distance splicing (both novel and known splice sites), complex variants and long indels. Besides, GSNAP is an SNV-tolerant aligner, where minor alleles are treated as matches to a reference space, rather than mismatches to a reference sequence. In order to do so, GSNAP employs a search process of merging and filtering position lists from a genomic index at the oligomer level [55].

As second, the STAR (Splice Transcripts Alignment to a Reference) aligner was selected [69]. STAR is an ultrafast aligner but on the other hand, demands a significant amount of RAM (~30GB for the human genome). STAR aligns non-contiguous sequences directly to a reference genome in two steps: a seed searching step followed by seed clustering and stitching [60]. In the seed searching step, the algorithm finds the Maximal Mappable Prefix (MMP), then takes the unmapped portion of the read and finds the MMP for that fragment. This approach represents a natural way of finding exact locations of splice junction without any a priori knowledge of splice junctions' loci or properties [55]. The MMP in the seed searching step is implemented through uncompressed suffix arrays, which increases speed but also memory usage. In the second phase of the algorithm, STAR stitches together all the seeds that were aligned to the genome in the first phase using a local alignment scoring scheme. The stitched combination with the highest score is selected as the best alignment of the read [69]. An interesting feature of the STAR aligner is the ability to operate in 2-pass mode. The 2-pass mode allows more sensitive novel junction discovery. The basic idea is to run a first pass of STAR mapping with the usual parameters. The junctions detected in this first pass will be used as 'annotated' junctions in a second run of the STAR aligner.

For each alignment tool, the parameter space is enormous. This makes it impossible to analyse each setting and to optimise these parameters in order to obtain the highest precision and sensitivity. Baruzzo *et al.* [66] used a heuristic strategy to search the parameter space of each alignment tool and concluded that most alignment tools (e.g. GSNAP, MapSplice2 and STAR) perform best with default settings. For this reason, the aligners evaluated in this thesis will mostly use default parameter settings.

The obtained SAM and BAM files need some further processing and filtering before proceeding to the variant calling step of the pipeline. As already mentioned before, PCR duplicates can violate assumptions of variant calling potentially resulting in false positive called variants. Hence, it is important to remove or mark these duplicates in a way they are not taken into account for variant calling. However, since it is impossible to distinguish true technical duplicates from serendipitous biological duplicates, variant calling tools will be rather conservative in calculating the confidence of variants. Other processing steps that will be evaluated involve the splitting of N cigar reads, the recalibration of base quality scores and indel realignment.

1.2.5 Variant Calling

Variant calling implies the identification of variants from sequence data. In order to detect only somatic mutations, RNA-Seq data from tumour tissue will be compared with WES data from a matched healthy tissue sample, thereby avoiding the incorrect classification of germline variants as neoantigens [7]. Various types of somatic mutations can result in T cell-recognised neoepitopes: SNVs, indels, splice site mutations and gene fusions. Most variant calling algorithms work well for SNVs since they are the most abundant type of tumour mutation [5] and because of the relative simplicity and reliability of identifying sequence changes of one base pair [41]. If SNVs resulting in a nonsynonymous mutation are expressed, they can generate T cell-recognised neoepitopes. Besides SNVs, indels and gene fusions can lead to highly immunogenic frameshifts [41], [70], [71]. Turajlic *et al.* [41] proved in a pan-cancer analysis that indels giving rise to a frameshift mutation are a highly immunogenic mutational class since this frameshift mutation may result in a highly divergent amino acid sequence in the resulting protein and hence may produce strong neoantigens. Indeed, it was shown that a high abundance of frameshift neoantigens was associated with upregulation of genes involved in the immune response, including MHC class I antigen presentation, CD8$^+$ T cell activation, and increased cytolytic activity. As a result, it was concluded that frameshift mutations creating novel open reading frames might be an important source of tumour-derived neoantigens and in this way induce multiple neoantigen reactive T cells, because of both an increased number of mutant peptides and a reduced susceptibility to self-tolerance mechanisms [41].

Nevertheless, the detection of indels longer than 2 base pairs appears to be more challenging than SNV detection from RNA-Seq data [55]. For this reason, many of the neoantigen detection pipelines previously described mainly focused on SNVs detection [32], [34]. Piskol *et al.* [32] developed a highly accurate approach termed SNPiR to identify SNVs in RNA-Seq data using consecutively a splice-aware aligner and the Genome Analysis Toolkit's (GATK's) [72] UnifiedGenotyper for calling variants. In a next step, additional filtering criteria are applied on the called variants for ensuring removal of artefacts that might have been introduced. These filters include the removal of false calls in duplicated regions, in homopolymeric regions, or close to splice junctions. In addition, known RNA editing sites are removed from the called variants. Removal of these false positive calls resulted in a high precision of SNV detection. However, indel detection was not considered. It is important to note that UnifiedGenotyper is a caller for germline mutations and not for somatic mutations. Therefore, it is more appropriate to use GATK's somatic variant caller MuTect2 instead. For example, Coudray *et al.* [33] suggested another approach using the STAR aligner's 2-pass procedure combined with MuTect2. Since the MuTect2 [73] algorithm uses some hard-coded filters to remove false positive variants, no additional criteria were applied. Using this pipeline, they were able to identify variants (both SNVs and indels) that were missed by WES. Finally, Neums *et al.* [56] recently developed a method called Variant Detection in RNA (VaDiR) that integrates 3 variant callers, namely: SNPiR [32], RVBoost [74], and MuTect2 [73]. In addition to the filtering procedures of the variant callers themselves, the results were further filtered by taking an intersection of called variants from the 3 callers and by applying a read depth (DP) and a VAF filter.

The variant calling on RNA-Seq data is more challenging because of splice junctions, RNA editing, allele-specific expression, variable levels of gene expression and the tumour heterogeneity, as already explained in section 1.2.3. Therefore, it is important to carefully select a variant calling tool that takes into account these limitations and to optimise additional filtering of the called variants. For our pipeline, variant callers identifying both SNVs and indels with a great accuracy and sensitivity were selected based on previously described pipelines [32], [33], [56] and performance evaluations of different variant callers [55], [75]. The three variant calling algorithms selected are: MuTect2, VarDict and Strelka2. The fundamental idea of these variant callers is to identify potential variants using the tumour sample and to distinguish somatic variants from germline variants using the matched normal sample.

It is important to note that sensitive variant calling largely depends on the alignment tool used for mapping. Soft-clipping reads (S in CIGAR string) at the deletion edge introduced by alignment tools increase mapping sensitivity and can potentially be used for some variant calling tools to identify indels through realignment. However, some variant callers rely on hard evidence indels marked in the CIGAR string (I or D in CIGAR string) and soft-clipped reads are generally ignored by these variant callers [55]. For this reason, different combinations of alignment tools and variant callers will be evaluated to find the optimal combination.

The first variant caller MuTect2 [73], [76] is a popular and well documented somatic variant caller developed by GATK [72] and is based on the original MuTect algorithm and the germline variant calling tool HaplotypeCaller. While the HaplotypeCaller relies on a ploidy assumption (diploid by default), MuTect2 allows for a varying allelic fraction for each variant, as is often seen in tumours with purity less than 100%, multiple subclones, and/or copy number variation. MuTect2 is a haplotype-based variant caller and therefore, indel realignment is no longer necessary because the original local alignment information is discarded and reads are assembled and realigned [77]. Since MuTect2 is designed to call somatic variants only, the algorithm skips variant sites that are clearly identified as germline variants based on a comparison of the normal and tumour sample. Performing this step at an early phase avoids spending computational resources on germline events. MuTect2 infers genotypes based on two log-odd ratios: TLOD and NLOD. The former scores the confidence that a mutation is present in the tumour sample and the latter scores the confidence that a mutation is absent from the matched normal sample. The thresholds that are used by MuTect2 to consider a variant as being real and somatic (resulting in the annotation “PASS”) are by default TLOD > 6.3 and NLOD > 2.2 [33]. Besides these thresholds, MuTect2 applies a number of hard-coded quality filters to select true variants with a high specificity. An additional option for MuTect2 is the use of a Panel of Normals which allows to call somatic variants without the need for a paired normal sample.

The second variant caller that will be evaluated is VarDict [78]. VarDict is a sensitive variant caller for both single (tumour sample only) and paired sample (tumour and normal sample) variant calling from alignment files and can be used for both DNA and RNA sequencing data. Just like MuTect2, VarDict performs local realignment to improve indel identification, by which soft-clipped reads are taken into account. The VarDict algorithm employs a heuristic approach to identify variants whose supporting reads meet a defined threshold to filter out sequencing artefacts. These potential variants are analysed in the matched normal sample to filter out germline variants applying Fisher’s exact test [77]. In addition, VarDict performs amplicon aware variant calling for PCR-based targeted sequencing and has a built-in option to perform de-duplication on the fly, removing the necessity for an additional step and so improving efficiency. When using VarDict in paired sample mode, somatic variants can be detected. In contrast to MuTect2, VarDict uses only a limited number of hard-coded quality filters. Several downstream strategies have been developed to filter variants. For example, Blue Collar Bioinformatics (bcbio) provides an overview of how to develop further filters for VarDict [79]. Previous evaluation of several variant callers [75] appointed VarDict as best performing variant caller, considering both sensitivity and positive predicted value.

The third variant caller that was selected is Strelka2 [80]. Strelka2 is an accurate and fast variant calling tool build upon the original Strelka somatic variant calling algorithm and can be used for both germline and somatic variant calling. This algorithm relies on a series of successive steps: parameter estimation from sample data, candidate variant discovery, realignment, variant probability inference, and empirical re-scoring and filtration. Strelka2 is capable of detecting SNVs and indels up to a predefined maximum size of 49 bases. Besides, good variant calling results are provided down to about 5-10% tumour purity given sufficient normal and tumour sequencing depth. Similar to MuTect2 and VarDict, Strelka2 requires a matched normal sample to be able to distinguish between both germline variants and true somatic variants in the tumour sample. Strelka has already been used by several research groups [35], [39] to detect variants for the development of a personal neoantigen vaccine.

The output of these variant calling tools is a variant call format (VCF) file that stores all variant information. Based on the genomic location of a variant, additional information can be obtained using various annotation tools like Variant Effect Predictor (VEP) [81] from Ensembl, ANNOVAR [82], Oncotator [83] or SnpEff [84]. This information includes genes and transcripts affected by the variants, consequence of the variants, known variant from the 1000 Genomes Projects and SIFT and PolyPhen scores. Once variant detection is completed, each variant is annotated to predict the amino acid change that resulted from the altered RNA sequence.

It is important to note that some of the identified variants can be RNA editing sites. As already explained before, RNA editing involves mostly A:T→G:C transition. Several databases of RNA editing sites in humans exist: The Inosinome Atlas [62], RADAR [85] and REDIportal [86]. These databases can be used to identify known RNA editing sites in order to separate them from genomic variants. Filtering out RNA editing sites will increase the overlap with the WES called variants and as a result enhance the precision of the variant calling using RNA-Seq.

1.2.6 Neoepitope Prediction and Selection

After sequencing, alignment and variant calling have been performed, an annotated VCF file containing information about the identified variants is obtained. Not every identified neoepitope will be presented on an MHC class I or II molecule and/or induce an immune response. As already explained in section 1.1.3, the processing and presentation of antigens is a complex, multistep process and only specific peptides will fit in the peptide-binding groove. For this reason, only a fraction of the neoantigens is presented on MHC molecules at sufficient levels to induce an effective T cell response. Therefore, it is critical to select neoantigens with the highest likelihood of immunogenicity. Critical components of this neoepitope selection process are the in-silico prediction of MHC class I and II binding affinities for specific peptides and RNA expression analysis [7].

Interactions between a specific peptide and the binding pocket residues of an MHC molecule depends on the type of HLA proteins and which amino acids interact in the binding groove. Therefore, the binding affinity of any peptide is sequence-specific relative to that patient's HLA proteins. Based on sequence data, the HLA haplotype-specific to the patient can be identified. This HLA haplotype can be used for the prediction of binding affinity using an artificial neural network based learning method developed from a training set of experimentally derived binding affinities [7], [87].

Most commonly used software package for the prediction of peptide-MHC class I binding is NetMHC[88], [89] that uses artificial neural networks that have been trained for 81 different human MHC alleles. Variant-containing peptides with a length between 8 and 11 amino acids are parsed as input data for NetMHC along with corresponding wild type peptides and HLA class I haplotypes. For each peptide, the algorithm outputs a predicted half maximal inhibitory concentration (IC_{50}), which measures the concentration of a given peptide needed to compete with a standardized peptide already bound to a given MHC I allele [87]. A widely used threshold to determine whether a neoepitope will have a strong to intermediate binding affinity and thus will most likely elicit a CTL response is an IC_{50} lower than 500 nM. However, this arbitrary threshold has been questioned, since it has been proven that several peptides with a predicted IC_{50} well over 500 nM elicited a $CD8^+$ T cell response [37]. Besides the prediction of binding affinity, also other steps of antigen processing and presentation, stability, and TCR recognition can be incorporated in prediction algorithms. Unlike MHC I molecules, predictions for MCH class II molecules are significantly more challenging due to extensive HLA class II polymorphism in the general population and the open binding groove [90].

Besides neoepitope selection based on predicted binding affinity values ($IC_{50} \leq 500$ nM) additional filtering steps can be applied. An important criterium is the expression level of the mutated gene in the tumour tissue. It has already been observed that gene expression levels, the amount of translated protein, the cell surface density of MHC ligands derived from it, immune recognition and lysis of the respective cell are all positively correlated [7], [90], [91]. In this way, it is possible that a highly expressed protein can account for a high T cell response even if its IC_{50} is higher than 500 nM. The RNA-Seq data of the tumour tissue can be used for expression analysis.

Besides the IC_{50} value and the expression level, the VAF and coverage are also useful criteria to further select variants. It is important to note that not only the gene but also the variant allele should be expressed. The VAF measures how many reads contain the mutations compared to reads containing the normal sequence. In theory, heterozygous mutations should have a VAF around 0.50. However, somatic mutations in tumour tissue appear at lower frequencies since tumour tissue is heterogeneous. First of all, because tumour tissue partially consists of healthy cells, and second, because tumour cells contain both clonal and subclonal mutations [33]. Moreover, copy number variations can lead to gain or loss of chromosomal regions, and duplication or deletion of genes [92].

1.3 Personalised Immunotherapy

Lung cancer is the leading cause of cancer-related deaths worldwide [1]. Although various treatment methods, such as surgery, radiation therapy and chemotherapy have been used to treat lung cancer patients, a high mortality is still observed in patients with advanced stages of tumours. Recently, a new approach has been studied to treat cancer using patient's own immune system: immunotherapy. The currently most effective cancer immunotherapies include adoptive cell transfer, immune checkpoint inhibition and therapeutic cancer vaccines [93]. These immunotherapies are aimed at (re)activating and expanding tumour-specific CTLs, with the ultimate goal being the destruction of primary cancer tumours and metastatic tumours. CD8$^+$ T cells play a central role in immunity to cancer. When a CTL recognises a specific antigen presented on the surface of a malignant cancer cell through interaction between the TCR and the MHC I molecule, the cancer cells are killed by synaptic exocytosis of cytotoxic perforins and granzymes. However, when a tumour enters the escape phase, several immune evasion strategies are adapted by the cancer cells that hinder the recognition and killing by CTLs. The goal of immunotherapies is to circumvent this immune escape by reactivating and enhancing the patient's immune system.

ACT involves the *in vitro* expansion of tumour infiltrating lymphocytes (TIL) obtained from the patient's tumour tissue. When the expanded TIL are re-infused in the patient, they will display an increased specificity towards the cancer cells which will result in an enhanced spontaneous T-cell response and degradation of the cancer cells. Nowadays, an even more advanced technique is being developed: CAR-T cell therapy. For this therapy, isolated T cells are genetically engineered to express a chimeric antigen receptor (CAR) that specifically recognises TSA on cancer cells and enhances T cell proliferation. Another possible approach is to circumvent cancer's immune-evading strategies using immune checkpoint blockers, this is called immune checkpoint inhibition (ICI). One possible target is the PD-1 receptor, an immune checkpoint receptor expressed by activated T cells which induces immunosuppression through interaction with its ligands PD-L1 and PD-L2. Tumour cells can express these ligands and, in this way, escape the immune response. Immune checkpoint blockers are antibodies that interact with PD-1 or PD-L1/2 and in this way restore the cytotoxicity of pre-existing cancer-specific T cells in order to destroy the cancer cell. Another possible target for ICI is CTLA-4, a CD28 homolog expressed on T cells that can inhibit T-cell proliferation. Clinical studies have already proven the effectiveness of these therapies. However, only a limited number of patients respond to the treatment, indicating the need for an additional component in the treatment like for instance a tumour-specific vaccine.

The detection and selection of neoantigens, as previously described in detail, allows the development of a personalised therapeutic cancer vaccine specifically targeting neoepitopes presented on the analysed cancer cells. Different types of therapeutic cancer vaccine exist. Nevertheless, for this pipeline, an mRNA-loaded dendritic cell vaccine will be used to target and kill malignant cancer cells [94]. Mature DCs play a major role in the Cancer-Immunity Cycle since they evoke T-cell priming and activation, leading to the recognition and eradication of cancer cells [18]. However, this is a highly complex process involving different immune cells, cytokines and co-stimulatory molecules. Brabants *et al.* [94] developed a good manufacturing practice (GMP)-compliant culture protocol that generates high yield of mature DCs in a short period of time. The produced DCs can translate incorporated mRNA and present the resulting neoepitopes bound to an MHC molecule on the cell surface. In this way, the DCs can prime and activate CD8$^+$ T cells specifically targeting cancer cells based on the somatic variants that were selected in the bioinformatics pipeline.

Personalised cancer vaccines are a promising new approach of cancer treatment that offer great improvements over current treatment options. One major advantage of this approach is the possibility to produce DCs targeting different neoepitopes, thereby overcoming the problem of tumour heterogeneity. In this way, not only the treatment response is enhanced but also the risk of treatment resistance is lowered, and a higher chance of patient survival can be achieved. Although it is becoming increasingly evident that in the future personalised immunotherapy has the potential to replace conventional therapies, it will likely be a combination therapy that achieves optimal treatment outcomes. A combination therapy of a personalised dendritic cell vaccine and immune checkpoint blockers offers a great potential. Induction of the immune response by DC vaccination can evoke up-regulation of PD-L1 in the tumour microenvironment. As a result, efficacy of immune checkpoint blockers anti-PD-1 and anti-PD-L1 is improved and a higher treatment response can be obtained. In addition, vaccine induced memory T cells may enhance durability of the anti-PD-1 and anti-PD-L1 effects by promoting robust memory responses [7]. However, the implementation of such a personalised combination therapy as standard therapy is still ongoing research and many challenges have to be solved.

2 MATERIALS AND METHODS

2.1 Data Collection

In this section, the extraction of nucleic acids and the library preparation will be explained. The tumour samples were provided by Prof. Karim Vermaelen at the UZ Gent. Lung tissue was obtained directly from a lung squamous cell carcinoma during a lobectomy. Two sample types with an estimated tumour purity of 50% were analysed: FFPE samples and FF samples.

2.1.1 RNA Extraction

For the extraction of RNA from the FFPE samples the RNeasy® FFPE Kit (Qiagen) was used [95]. This kit is specially designed for purifying total RNA from FFPE tissue sections. The paraffin embedded tissue was trimmed of the glass slides using a scalpel and a Deparaffinization Solution was added to remove the paraffin. In a next step, optimised lysis buffer was added to allow sample lysis with proteinase K digestion. After lysis, samples were incubated at 80°C for 15 minutes to reverse formalin crosslinking. In a next step, DNA was removed from the FFPE sample using DNase Booster Buffer and DNAse I. Finally, the concentrated RNA was purified using RNeasy® MinElute spin columns. Quality metrics of this extraction protocol can be found in Appendix 8.1.

The extraction of RNA from the FF samples was performed using the Maxwell® RSC simplyRNA Tissue Kit (Promega) [96]. This procedure purifies total RNA with minimal sample handling before automated purification on a Maxwell® RSC Instrument (Promega). The automated procedure involves paramagnetic particles which provide a mobile solid phase to optimise sample capture, washing and purification of nucleic acids. Sample pre-processing was performed manually and involved a tissue homogenization step (1-Thioglycerol/Homogenization Solution) and a DNA removal step (DNase I Solution).

2.1.2 Library Preparation and Sequencing

For the library preparation of the extracted RNA material the TruSeq® RNA Exome Kit (Illumina) was used [97], [98]. This library preparation kit does not rely on the presence of the polyadenylation signal since it specifically targets the RNA coding region (exonic regions) using sequence-specific capture. Library preparation for RNA-Seq involves the conversion of cellular RNA into molecules that are suitable for sequencing (Figure 5). Some abundant RNA such as ribosomal RNA (rRNA), transfer-RNA (tRNA) and small nucleolar RNA (snoRNA) can comprise up to 80% of the total cellular RNA and in this way reduce the depth of sequence coverage, resulting in less detection of lowly expressed RNA [64]. However, when using the TruSeq® RNA Exome Kit this problem is minimised. First, RNA molecules were fragmented into a smaller size to be suitable for sequencing by the Illumina platform. Once RNA of the appropriate size was obtained, the single-stranded RNA molecules are converted into double-stranded complementary DNA (cDNA) using reverse transcriptase (RT) with random primers during first strand synthesis. The second strand synthesis removes the RNA template and synthesizes a replacement strand, incorporating dUTP in place of dTTP to generate double-stranded cDNA. During the Adenylate 3' End step a single ‘A’ nucleotide is added to the 3’ ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single ‘T’ nucleotide on the 3’ end of the adapter provides a complementary overhang for ligating the adapter to the fragment [98]. In a next step, adapter oligonucleotides are ligated to the cDNA to allow amplification and sequencing. These adapters consist of a sequencing binding site, indices and nucleotides complementary to the flow cell oligonucleotides. In paired-end sequencing, the adapter sequences contain a ‘read 1’ index at one end of the cDNA fragment and a ‘read 2’ index at the other end of the cDNA fragment. In this way, 2 reads are amplified from one cDNA fragment. Paired-end sequencing provides additional information for alignment since the length of both the total cDNA fragment and the reads are known which increases the probability of mapping across splice junctions and indels. Therefore, paired-end data can be very useful for estimating alternative splicing, identification of novel transcripts, identification of chromosomal rearrangements and for mapping to high-homology regions. After adaptor ligation, a PCR step using a PCR Primer Cocktail is performed to selectively enrich those cDNA fragments that have adapter molecules on both ends and to amplify the library. In a final step of the library preparation protocol, the coding regions of the transcriptome are then captured using sequence-specific probes. In order to do so, streptavidin beads are used to capture the probes hybridized to the targeted regions of interest [98]. The TruSeq® RNA Exome protocol involves two hybridization and two capture rounds.

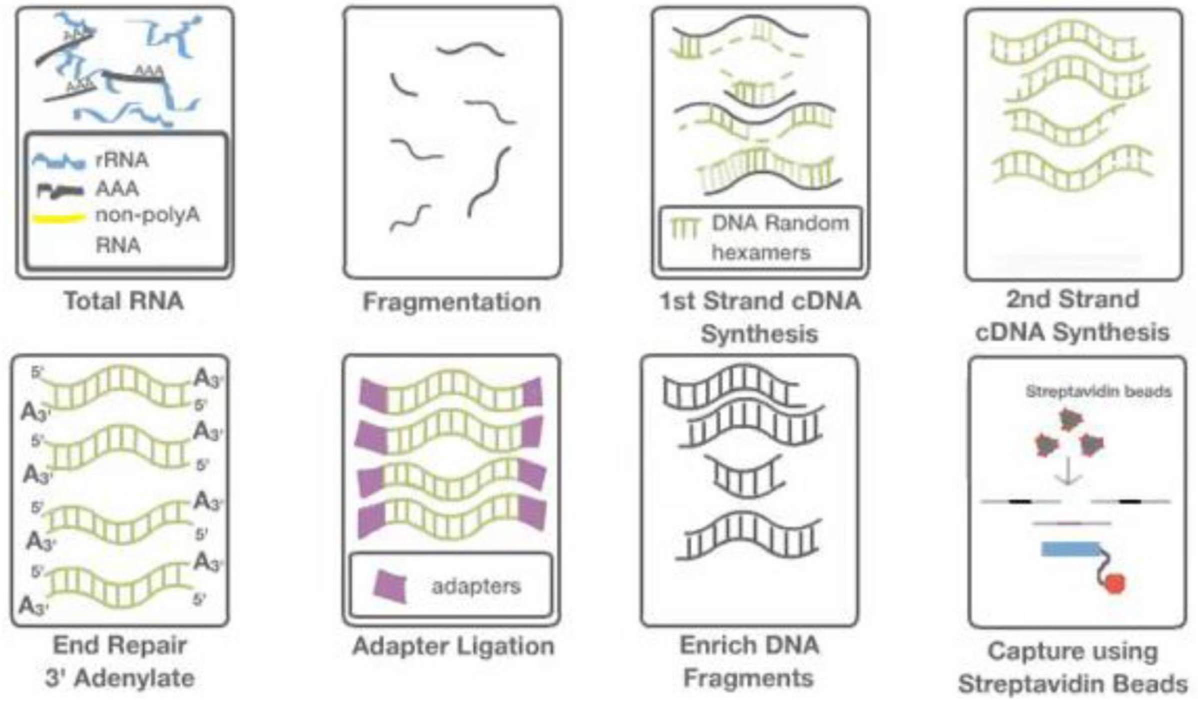


Figure 5. An overview of Illumina’s TruSeq® RNA Exome Kit [87].

2.2 Data Processing

The different tools and scripts described in this chapter ran on either the HPC infrastructure that consists of several Tier2 clusters which are hosted in the S10 datacentre of Ghent University, or locally on a laptop. The laptop ran on Windows 10 Home and the Ubuntu 16.04.4 LTS terminal. The connection with HPC server was made using the Secure Shell (SSH) protocol ensuring secure network services. Two different SSH clients where used: PuTTY (version 0.70) and WinSCP (version 5.13.3). PuTTY was used to connect to the HPC server and WinSCP was needed for the file transfer between the local computer and the remote HPC server.

The sequencing data was processed using GSNAP [68], STAR [69], GATK [72], Picard [99], Biobambam [100], R [101], Python [102], MuTect2 [73], VarDict [78], Strelka2 [80], VEP [81], RTG Tools [103], Hap.py [104], BEDTools [105], VAtools [106], SAMtools [107], [108] and BCFtools [107], [108], which is a part of the SAMtools package. A detailed overview of the used version of these tools can be found in Appendix 8.2. Example scripts mentioned in the sections below are available in my personal GitHub repository (https://github.ugent.be/lovouden/Thesis_Lore_Van_Oudenhove).

2.2.1 Alignment

After sequencing, fastq files containing raw sequencing data were obtained. To assess the quality of this sequencing data, FastQC [109] was used to analyse the fastq files. Some important quality control metrics are base quality over read length, per sequence quality score, QC content and sequence duplication levels. Reads containing a low-quality score or other sequence artefacts may be trimmed or corrected to improve the alignment quality. Following quality control, the processed data can be aligned to a reference genome using an appropriate aligner.

For RNA-Seq, two aligners were evaluated: GSNAP and STAR. The reference genome used is hg38, this is the assembly of the human genome released December of 2013 which uses alternate or ALT contigs to represent common complex variation in the human genome. However, GSNAP and STAR cannot handle these ALT contigs. Therefore, ALT contigs were removed and hg38-noalt was used as reference genome.

Before aligning the raw RNA-Seq data to the reference genome, the total cellular RNA was filtered. In order to do so, the raw fastq files were aligned to known human tRNA, rRNA and snoRNA sequences. The fasta files containing these sequences were provided by BioBix. The remaining unmapped reads were used for further mapping against hg38-noalt.

STAR

In a first step, STAR genome indexes were created using `--runMode genomeGenerate`. The index file for hg38-noalt was created using the hg38 GTF annotation file hg38-noalt.93.gtf and the option `--sjdbOverhang 150` to enhance mapping of known splice junctions. The STAR aligner was used in 2-pass mode with adjusted parameters `--outFilterMismatchNmax 2`, `--outFilterMultimapNmax 10`, `seedSearhStartLmaxOverLread 0.5` and `--outSAMmapqUnique 60` in order to allow further processing using GATK. For other parameters default values were maintained. Example scripts with the implementation of these commands are available in my personal GitHub repository: RNA-Seq_STAR_Index_Creation.sh and RNA-Seq_STAR_Alignment_FFPE_sample.sh.

GSNAP

GSNAP genome indexes were created using `gmap_build`. Next, the filtered fastq files were aligned to hg38-noalt using `gsnap` with additional option `-s hg38-noalt.splicesitesfile.iit` to supply the known splice junctions file. This file was created using the GSNAP function `gtf_splicesites` and the annotation file hg38-noalt.93.gtf. For the additional options the default values were maintained. Example scripts with the implementation of these commands are available in my personal GitHub repository: RNA-Seq_GSNAP_Index_Creation.sh and RNA-Seq_GSNAP_Alignment_FFPE_sample.sh.

2.2.2 Pre-Processing

In general, the variant calling process consists of three components: pre-processing, variant evaluation, and post-filtering. The main purpose of pre-processing is to make the alignment files suitable for variant calling and to reduce low-quality reads. First of all, Picard [99] was used to validate the obtained BAM file (`ValidateSamFile`) and to add read groups (`AddOrReplaceReadGroups`). Next, the BAM file was sorted using `samtools sort` and duplicate reads were marked using Biobambam2's `bammarkduplicates`. Duplicate reads were marked by adding the 0x400 bit (1024) flag to the second column of a SAM record, for each mate of a pair. Before somatic variant calling, GATK's SplitNCigarReads, BaseRecalibrator and/or IndelRealigner were applied to enhance accuracy of variant calling. An example script with the implementation of these tools is available in my personal GitHub repository: RNA-Seq_PreProcessing_Frozen_GSNAP_ExampleScript.sh.

SplitNCigarReads

The SplitNCigarReads tool [110] was developed specifically for RNA-Seq data to splits reads into exon segments. This tool identifies all N cigar elements (base skipped from the reference) in sequence reads and creates k+1 new reads, where k is the number of N cigar elements. The part of the read that is right of the N (the intronic part) is hard clipped. In this way, the number of called false variants can be reduced. The input for SplitNCigarReads consisted of the alignment file, the reference genome and the additional option `-U ALLOW_N_CIGAR_READS` in order to be able to handle RNA-Seq data as input.

BaseRecalibrator

The BaseRecalibrator tool [111] can be applied to acquire more accurate base quality scores, which in turn improves the accuracy of the called variants. Since variant calling algorithms rely on these base quality scores to call variants, it is important to estimate the systematic technical error introduced by the sequencing machine. Base quality score recalibration (BQSR) is a machine learning algorithm that involves two key steps: first the program builds a model of covariation based on the data and a set of known variants; next it adjusts the base quality scores in the data based on this model. The base score recalibration involves two functions. First, `BaseRecalibrator` was applied to the alignment file in order to calculate the recalibration table. The input for `BaseRecalibrator` consisted of the reference genome, a VCF file with known human variants (`--knownSites dbsnp-150.vcf.gz`) and the additional option `-U ALLOW_N_CIGAR_READS`. In a second step, these recalibration values were applied to the alignment file using `PrintReads`.

IndelRealigner

The IndelRealigner [112] tool locally realigns reads to minimize the number of mismatching bases across all the reads. Mismatching bases can accumulate due to the presence of an indel. Local realignment of this regions transforms reads with misalignment caused by indels into clean reads containing a consensus indel which eventually results in a more accurate detection of indels and SNVs by variant calling tools. There are 2 steps to the realignment process. In a first step, (small) suspicious intervals which are likely in need of realignment are identified using the `RealignerTargetCreator` tool. The input for this tool consisted of the reference genome, the alignment file, a file containing knowns SNPs (`-known resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf`) and a file containing known indels (`-known resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf`). The resulting list of target intervals are used as an input for the `IndelRealigner` tool to produce a realigned version of the input alignment file.

Each pre-processing step has an impact on the number of variants identified and the precision of variant calling. Therefore, different combinations of pre-processing steps will be evaluated.

2.2.3 Variant Calling

After pre-processing the alignment files, somatic variant calling was performed. Three different aligners were evaluated: MuTect2, VarDict and Strelka2. All variant callers operated in paired tumour-normal mode comparing a tumour BAM file (RNA-Seq) and a normal BAM file (WES). The normal BAM file consisted of WES data aligned using BWA-MEM, and pre-processed using `samtools sort` and Biobambam2's `bammarkduplicates`. Each variant calling algorithm produces a VCF file containing the identified somatic variants. This VCF file consists of 11 columns: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, TUMOR and NORMAL.

MuTect2

In order to enhance the speed and efficiency of variant calling, the regions of interest were divided over several interval lists. The input for MuTect2 consisted of the processed WES data from the normal sample and the RNA-Seq BAM file from the tumour sample, the reference genome, the interval list and additional option `-U ALLOW_N_CIGAR_READS` in order to be able to handle RNA-Seq data as input. The resulting VCF files of different intervals were zipped, indexed and merged using `rtg bgzip`, `rtg index` and `rtg vcfmerge` from the RTG Tools package respectively. The identified variants are subjected to MuTect2 built-in hard-coded quality filters (Table 1). A disadvantage of these MuTect2 quality filters is the fact that thresholds used for filtering are often hard-coded and therefore it is not possible to modify them. An example script with the implementation of MuTect2 variant calling is available in my personal GitHub repository: RNA-Seq_Frozen_GSNAP_Mutect2_batch1_ExampleScript.sh.

Table 1. MuTect2 built-in hard-coded quality filters.

Filter	Description
PASS	The variant passed all filters.
alt_allele_in_normal	Reject false positives in the tumour data by looking at the normal data for evidence of the alternate allele beyond what is expected from random sequencing error [73].
clustered_events	When several mutations are close together, they are filtered out because this is often a sign of being an artefact [73].
clustered_read_position	Reject false positives caused by misalignments originating from the alternate alleles being clustered at a consistent distance from the start or end of the read alignment [73].

germline_risk	Reject variants that show sufficient evidence of being germline based on dbSNP, COSMIC and the matched-normal sample (NLOD value) [73].
homologous_mapping_event	This filter detects homologous sequences and filters out variants falling into sequences that have three or more events observed in the tumour [73].
multi_event_alt_allele_in_normal	Reject variants when multiple events are detected at the same position in the matched-normal sample [73].
panel_of_normals	Reject variants seen in at least 2 samples in the Panel of Normals (when provided) [73].
str_contraction	Reject variants from short tandem repeat regions [73].
strand_artifact	Reject false positives caused by context-specific sequencing errors where the vast majority of the alternate alleles are observed in a single direction of reads [73].
t_lod_fstar	Reject a variant when the specific TLOD > 6.3 threshold is not reached, suggesting insufficient evidence of its presence in the tumour sample [73].
triallelic_site	Variant filtered because more than two alternate alleles pass the TLOD > 6.3 threshold [73].

VarDict

Similar to MuTect2, interval list were provided in order to enhance speed and efficiency of variant calling. The input for VarDict consisted of the processed WES data from the normal sample, the RNA-Seq BAM file from the tumour sample, the reference genome and the interval list. To run VarDict in paired normal-tumour mode the VarDict scripts `testsomatic.R` and `var2vcf_paired.pl` were used and the minimum allele frequency was set at 0.01 (default value). The resulting VCF files of different intervals were zipped, indexed and merged using `rtg bgzip`, `rtg index` and `rtg vcfmerge` from the RTG Tools package respectively. Since VarDict not only calls SNVs and indels but also multi-nucleotide polymorphisms, SNVs and indels need to be extracted in order to be able to compare VarDict to the other variant callers. This can be done using `bcftools view --types snps,indels`. Opposite to MuTect2, VarDict has some filters that are not hard-coded which means the filtering thresholds can be changed using different parameters. When using default parameters, these VarDict filters are less stringent and results in more false positive identified variants. Therefore, additional filtering of the identified variants is required to enhance the precision of variant calling. An example script with the implementation of VarDict is available in my personal GitHub repository: RNA-Seq_Frozen_GSNAP_VarDict_batch1_ExampleScript.sh.

The aim of implementing additional filters is to reduce the false positive identified variants using VarDict. The Blue Collar Bioinformatics (bcbio) provides a guideline for the development of additional filters [79]. The additional filters that were applied can be found in Table 2. The implementation of these filters can be found in Appendix 8.3 or in my personal GitHub repository: RNA-Seq_FFPE_GSNAP_VarDict_AdditionalFilters_ExampleScript.sh.

*Table 2. VarDict additional quality filters. *filter described in the bcbio guidelines* [79].

Filters	Description
PASS	The variant passed all filters.
LowFreq_with_LowDepth*	Reject variants with a low allele frequency and a low total read depth (DP) or coverage. This filter also includes criteria for mapping quality, number of reads mismatches, low depth and low quality [79].
LowFreq_with_PoorQual*	Reject variants that have a combination of a low allele frequency, a low quality and high p-values [79].
LowFreqBias*	Reject variants that have a combination of a low allele frequency, a low depth, a low p-value for strand bias and more than two mismatches [79].
Exon-exon Junctions	Reject variant in close proximity of exon-exon junctions.
Likely Germline	Reject variants classified as germline according to VarDict.
Likely Somatic + Strong Somatic	Select variants classified as somatic or strong somatic variants according to VarDict.
Strong Somatic	Select variants classified as strong somatic variants according to VarDict.

Strelka2
For Strelka2, no interval lists were provided since this variant caller has a multithreading option. The input for Strelka2 consisted of the processed WES data from the normal sample and the RNA-Seq BAM file from the tumour sample, the reference genome and the `--exome` option. The Strelka2 output consisted of two separate VCF files for somatic SNVs and somatic indels. These VCF were merged using `rtg vcfmerge --force-merge=DP`. The resulting VCF file was zipped and indexed using `rtg bgzip`, and `rtg index` from the RTG Tools package respectively. Strelka2 has only two hard-coded built-in quality filters (Table 3). An example script with the implementation of Strelka2 is available in my personal GitHub repository: RNA-Seq_Frozen_GSNAP_Strelka2_Protocol_1.4_ExampleScript.sh.

Table 3. Strelka2 built-in quality filters.

Filters	Description
LowEVS	Reject variants when the Somatic Empirical Variant Score (SomaticEVS) is below the threshold [80].
LowDepth	Reject a variant when the tumour or normal sample read depth at this locus is below 2 [80].

3.1.3 Annotation

Functional annotation is a key step to understand the potential clinical impacts of identified variants. The Variant Effect Predictor [81] (VEP) is an Ensembl annotation tool that determines the effect of variants on genes, transcripts, and protein sequences, as well as regulatory regions. The additional option`--stats_file` was used to produce an HTML file containing annotation statistics. An example script with the implementation of VEP is available in my personal GitHub repository: RNA-seq_Frozen_GSNAP_Case1_Mutect2_GATK3-Annotate_VEP_ExampleScript.sh.

2.3 Performance Evaluation

The aim of this master thesis is to evaluate the different pre-processing steps, alignment tools and variant calling algorithms in order to improve the precision of the identification of somatic variants from RNA-Seq data obtained from FF and FFPE tissue. To do so, a comprehensive benchmarking of different workflows was performed. The performance of each workflow was evaluated based on the intersection of the identified variants with a Gold Standard set of variants.

2.3.1 Performance Metrics

In order to compare different pre-processing steps, aligners, and variant callers, the precision and sensitivity of variant detection was calculated using a Gold Standard set of variants. To obtain this Gold Standard set, somatic variant calling was performed on WES data from the patient's tumour tissue. This analysis was conducted and validated at the UZ Gent and involved BWA-MEM alignment and MuTect2 somatic variant calling. The Gold Standard set of variants is available on my personal GitHub: batch1-mutect2-annotated-decomposed.vcf.gz. The DNA variants in this Gold Standard set can be regarded as true positive (TP) variants. The Gold Standard variant set consists of 1254 variants, including 1113 SNVs and 141 indels. However, since it is impossible to verify whether the Gold Standard set of variants includes all true variants, it is more correct to refer to these Gold Standard variants as DNA concordant variants. The performance metrics used to evaluate different workflows are precision and sensitivity. The performance metrics were calculated using the following formulas:

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

2.3.2 Performance Evaluation of the Pre-Processing Steps

A first step in the optimisation of the somatic variant detection was the selection of the appropriate pre-processing steps. As already described before, three different pre-processing tools were analysed: SplitNCigarReads, BaseRecalibrator and IndelRealigner. Six protocols with different combinations of these pre-processing tools were established (Figure 6). Protocol 0 performs no additional pre-processing, Protocol 2 and 5 include only one pre-processing tool, Protocol 1 and 4 include two pre-processing tools and Protocol 3 includes all pre-processing tools. After pre-processing, the resulting BAM file was used for variant calling with MuTect2. The output VCF files were compared to the Gold Standard variant set. Two tools were used to identify TP, TN, FN and FN variants: som.py from the Hap.py package and `bcftools isec`. The `bcftools isec` command was used to obtain an intersecting VCF file that contained only the DNA concordant somatic variants. Finally, the performance metrics were calculated as explained in section 2.3.1. For the comparison of these six protocols mainly precision is taken into consideration since the main goal is to identify variants that are true positive. Currently, the best way to verify whether an RNA variants is true positive is to check if this variant is supported by DNA sequencing. In other words, when an RNA variant is included in the Gold Standard variant set, it can be considered true positive.

To evaluate whether different pre-processing protocols identified the same somatic variants, the concordance between different VCF files was analysed using `bcftools isec` and `bcftools stats`. Venn diagrams were created using the Python matplotlib_venn package (version 0.11.5).

2.3.3 Performance Evaluation of the Alignment Algorithms

A second step in the optimisation of the bioinformatics workflow is the evaluation of the alignment algorithms GSNAP and STAR. To select the most appropriate aligner, again the precision of variant calling was calculated and compared. In order to do so, pre-processing Protocol 4 and variant caller MuTect2 were used and only the alignment tool was changed. Similarly as before, the DNA concordant variants were identified using som.py and `bcftools isec`, and the precision and sensitivity were calculated.

To evaluate the concordance between GSNAP and STAR, the overlap between different VCF files was analysed using `bcftools isec` and `bcftools stats`. Venn diagrams were created using the Python matplotlib_venn package (version 0.11.5).

2.3.4 Performance Evaluation of the Variant Calling Algorithms

After alignment and pre-processing of the RNA-Seq data the actual variant calling was performed. As already described before three variant calling algorithms were compared: MuTect2, VarDict and Strelka2. The output VCF files were analysed using som.py from the Hap.py package. To include variants that did not pass the built-in quality filters the option `--include-nonpass` was used.

The results indicated that any single caller is not adequate in discovering variants with high precision. Therefore, it was tested if the combination of three variant calling algorithms would provide a higher rate of variant calls supported by WES. The overlap between the VCF files of MuTect2, VarDict (with additional filters) and Strelka2 was made using `bcftools -isec`. The resulting VCF files, containing only variants identified by all three variant callers, were analysed using `bcftools stats –apply-filters PASS` to obtain the total number of variants, the number of SNVs and the number of indels that passed the quality filters. An example script to evaluate this overlap is available in my personal GitHub repository: RNA-Seq_FFPE_GSNAP_MuVaSt_ExampleScript.sh. The precision was calculated similarly as before. To summarise these result, a Venn diagram was created using the Python matplotlib_venn package (version 0.11.5). Variants identified in all three variant callers were regarded as the most reliable variants.

To assess the difference between variants identified using RNA-Seq or WES, the expression of somatic variants called in WES was analysed using Kallisto data provided by the Center for Medical Genetics Ghent (CMGG). This Kallisto file included expression information (transcripts per kilobase million, TPM) on the gene level. The VCF Expression Annotator tool (VAtools) was used to add the expression data to the Gold Standard VCF file. Based on Supplementary Figure 1, a threshold of 1 TPM was used. In addition, the allele-specific expression of SNVs was evaluated using GATK's ASEReadCounter [113]. Indels were not evaluated since ASEReadCounter was unable to handle indels correctly. A VAF threshold of 0.04 was adapted based on a previous study by Karasaki *et al.* [34].

Besides expression, also the VAF and DP were analysed. The VAF was calculated separately for RNA and DNA variants using the following formulas:

$$VAF_{RNA} = \frac{AD_{alt}\ in\ RNA - Seq}{DP\ in\ RNA - Seq} \qquad VAF_{DNA} = \frac{AD_{alt}\ in\ WES}{DP\ in\ WES}$$

where AD_{alt} is the allelic depth of the alternate allele and DP the total read depth. This analysis was performed in Jupyter Notebook. An example script can be found in my personal GitHub repository: VAF_ExampleScript.py and DP_ExampleScript.py.

Finally, BEDTools was used determine whether the variants identified using RNA-Seq were covered by the WES SureSelect Human All Exon V6 (Agilent) capture kit using the ‘‘-intersectBed’’ function.

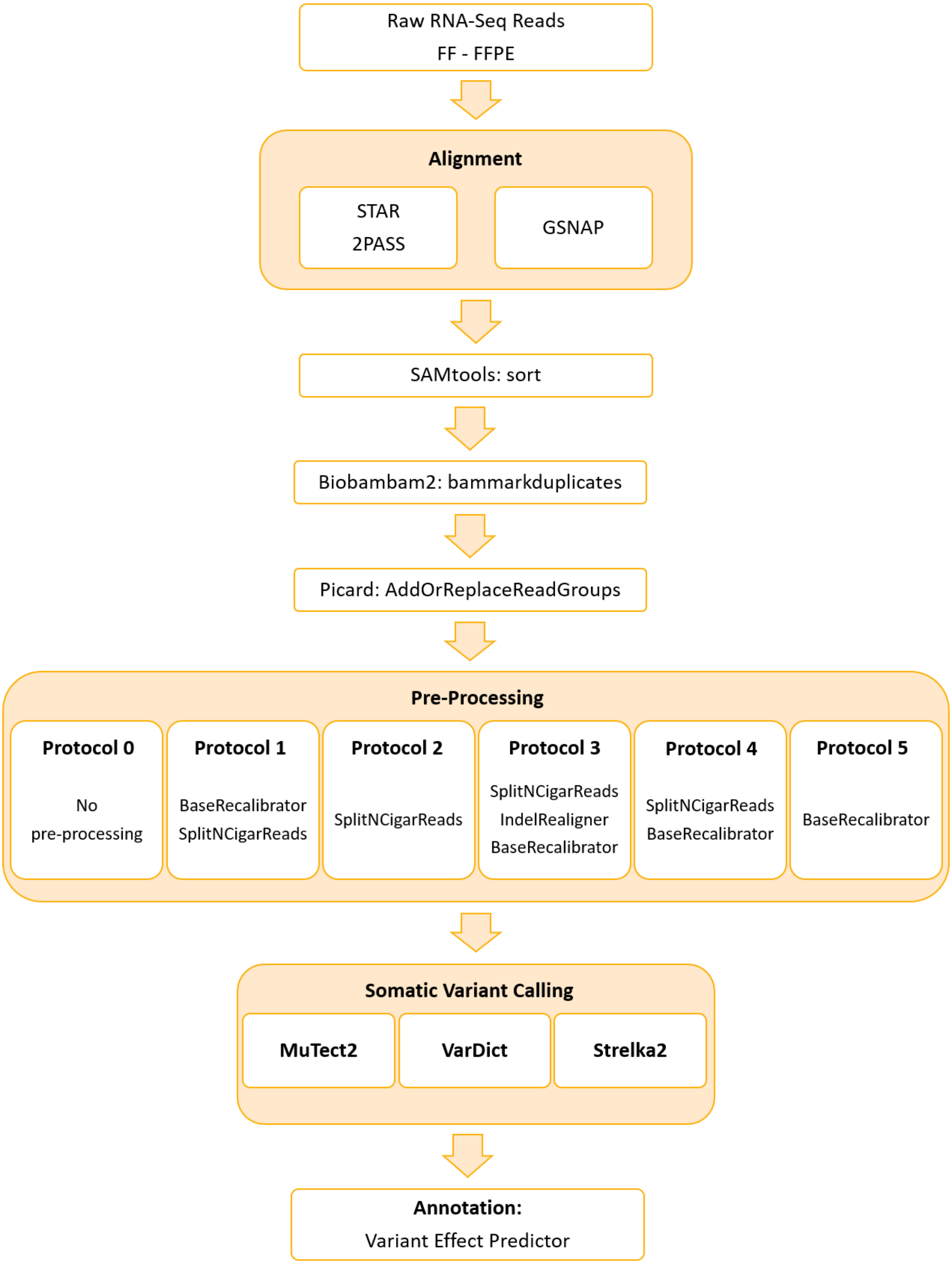


Figure 6. Overview Bioinformatics Pipeline.

3 RESULTS

For the performance evaluation of the pre-processing protocols, alignment algorithms and variant callers, one FF sample and one FFPE sample were used. The sequencing data consisted of fastq files containing paired-end reads. The quality of this sequencing data was assessed using FastQC [109]. The mapping statistics of both samples can be found in Appendix 8.5.

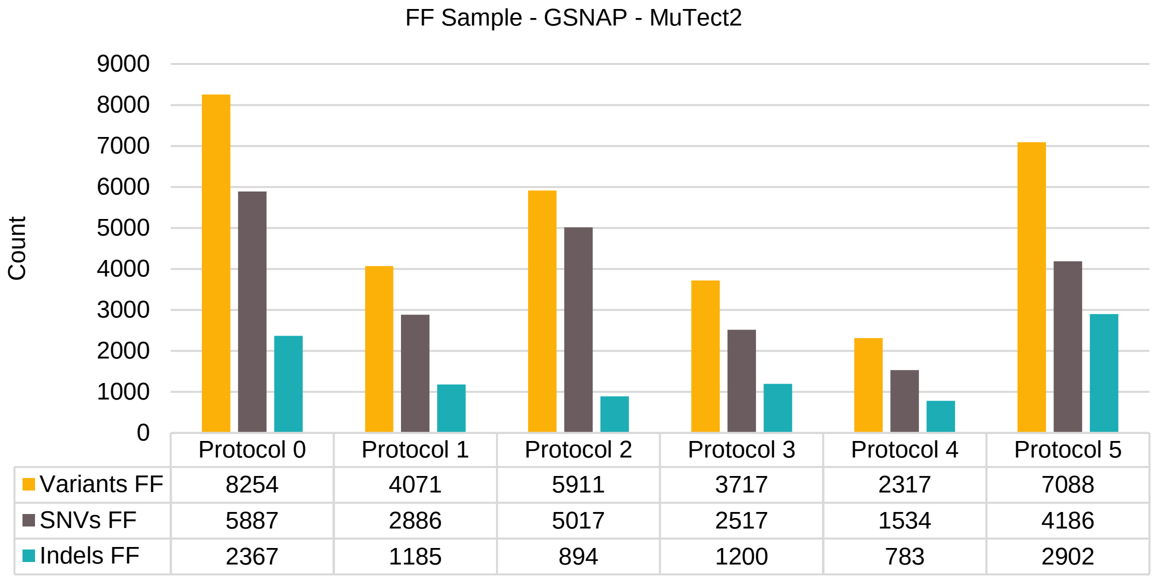
3.1 Performance Evaluation of the Pre-Processing Steps

A first step in the optimisation of the bioinformatics workflow for the identification of somatic variants is the selection of the appropriate pre-processing steps. To evaluate the performance of the different pre-processing steps, six pre-processing protocols (Figure 6) were compared for both sample types. This comparison was conducted for both aligners and one variant caller, MuTect2. The section below only describes the results for the GSNAP alignment since the same conclusion was drawn from the STAR alignment. The detailed results for STAR can be found in Appendix 8.6. The number of variants, the number of SNVs and the number of indels were visualised using a histogram. Only variants that have passed MuTect2's built-in quality filters (Table 1, section 2.2.3) were considered. A second graph shows the overall precision, the precision of SNV detection and the precision of indel detection for each protocol calculated as explained in section 2.3.1. Again, only variants that have passed MuTect2's built-in quality filters were considered.

3.1.1 GSNAP

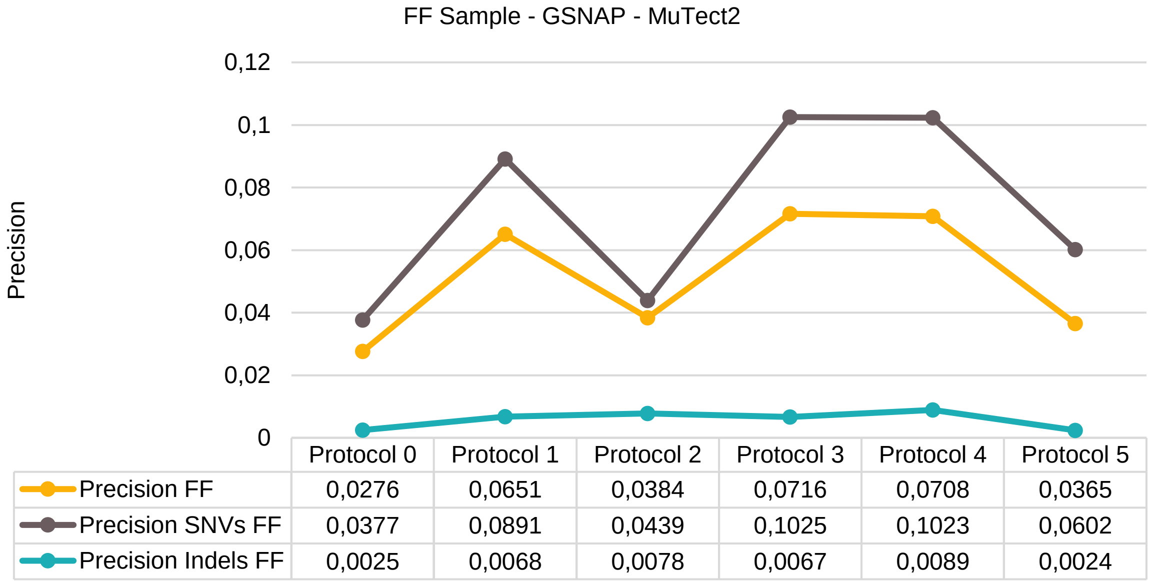
FF Sample

The first workflow, used to compare the six pre-processing protocols, was comprised of GSNAP alignment of the FF data and somatic variant calling with MuTect2. An overview of the number of called variants can be found in Figure 7. When no pre-processing steps were applied (Protocol 0, see Figure 6), the highest number of variants was called (8254). When only one pre-processing step was applied (Protocol 2 and 5, see Figure 6), fewer variants were called. Moreover, applying two or three pre-processing steps (Protocol 1, 3 and 4, see Figure 6) resulted in less variants. For each protocol, fewer indels than SNVs were detected. Figure 8 contains information on the precision of the called variants. Protocol 1, 3 and 4 (see Figure 6) obtained a higher overall precision than Protocol 0, 2 and 5 (see Figure 6). Generally, the precision of SNV detection was higher than the precision of indel detection. Moreover, application of the different pre-processing protocols had only a limited effect on the precision of indel detection. As a result, the improvement in overall precision for Protocol 1, 3 and 4 was mainly due to a more precise detection of SNVs. It can be concluded that applying no or only one pre-processing step (Protocol 0, 2 and 5, see Figure 6) resulted in a low overall precision. Furthermore, Protocol 3 and 4 resulted in the highest overall precision and seem to be the most appropriate pre-processing protocols for an accurate somatic variant calling. Both Protocol 3 and 4 include GATK's SplitNCigarReads and BaseRecalibrator. In addition, Protocol 3 also includes IndelRealigner. Nevertheless, the additional application of the IndelRealigner tool did not enhance the precision of indel detection for Protocol 3 compared to Protocol 4 (Figure 8).



	Protocol 0	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
■ Variants FF	8254	4071	5911	3717	2317	7088
■ SNVs FF	5887	2886	5017	2517	1534	4186
■ Indels FF	2367	1185	894	1200	783	2902

Figure 7. Histogram depicting the number of variants, SNVs and indels obtained using the different pre-processing protocols depicted in Figure 6, based on GSNAP alignment and MuTect2 somatic variant calling for the FF sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

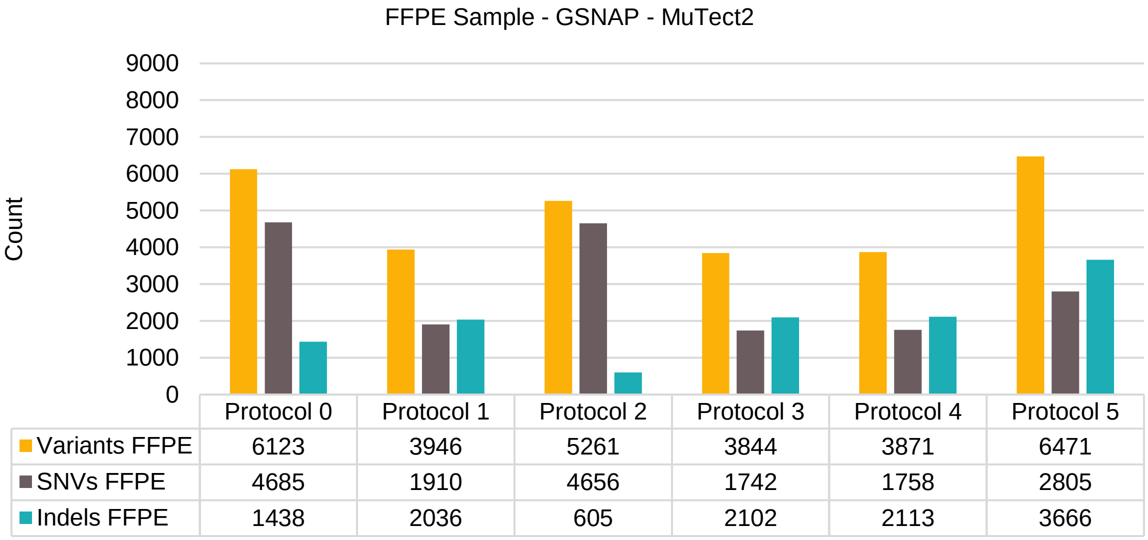


	Protocol 0	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
● Precision FF	0,0276	0,0651	0,0384	0,0716	0,0708	0,0365
● Precision SNVs FF	0,0377	0,0891	0,0439	0,1025	0,1023	0,0602
● Precision Indels FF	0,0025	0,0068	0,0078	0,0067	0,0089	0,0024

Figure 8.Graph depicting the precision obtained using the different pre-processing protocols depicted in Figure 6, based on GSNAP alignment and MuTect2 somatic variant calling for the FF sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

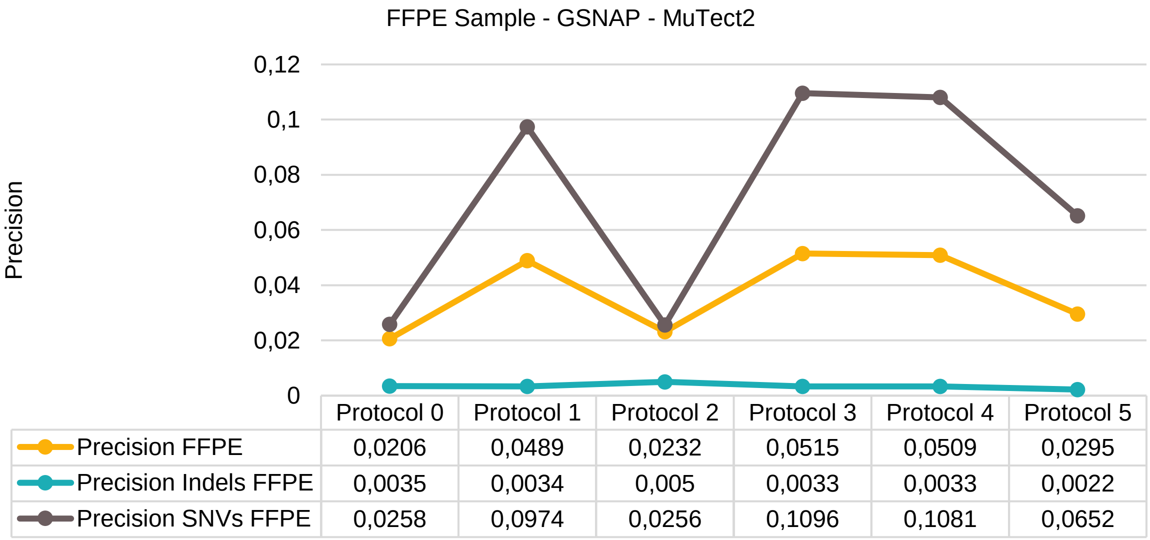
FFPE Sample

The same workflow was applied to the FFPE data. Protocol 0, 2 and 5 (see Figure 6) identified a higher number of variants (Figure 9). Again, this higher number of variants resulted in a lower precision (Figure 10). Protocol 1, 3 and 4 (see Figure 6) detected a slightly higher number of indels than SNVs, however, the precision of indel detection remained very low for all protocols. Similar to the FF data, Protocol 3 and 4 (see Figure 6) had the highest precision and seem to be the best pre-processing protocols. The improvement in overall precision was mainly due to an enhancement in the precision of SNV detection, since only a limited effect was observed for the precision of indel detection when different pre-processing protocols were applied. Generally, the overall precision was higher for the detection of somatic variants in the FF sample.



	Protocol 0	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
■ Variants FFPE	6123	3946	5261	3844	3871	6471
■ SNVs FFPE	4685	1910	4656	1742	1758	2805
■ Indels FFPE	1438	2036	605	2102	2113	3666

Figure 9. Histogram depicting the number of variants, SNVs and indels obtained using the different pre-processing protocols depicted in Figure 6, based on GSNAP alignment and MuTect2 somatic variant calling for the FFPE sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

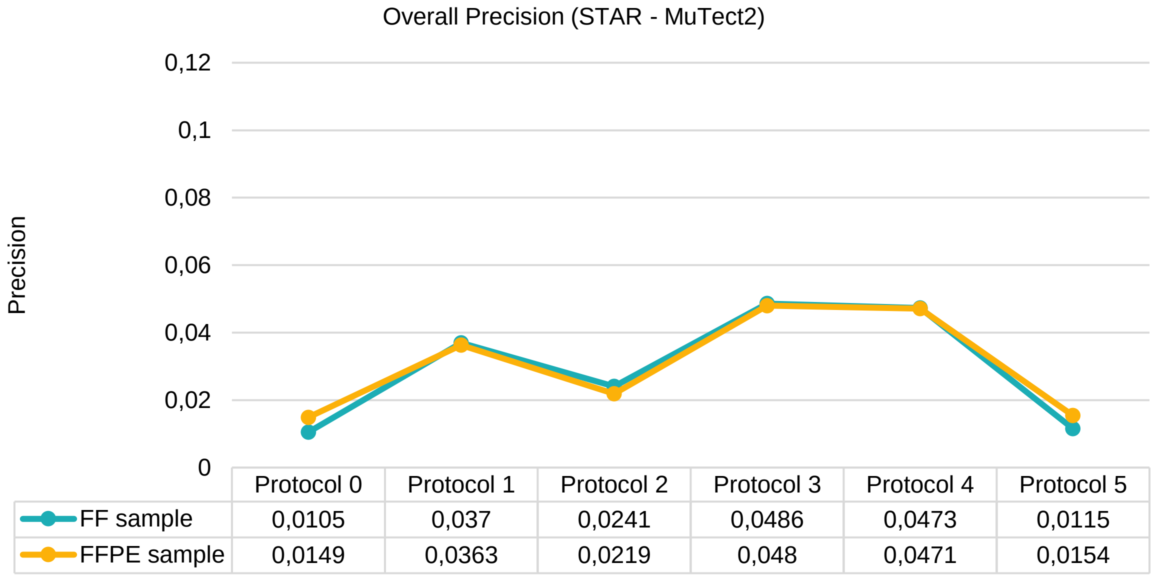


	Protocol 0	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
●Precision FFPE	0,0206	0,0489	0,0232	0,0515	0,0509	0,0295
●Precision Indels FFPE	0,0035	0,0034	0,005	0,0033	0,0033	0,0022
●Precision SNVs FFPE	0,0258	0,0974	0,0256	0,1096	0,1081	0,0652

Figure 10. Graph depicting the precision obtained using the different pre-processing protocols depicted in Figure 6, based on GSNAP alignment and MuTect2 somatic variant calling for the FFPE sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

3.1.2 STAR

The same analysis was performed using STAR as alignment algorithm. Since these results also indicated that Protocol 3 and 4 outperformed the other pre-processing protocols, only an overview of the overall precision for both sample types is depicted in Figure 11. More detailed results can be found in Appendix 8.6. For both sample types, the overall precision was in the same range. Protocol 3 and 4 obtained a higher precision for the variant calling of somatic variants. It should be noted however, that the overall precision for STAR is lower than the overall precision obtained using GSNAP.



	Protocol 0	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
FF sample	0,0105	0,037	0,0241	0,0486	0,0473	0,0115
FFPE sample	0,0149	0,0363	0,0219	0,048	0,0471	0,0154

Figure 11. Graph depicting the overall precision of variant calling using the different pre-processing protocols, based on STAR alignment and MuTect2 somatic variant calling for both sample types. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

3.1.3 Protocol 3 versus Protocol 4

To verify whether Protocol 3 and 4 identified the same variants, the overlap between these VCF files was analysed. Table 4 and Table 5 give an overview of both protocols and their overlap for both sample types. Only variants that have passed all built-in quality filters from MuTect2 were considered. Figure 12 summarises the overlapping variants in Venn diagrams.

Table 4. Overlapping variants between Protocol 3 and Protocol 4 using GSNAP or STAR alignment and MuTect2 somatic variant calling for the FF sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

Aligner	Type	Protocol 3	Protocol 4	Overlap
GSNAP	Variants	3717	2317	2184
	SNVs	2517	1534	1425
	Indels	1200	783	759
STAR	Variants	5655	5705	4533
	SNVs	4591	4639	3595
	Indels	1064	1066	938

Table 5. Overlapping variants between Protocol 3 and Protocol 4 using GSNAP or STAR alignment and MuTect2 somatic variant calling for the FFPE sample. Only variants that have passed the built-in quality filters from MuTect2 were taken into account.

Aligner	Type	Protocol 3	Protocol 4	Overlap
GSNAP	Variants	3844	3871	3597
	SNVs	1742	1758	1641
	Indels	2102	2113	1956
STAR	Variants	3936	3988	3494
	SNVs	2384	2411	2067
	Indels	1552	1577	1427

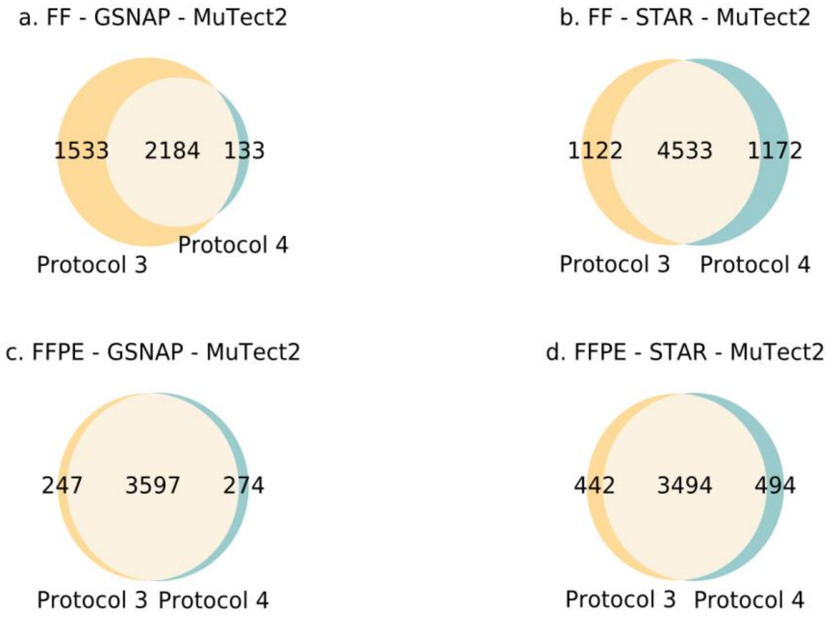


Figure 12. Overlap between Protocol 3 and Protocol 4. (a) Overlap for the FF sample, GSNAP alignment and MuTect2 somatic variant calling. (b) Overlap for the FF sample, STAR alignment and MuTect2 somatic variant calling. (c) Overlap for the FFPE sample, GSNAP alignment and MuTect2 somatic variant calling. (d) Overlap for the FFPE sample, STAR alignment and Mutect2 somatic variant calling.

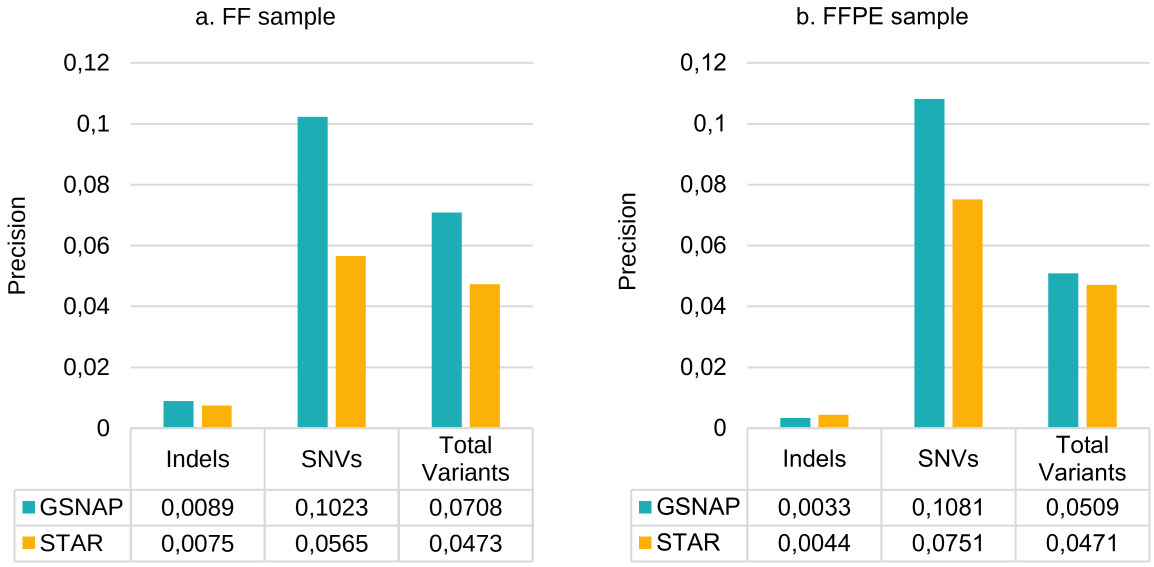
Figure 12 shows a high concordance between variants called by either Protocol 3 or Protocol 4. Nevertheless, a small portion of the variants seem to be unique either for Protocol 3 or 4. This can be the result of the IndelRealigner tool used in Protocol 3. Figure 12a shows a larger fraction of Protocol 3 unique variants, this is not surprising since for this workflow Protocol 3 identified a higher number of variants (3717) than Protocol 4 (2317). Despite these difference, the precision of both protocols were in the same range: 0.0716 for Protocol 3 and 0.0708 for Protocol 4. For the other workflows (Figure 12b, c, d), the number of called variants was more in the same range and, therefore, a larger overlap between Protocol 3 and 4 was observed. For example, Figure 12c depicts a large overlap between Protocol 3 and 4. Only 247 variants (101 SNVs and 146 indels) were 'Protocol 3 unique' and only 274 variants (117 SNVs and 157 indels) were 'Protocol 4 unique'. From these 247 'Protocol 3 unique' variants only 1 SNV was supported by WES and from the 274 'Protocol 4 unique' variants no variants could be verified in WES data (Supplementary Figure 6c). Similar results were found for the other workflows (Appendix 8.7).

3.1.4 Conclusion

The workflows analysed above indicate that Protocol 3 and Protocol 4 obtained the highest overall precision for the detection of somatic variants. Protocol 3 consisted of three pre-processing steps: SplitNCigarReads, IndelRealigner and BaseRecalibrator. Protocol 4, on the other hand, only consisted of two pre-processing steps: SplitNCigarReads and BaseRecalibrator. It can be concluded that applying SplitNCigarReads and BaseRecalibrator in this order had a positive effect on the overall precision of variant calling when MuTect2 was used. Therefore, it is considered advantageous to apply these pre-processing steps in order to enhance the accuracy of somatic variant calling in RNA-Seq data. IndelRealigner, on the other hand, only had a limited effect since no vast improvement was observed in the precision of indel detection. A high overlap was observed between Protocol 3 and Protocol 4, nevertheless, both protocols identified some unique variants. However, only a limited number of these unique variants were supported by WES (Appendix 8.7). To select the optimal pre-processing protocol, one should consider the variant calling algorithm that will be used. However, since MuTect2, VarDict and Strelka2 perform a realignment step in their algorithm, Protocol 4 can be selected as the most appropriate pre-processing protocol.

3.2 Performance Evaluation of the Alignment Algorithms

The next step in the optimisation of the bioinformatics pipeline is the evaluation of the alignment algorithms GSNAP and STAR. Since MuTect2 will be used for somatic variant calling, it is appropriate to use Protocol 4 for pre-processing of the alignment files. Nevertheless, the same analysis was performed for Protocol 3, these results can be found in Appendix 8.8. Figure 13 gives an overview of the precision for the detection of all variants, SNVs and indels for both sample types when Protocol 4 and MuTect2 were applied. These histograms indicate that alignment with GSNAP resulted in a higher precision than if STAR was used, particularly SNVs were called with a higher precision. This can be related to the fact that STAR generally called more variants than GSNAP which resulted in a larger false positive fraction. For example, when looking at variants identified in the FF sample using Protocol 4 and MuTect2, GSNAP alignment resulted in 2317 variants (Figure 7) and STAR in 5705 variants (Supplementary Figure 3). For both GSNAP and STAR the precision of SNV detection was higher than for indel detection. For the FFPE sample (Figure 13b) only a small difference in overall precision was observed, nevertheless, SNV precision was much higher for GSNAP alignment. The explanation for this is the fact that for the FFPE sample GSNAP alignment resulted a large indel/SNV ratio (1.20) which subsequently led to a lowered overall precision. For the STAR alignment the indel/SNV ratio was lower (0.65) which resulted in reduced effect of the low indel precision. For the FF sample this effect was less pronounced due to lower indel/SNV ratios: 0.51 for GSNAP and 0.23 for STAR alignment.



a. FF sample

	Indels	SNVs	Total Variants
■ GSNAP	0,0089	0,1023	0,0708
■ STAR	0,0075	0,0565	0,0473

b. FFPE sample

	Indels	SNVs	Total Variants
■ GSNAP	0,0033	0,1081	0,0509
■ STAR	0,0044	0,0751	0,0471

Figure 13. Precision of somatic variant calling for all variants, SNVs and indels. The alignment tools used were GSNAP and STAR and the somatic variant calling was performed using MuTect2 after pre-processing with Protocol 4. Only variants that have passed built-in quality filters from MuTect2 were taken into account. (a) FF sample. (b) FFPE sample.

To verify whether GSNAP and STAR alignment resulted in the identification of the same somatic variants, again the overlap between the VCF file was calculated. Figure 14 gives information on the overlap between called variants, SNVs, indels and variants supported by WES (DNA concordant). It appears that only a limited fraction of the variants was called by both alignment tools. This is surprising since in both cases the same pre-processing steps and variant calling tool were used. To further investigate this difference, the unique fractions in Figure 14a were annotated and analysed. An overview of the most severe consequences of the variants uniquely identified by GSNAP and STAR can be found in Figure 15. It was observed that the alignment algorithm used seem to affect the type of variants identified. GSNAP alignment mainly identified splice site variants (50.8%), frameshift variants (20%), missense variants (17.3%) and a smaller fraction of synonymous variants (8%). STAR alignment, on the other hand, mainly identified missense variants (53%), synonymous variants (22.1%) and a smaller fraction of frameshift variants (10.7%). The inconsistent identification of somatic variants may be attributed to the algorithmic differences between GSNAP and STAR. Further research involving more samples is required to further investigate this discrepancy, however, this was out of scope of this master thesis.

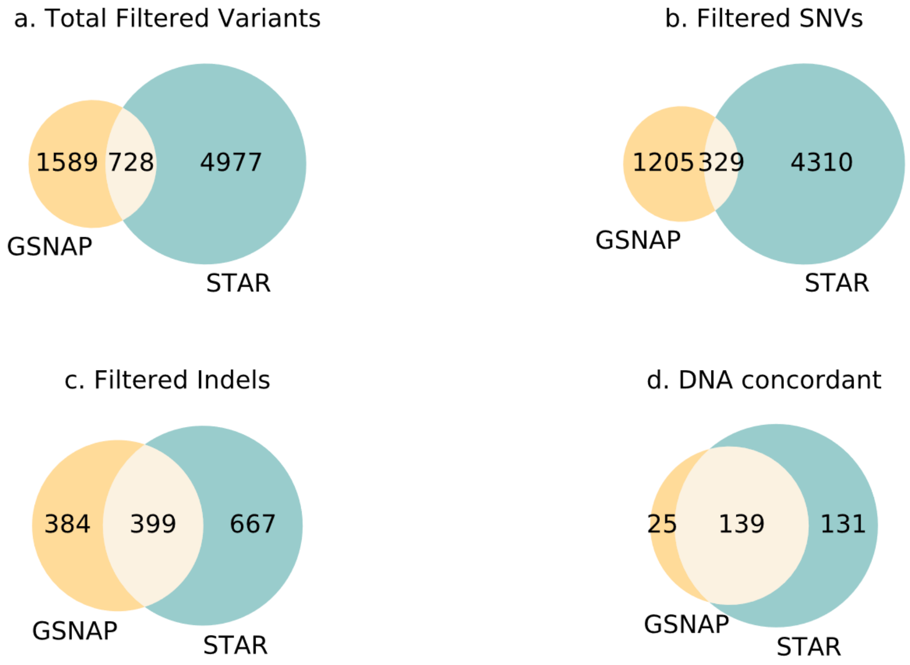


Figure 14. Venn diagrams depicting the overlap between somatic variants called using either GSNAP or STAR alignment for both sample types when pre-processing Protocol 4 and MuTect2 somatic variant calling were used. (a) Venn diagram for all variants that have passed built-in quality filters from MuTect2. (b) Venn diagram for SNVs that have passed built-in quality filters from MuTect2.(c) Venn diagram for all indels that have passed built-in quality filters from MuTect2. (d) Venn diagram for variant calls that were supported by WES (DNA concordant).

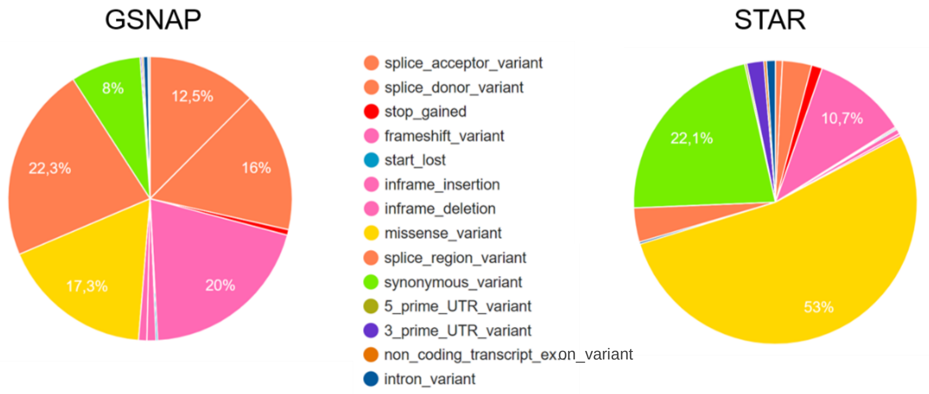


Figure 15. Most severe consequences of somatic variants called by MuTect2 after GSNAP or STAR alignment, only aligner unique variants from Figure 14a were considered. The VCF files were annotated using VEP. For GSNAP 1589 variants were taken into account, for STAR 4977.

3.3 Performance Evaluation of the Variant Calling Algorithms

In this section, three variant calling algorithms are analysed for both sample types: MuTect2, VarDict and Strelka2. Here, only results for GSNAP alignment and pre-processing Protocol 4 are shown. For the sake of completeness, Appendix 8.10 provides detailed results on the GSNAP alignment combined with pre-processing Protocol 3 and Appendix 8.11 for the STAR alignment combined with pre-processing Protocol 4.

3.3.1 MuTect2

Table 6 gives an overview of the number of called variants using MuTect2 and the resulting precision for the FF sample when GSNAP alignment and pre-processing Protocol 4 were applied. A large fraction (73.22%) of the variants was rejected by built-in quality filters from MuTect2 (Table 1, section 2.2.3) resulting in 2317 variants, including 1534 SNVs and 783 indels. Applying these filters had a positive effect on the precision, particularly on the precision of SNV detection (0.1023). Nevertheless, an overall precision of 0.0708 is considerably low. This low overall precision is partly attributable to the high fraction of DNA discordant indels which resulted in a poor precision for the detection of indels (0.0089). As for the FF sample, somatic variant calling of the FFPE sample revealed a large fraction (91.93%) of called variants that was rejected by built-in quality filters from MuTect2 (Table 1, section 2.2.3) resulting in 3871 variants, including 1758 SNVs and 2113 indels (Table 7). Again, applying these quality filters had a positive effect on the precision, particularly on the SNV precision (0.1081). Nevertheless, due to the low precision of indel detection (0.0033), the overall precision (0.0509) remained considerably low. Compared to the FF sample, a lower overall precision was obtained for the FFPE sample, however, the precision of SNV detection was in the same range.

Similar results were found when pre-processing Protocol 3 (Appendix 8.10) was applied. As already described in section 3.2, when STAR was used as alignment tool (Appendix 8.11), a lower overall precision was obtained.

Table 6. Number of (filtered) variants, SNVs and indels called for the FF sample when GSNAP alignment, pre-processing Protocol 4 and MuTect2 somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	**SNVs**	**Indels**	**DNA Concordant Variants**	**DNA Concordant SNVs**	**DNA Concordant Indels**
No Filters	8652	6023	2629	226 (0.0261)	201 (0.0334)	27 (0.0095)
MuTect2 Filters	2317 (73.22%)	1534 (74.53%)	783 (70.22%)	164 (0.0708)	157 (0.1023)	7 (0.0089)

Table 7. Number of (filtered) variants, SNVs and indels called for the FFPE sample when GSNAP alignment, pre-processing Protocol 4 and MuTect2 somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	**SNVs**	**Indels**	**DNA Concordant Variants**	**DNA Concordant SNVs**	**DNA Concordant Indels**
No Filters	43903	15841	28062	308 (0.0070)	278 (0.0175)	30 (0.0011)
MuTect2 Filters	3871 (91.93%)	1758 (88.90%)	2113 (92.47%)	197 (0.0509)	190 (0.1081)	7 (0.0033)

3.3.2 VarDict

Table 8 gives an overview of the number of variants called using VarDict and the resulting precision for the FF sample when GSNAP alignment and pre-processing Protocol 4 were applied. VarDict called more variants than MuTect2, even after applying built-in quality filters from VarDict 59536 variants, including 49642 SNVs and 9894 indels, were retained. Due to the limited number of DNA concordant variants, the overall precision of variant calling remained very low (0.0051). More variants were identified in the FFPE data (99384) and the overall precision was only 0.0020 after applying the standard quality filters from VarDict (Table 9). Somatic variant calling using VarDict appeared to be less precise for the FFPE sample than for the FF sample.

Table 8. Number of (filtered) variants, SNVs and indels called for the FF sample when GSNAP alignment, pre-processing Protocol 4 and VarDict somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	**SNVs**	**Indels**	**DNA Concordant Variants**	**DNA Concordant SNVs**	**DNA Concordant Indels**
No Filters	218560	193571	24989	341 (0.0016)	280 (0.0014)	61 (0.0024)
VarDict Filters	59536 (72.76%)	49642 (74.35%)	9894 (60.41%)	305 (0.0051)	249 (0.0050)	56 (0.0057)
Additional Filters	3303 (98.49%)	750 (99.61%)	2553 (89.78%)	110 (0.0333)	104 (0.1387)	6 (0.0024)

Table 9. Number of (filtered) variants, SNVs and indels called for the FFPE sample when GSNAP alignment, pre-processing Protocol 4 and VarDict somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	SNVs	Indels	DNA Concordant Variants	DNA Concordant SNVs	DNA Concordant Indels
No Filters	525069	368312	156757	237 (0.0005)	194 (0.0005)	43 (0.0003)
VarDict Filters	99384 (81.07%)	68431 (81.42%)	30953 (80.25%)	201 (0.0020)	167 (0.0024)	34 (0.0011)
Additional Filters	8599 (98.36%)	3749 (98.98%)	4850 (96.91%)	83 (0.0097)	76 (0.0203)	7 (0.0014)

To enhance precision, additional filters described in Table 2 (section 2.2.3) were applied. Figure 16 and Figure 17 summarise the implementation of these additional filters for a workflow comprising GSNAP alignment and pre-processing Protocol 4 of the FF data. When the seven additional filters were applied, the number of variants decreased from 59536 to a total of 3303 (Figure 16), including 750 SNVs and 2253 indels (Table 8). The overall precision increased from 0.0051 to 0.0333 (Figure 17). The most pronounced enhancement was found for the precision of SNV detection: from 0.0050 to 0.1387. The precision of indel detection, on the other hand, did not increase. It can be concluded that additional filtering resulted in a more precise detection of SNVs but further optimisation remains necessary for indels. The results for the FFPE sample are summarised in Table 9.

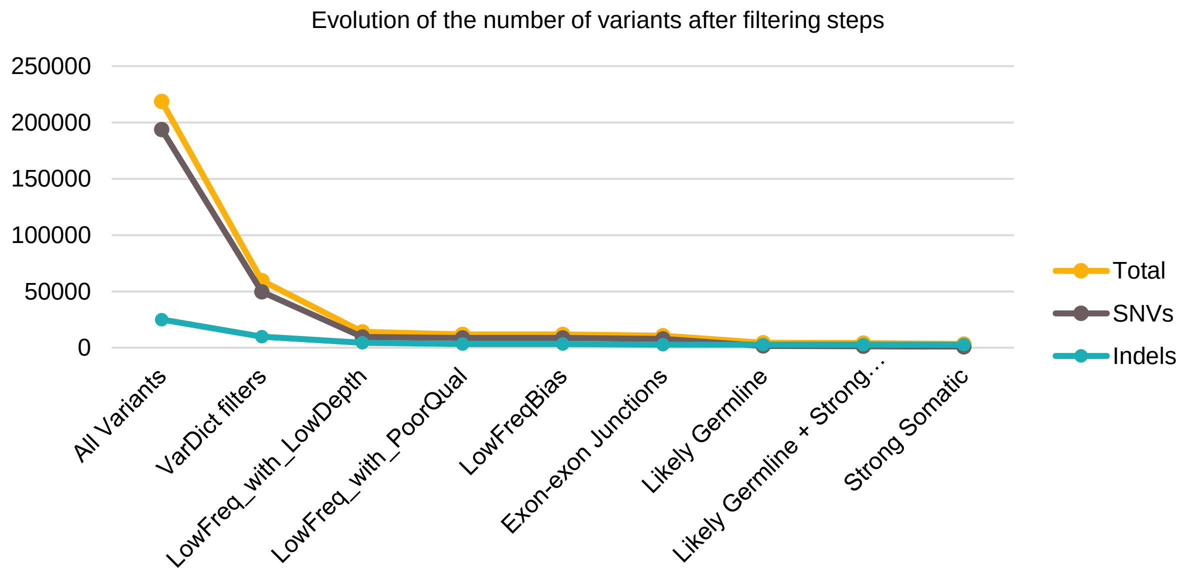


Figure 16. The evolution of the number of variants, including SNVs and indels, after applying built-in quality filters from VarDict and seven additional filters (Table 2, section 2.2.3). The workflow evaluated comprised GSNAP alignment, pre-processing Protocol 4 and VarDict somatic variant calling of the FF data.

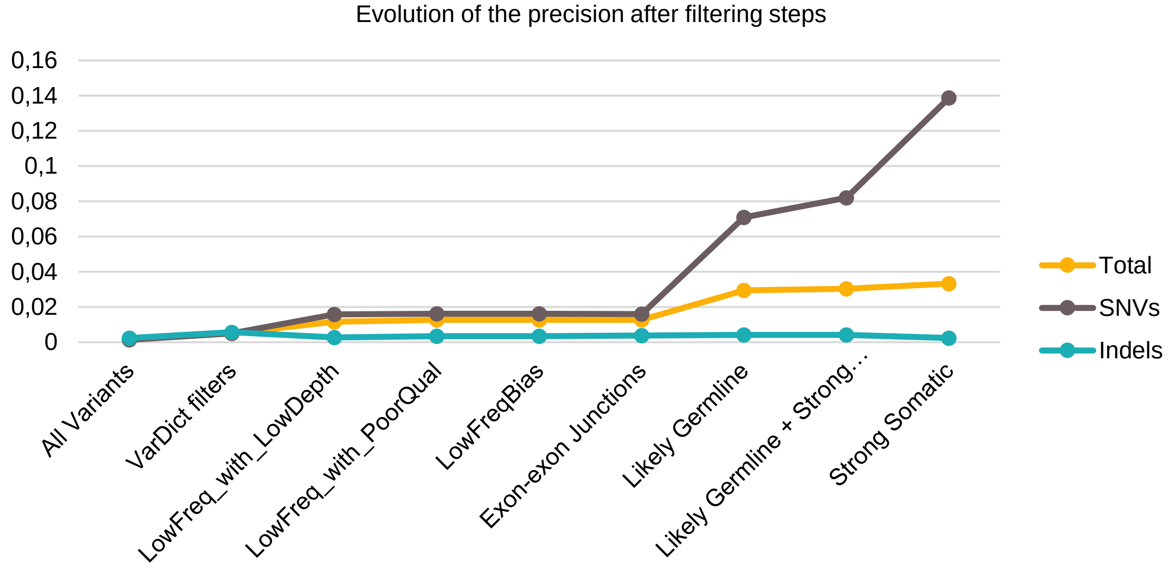


Figure 17. The evolution of the overall precision of variant detection, including SNV and indel precision, after applying built-in quality filters from VarDict and seven additional filters (Table 2, section 2.2.3). The workflow evaluated comprised GSNAP alignment, pre-processing Protocol 4 and VarDict somatic variant calling of the FF data.

3.3.3 Strelka2

Table 10 gives an overview of the number of variants called using Strelka2 and the resulting precision for the FF sample. A large fraction (94.08%) of the variants was rejected by built-in quality filters from Strelka2 (Table 3, section 2.2.3) resulting in a remaining 12602 variants, including 10498 SNVs and 2104 indels. Nevertheless, application of these standard quality filters had only a limited effect on the overall precision (0.0145). Again, a considerably low precision of indel detection was observed (0.0048). To further enhance the precision it would be possible to apply additional filters similar to the VarDict filters. However, no guidelines were provided for the implementation of additional filters. Besides, the VCF files produced by Strelka2 did not provide similar quality metrics to VarDict. For example, Strelka2 did not include a NM (number of mismatches in reads) and SBF (Strand Bias Fisher p-value) info field in the VCF file. As a result, no additional filters were applied to the Strelka2 variants. For the FFPE sample likewise a large fraction of the variants was rejected by Strelka2's standard quality filters (Table 11). However, the precision remained very low.

Table 10. Number of (filtered) variants, SNVs and indels called for the FF sample when GSNAP alignment, pre-processing Protocol 4 and Strelka2 somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	SNVs	Indels	DNA Concordant Variants	DNA Concordant Variants	DNA Concordant Variants
No Filters	212776	204685	8091	388 (0.0018)	337 (0.0016)	51 (0.0063)
Strelka2 Filters	12602 (94.08%)	10498 (94.87%)	2104 (74.00%)	183 (0.0145)	173 (0.0165)	10 (0.0048)

Table 11. Number of (filtered) variants, SNVs and indels called for the FFPE sample when GSNAP alignment, pre-processing Protocol 4 and Strelka2 somatic variant calling were applied and corresponding number of variants that were supported by WES (DNA concordant). From column 2 to 4, between brackets the percentage of rejected variants in relation to the initial number of called variants can be found. From column 4 to 6, between brackets the precision can be found.

	Variants	SNVs	Indels	DNA Concordant Variants	DNA Concordant SNVs	DNA Concordant Indels
No Filters	493766	447889	45877	468 (0.0009)	417 (0.0009)	51 (0.0011)
Strelka2 Filters	30757 (93.77%)	25702 (94.26%)	5055 (88.98%)	261 (0.0085)	252 (0.0098)	9 (0.0018)

3.3.4 Conclusion

The goal of this master thesis is to implement a bioinformatics workflow to identify somatic variants in tumour tissue with a high precision. In other words, it is important that a somatic variant called using RNA-Seq data, can be verified using WES data (Gold Standard variant set), which means the number of FP should be as low as possible. Therefore, the precision (as calculated in section 2.3.1) was compared in order to select the most appropriate variant calling algorithm. Sensitivity (as calculated in section 2.3.1) was not considered since this metric takes into account the FN variants, which consists of variants called in WES but not in RNA-Seq. However, the Gold Standard variant set was acquired from WES data, hence, it is possible that these Gold Standard variants were not expressed in the tumour tissue. As a result, some of the FN variants might be TN variants as these variants were not covered in RNA-Seq. Therefore, the calculated sensitivity is only a conservative estimate. An overview of the sensitivity for different variant callers can be found in Appendix 8.12.

The precision of different variant calling algorithms for the FF sample is summarised in Figure 18, only filtered variants were considered. MuTect2 somatic variant calling resulted the highest overall precision (0.0708) and the highest precision of indel detection (0.0089). When additional filters were applied, VarDict acquired the highest precision for SNV detection (0.1387). Strelka2, on the other hand, detected variants with the lowest overall precision (0.0145).

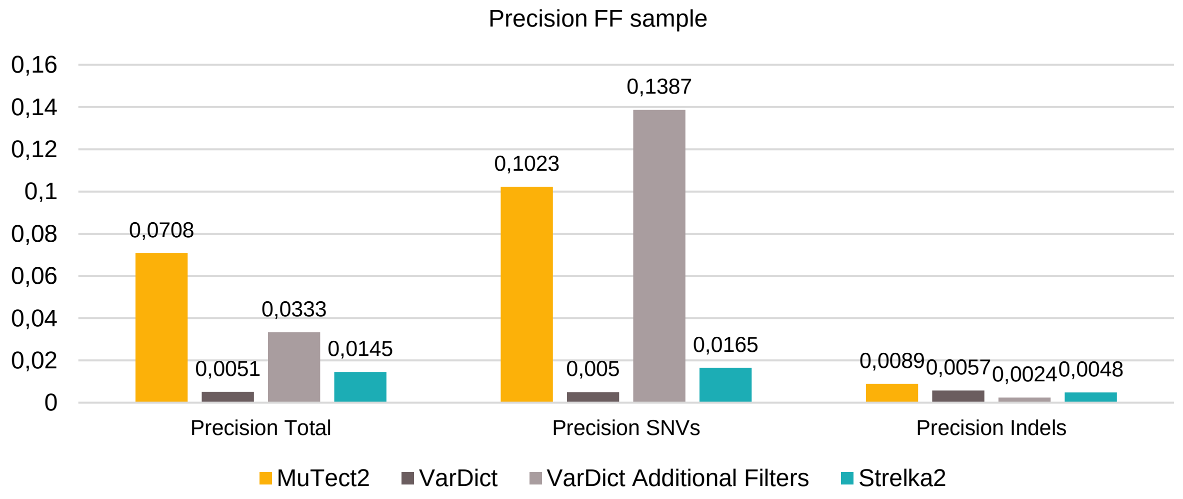


Figure 18. Overview of the precision for MuTect2, VarDict, VarDict with additional filters and Strelka2 for the FF sample. The workflow evaluated comprised GSNAP alignment and pre-processing Protocol 4.

The same comparison was made for the FFPE sample (Figure 19). It appeared that the implementation of additional filters for VarDict did not have a similar improvement as for the FF sample. Moreover, MuTect2 was the only variant caller that obtained a similar high precision for both sample types. Therefore, MuTect2 seems to be the most appropriate variant caller for the both sample types.

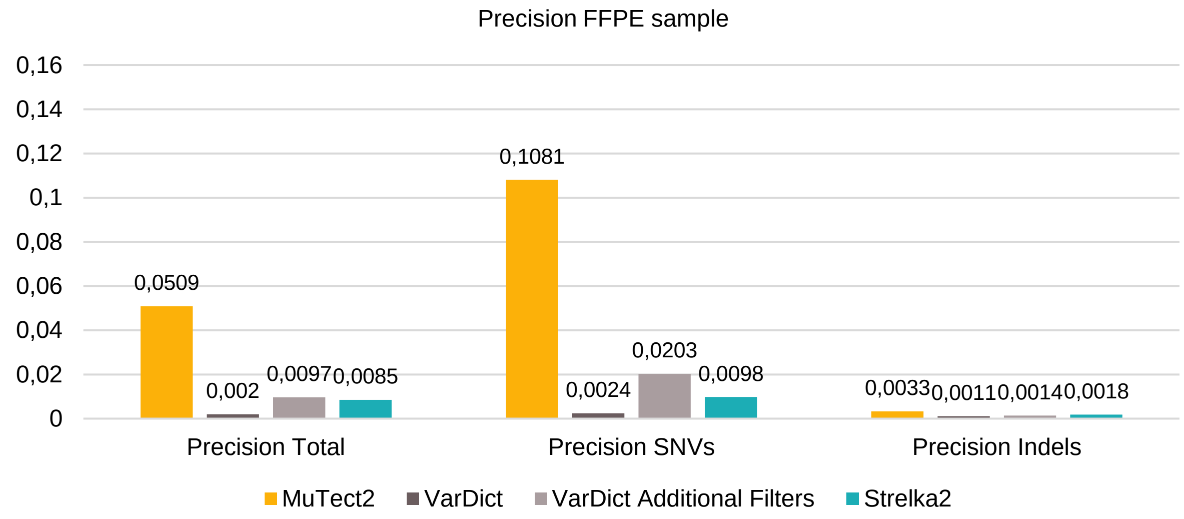


Figure 19. Overview of the precision for MuTect2, VarDict, VarDict with additional filters and Strelka2 for the FFPE sample. The workflow evaluated comprised GSNAP alignment and pre-processing Protocol 4.

For the selection of the most appropriate variant calling algorithm, the precision was considered the decisive performance metric. Each caller alone called many variants that were not validated by DNA somatic variants in the Gold Standard set (discordant calls). MuTect2 provided the least amount of variant calls not supported by WES compared to the other 2 methods and, therefore, MuTect2 seems the most appropriate variant caller. Nevertheless, only 7.08% of variant calls made by MuTect2 were supported by WES. The results in Figure 18 and Figure 19 indicate that any single caller was not adequate in discovering variants with a high precision. Moreover, only a limited number of variants were identified by all three variant callers (Figure 20, Figure 22), which indicated that a large portion of the variant identified by a single variant caller were false positive calls. Therefore, it was tested if the combination of three variant calling algorithms would provide a higher rate of variant calls supported by WES.

3.3.5 Combination MuVaSt

The combination of MuTect2, VarDict with additional filters and Strelka2 (further referred to as MuVaSt) was applied for both sample types, both alignment algorithms and both Protocol 3 and Protocol 4. Next, it was tested if the combination of three variant callers would provide a higher rate of variant calls supported by WES. Table 12 summarises the precision of the MuVaSt somatic variant calling method for both aligners and for both Protocol 3 and Protocol 4 for the FF sample. Table 13 shows the same results for the FFPE sample.

Table 12. Overview of the precision of the MuVaSt somatic variant calling method for the FF sample. It should be noted that only variants that have passed built-in quality filters and VarDict additional filters were taken into account.

	Total Precision	SNV Precision	Indel Precision
Protocol 3			
GSNAP	0.2457	0.7927	0.0090
STAR	0.2647	0.6560	0.0093
Protocol 4			
GSNAP	0.2401	0.8381	0.0109
STAR	0.2584	0.6522	0.0092

Table 13. Overview of the precision of the MuVaSt somatic variant calling method for the FFPE sample. It should be noted that only variants that have passed built-in quality filters and VarDict additional filters were taken into account.

	Total Precision	SNV Precision	Indel Precision
Protocol 3			
GSNAP	0.2215	0.6526	0.0197
STAR	0.1997	0.3284	0.0143
Protocol 4			
GSNAP	0.2231	0.6667	0.0231
STAR	0.2009	0.3283	0.0147

It is remarkable that the precision of indel detection remained very low. This poor indel precision had a negative effect on the overall precision. Therefore, the workflows were further evaluated solely based on the precision of SNV detection. For both sample types, STAR alignment resulted in a lower precision than GSNAP alignment when only SNVs were considered. When GSNAP was used, pre-processing Protocol 4 seemed to perform slightly better than Protocol 3. Moreover, the highest SNV precision was acquired when GSNAP and Protocol 4 were combined for both sample types. To further assess the performance of the MuVaSt somatic variant calling method, this workflow was evaluated based on the differences between variants identified using RNA-Seq or WES. These differences were assessed using gene expression, allele-specific expression, VAF, DP and coverage in WES.

FF Sample

Figure 20a gives an overview of the overlapping variants between MuTect2, VarDict (with additional filters) and Strelka2. It is remarkable that a large fraction of the called variants was unique for a single variant calling tool. For example, Strelka2 identified the highest number of unique variants (11210). Only a small fraction of 379 variants, including 105 SNVs and 274 indels, was identified by all three variant callers. Figure 20b depicts the number of DNA concordant variants for each variant caller. The intersection of the three VCF files resulted in the highest overall precision (0.2401): 91 of 379 variants were found in the Gold Standard variant set. Most of these DNA concordant variants appeared to be SNVs (88), only a minor fraction were indels (3) (Figure 23). Two of these indels were deletions of 1 base pair, the other one was an insertion of 1 base pair. This again indicated that this workflow is not yet suitable for the correct identification of indels. Conversely, a vast improvement was observed for the precision of SNV detection when the MuVaSt somatic variant calling method was employed: a precision of 0.8381 was obtained. From the 105 called SNVs, 88 were DNA concordant and only 17 were RNA unique (Figure 23b).

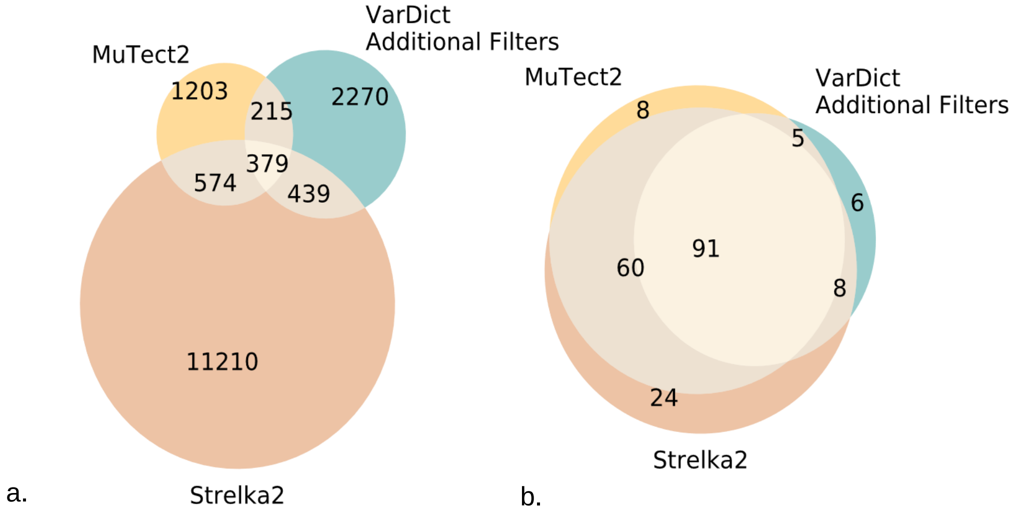


Figure 20. Overview of the overlapping variants between MuTect2, VarDict (with additional filters) and Strelka2 for the FF sample when GSNAP alignment and pre-processing Protocol 4 were applied. (a) This Venn diagram only takes into account variants that have passed the built-in quality filters. (b) This Venn diagram only includes variants that were validated by WES (DNA concordant).

To understand why a large fraction of the variant calls in WES (Gold Standard) were missed by RNA-Seq, the properties of DNA unique variants (1163, see Figure 23a) were analysed for the FF sample. An important reason why variants called in WES were missed by RNA-Seq is the gene expression. Analysis of the gene expression data obtained from RNA-Seq of the FF tumour tissue revealed that 537 of 1163 DNA unique variants (46.2%) had an expression below 1 TPM. These 537 lowly expressed variants consisted of 46 indels and 491 SNVs. Nevertheless, expression of the gene does not automatically imply expression of the variant. To overcome this limitation, the allele-specific expression of the DNA unique SNVs was analysed using ASEReadCounter [113] and expression data from RNA-Seq data of the FF sample. It should be noted that this analysis was limited to SNVs since ASEReadCounter was unable to handle indels correctly. It was observed that from the 1025 DNA unique SNVs only 358 were supported by at least one RNA read and only 309 SNVs had a VAF greater than 0.04. This means that at least 667 DNA unique SNVs (65.1%) were not supported by any RNA reads containing the mutation. It can be concluded that a large fraction of the DNA unique variants consisted of variants that were lowly, or not, expressed. As a result, it would be impossible to detect these variants using RNA-Seq only.

One important metric to consider is the VAF. The VAF denotes the fraction of reads containing the variant allele. Figure 21a depicts the $\mathrm{VAF}_{\mathrm{RNA}}$ (as calculated in section 2.3.4) for the RNA variants that were supported by WES (DNA concordant) and the RNA unique variants. It is noticeable that the $\mathrm{VAF}_{\mathrm{RNA}}$ for DNA concordant variants was higher than the $\mathrm{VAF}_{\mathrm{RNA}}$ of RNA variants that were missed by WES (RNA unique). Although it was possible to identify variants with a low $\mathrm{VAF}_{\mathrm{RNA}}$ in RNA-Seq data, a large fraction of these variants was not supported by WES. For example, it was observed that 83.7% (241/288) of the RNA unique variants had a $\mathrm{VAF}_{\mathrm{RNA}}$ below 0.2. Figure 21b reveals a more narrow peak between 0.17 and 0.4 for the $\mathrm{VAF}_{\mathrm{DNA}}$ distribution of the DNA concordant variants. The $\mathrm{VAF}_{\mathrm{DNA}}$ distribution for the DNA unique variants shows 2 peaks: a first peak between 0 and 0.17 and a second peak between 0.17 and 0.4. These two peaks may be attributed to the clonality of somatic mutations as it was already proven that the $\mathrm{VAF}_{\mathrm{DNA}}$ can be used to distinguish between the subpopulations of tumour cells [114]. Moreover, considering tumour purity and aneuploidy, somatic mutations with a high $\mathrm{VAF}_{\mathrm{DNA}}$ (>0.25) are more likely to be clonal mutations, while somatic mutations with a lower $\mathrm{VAF}_{\mathrm{DNA}}$ are more likely to be subclonal mutations [115]. As a result, the first peak of the VAF$_{DNA}$ (between 0 and 0.17) can be attributed to subclonal mutations; and the second peak (between 0.17 and 0.4) to clonal mutations. Consequently, based on the VAF$_{DNA}$ a high fraction of the DNA concordant variants may be considered clonal mutations. Correct identification of the subclonal mutations, on the other hand, appeared to be more difficult.

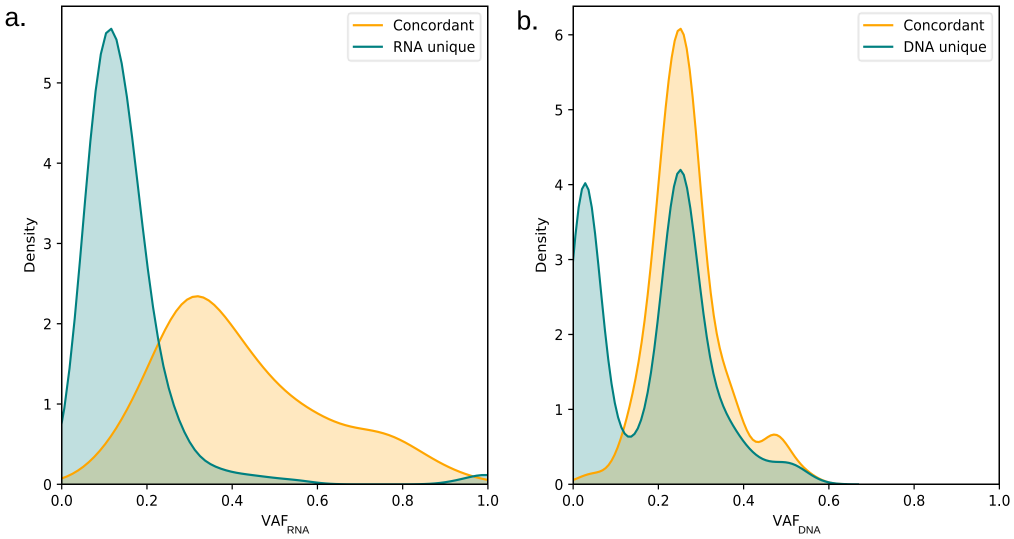


Figure 21. Distribution plots of the variant allele frequency (VAF) for the DNA concordant and discordant fractions for the FF sample when the MuVaSt variant calling method was used (GSNAP and Protocol 4). (a) VAF$_{RNA}$ for the RNA variants as calculated in section 2.3.4. The blue graph depicts the RNA variants that were not supported by WES, the yellow graph depicts variants that were supported by WES. (b) VAF$_{DNA}$ for the DNA variants as calculated in section 2.3.4. The blue graph depicts the DNA variants that were not supported by RNA-Seq, the yellow graph depicts variants that were supported by RNA-Seq.

Another important metric to consider is the read depth (DP). The average DP observed was 328, moreover, 92.2% of the RNA-Seq variants had a read depth of more than 10 reads. Due to the high sequencing coverage, no significant difference was observed between concordant and discordant variants (Appendix 8.13).

As it was already described in section 1.2.3, RNA-Seq has the potential to detect variants that could be missed by WES. To understand why a large fraction of the MuVaSt variants calls were missed by WES, these variants were analysed for the FF sample. WES only covers a target region of known genes and their flanking regions. Therefore, it is possible that some of the RNA unique variants were not identified in WES because they are not included in the exome target region. Indeed, using a BED file containing the SureSelect Human All Exon V6 (Agilent) capture regions, it was observed that only 15 (2 SNVs and 13 indels) of the 288 RNA unique variants were covered in WES. As a result, it is possible that these RNA unique variants contain true positive variants that cannot be detected using WES. The 288 RNA unique variants consisted of 17 SNVs and 271 indels (Figure 23). From the 17 SNVs, 1 A to G substitution and 4 T to C substitutions were observed. It is possible that these SNVs include some unknown RNA editing sites (section 1.2.3) that cannot be identified using WES. Nevertheless, analysis of a larger number of samples should be performed to prove this hypothesis. The indel fraction included 264 insertions of 1 base pair, 7 deletions of 1 base pair, 1 deletion of 29 base pairs, 1 deletion of 42 base pairs and 1 deletion of 43 base pairs. However, it is important to note that some of these RNA unique variants may also be attributed to artefacts originating from RNA-Seq related technical errors.

As described before, only a limited fraction of variants with a VAF_{RNA} below 0.2 was supported by WES (Figure 21a). This finding suggest that it might be beneficial to apply a threshold VAF_{RNA} value of 0.2 to enhance the precision of variant calling. When this threshold was applied, 130 of 379 called RNA variants were retained (34.3%) of which 83 variants, including 82 SNVs and 1 indel, were supported by WES. As a result, implementation of this additional VAF threshold resulted in an increase of overall precision from 0.2401 to 0.6385. The 130 RNA variants consisted of 92 SNVs and 38 indels which means 12.4% of the SNVs and 86.1% were rejected by the VAF_{RNA} threshold. This results in a precision of 0.8913 for the detection of SNVs and a precision of 0.0263 for the detection of indels. It is important to note that this threshold was adapted arbitrarily and solely based on this sample, therefore, it might be overfitted. As a result, a threshold VAF_{RNA} of 0.2 might not be useful for other samples with different tumour purities.

FFPE Sample

The MuVaSt variant calling method was also applied to the FFPE sample. Figure 22 gives an overview of the overlapping variants between MuTect2, VarDict (with additional filters) and Strelka2. Again, a large number of variants was called by only one variant caller. Only 251 variants were identified by all three variant calling tools of which 56 were also supported by the Gold Standard variant set resulting in an overall precision of 0.2231. The 56 DNA concordant variants consisted of 52 SNVs and only 4 indels (Figure 23). Consequently, the precision for SNV detection was 0.6667 and only 0.0231 for the detection of indels.

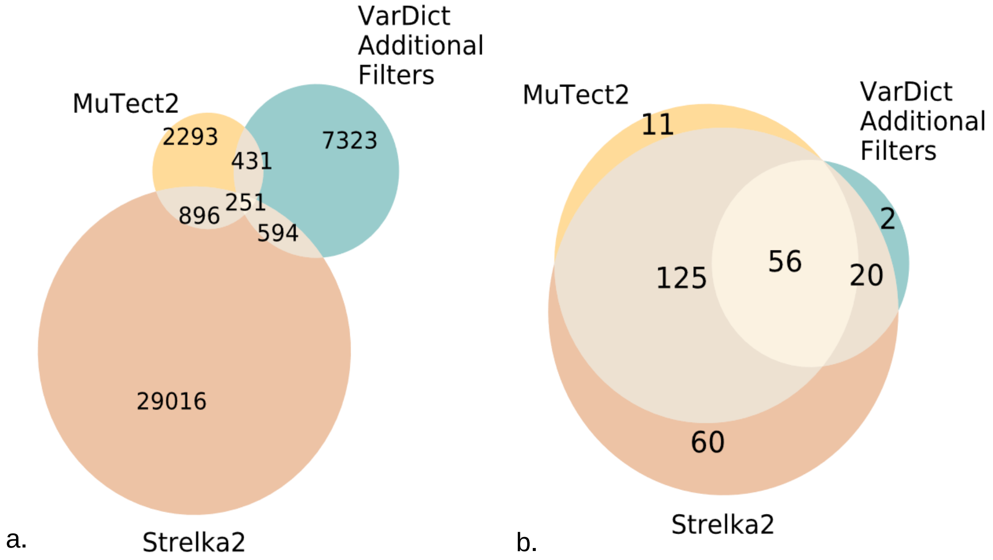


Figure 22. Overview of the overlapping variants between MuTect2, VarDict (with additional filters) and Strelka2 for the FFPE sample when GSNAP alignment and pre-processing Protocol 4 were applied. (a) This Venn diagram only takes into account variants that have passed the built-in quality filters. (b) This Venn diagram only includes variants that were validated by WES (DNA concordant).

Gene expression analysis of DNA unique variants indicated that 545 of 1198 DNA unique variants (45.5%) had an expression below 1 TPM. These 545 unexpressed variants consisted of 46 indels and 499 SNVs. Evaluation of the allele-specific expression revealed that from the 1061 DNA unique SNVs only 364 were supported by at least one RNA read and only 331 had a VAF greater than 0.04. As a result, at least 697 DNA unique SNVs (65.7%) were not supported by any RNA read containing the mutation. Similar to the FF sample, a large fraction of the DNA unique variants showed no, or only low, expression in the RNA-Seq data.

Analysis of the RNA variants missed by WES indicated that only 21 (7 SNVs and 14 indels) of the 195 RNA unique variants were covered in the SureSelect Human All Exon V6 (Agilent) target regions. The remaining 186 RNA variants possibly contain true positive variants that cannot be detected using WES. The 195 RNA unique variants consisted of 26 SNVs and 169 indels (Figure 23). From the 26 SNVs, only 2 A to G substitution and 2 T to C substitutions were observed. Again, these SNVs might contain some unknown RNA editing sites. The indel fraction included 145 insertions of 1 base pair, 3 deletions of 1 base pair, 6 deletion of 2 base pairs, 13 deletion of 3 base pairs and 2 deletion of 4 base pairs.

Analysis of the VAF for both RNA and DNA variants showed similar results as for the FF sample (Appendix 8.14). Application of the threshold VAF_{RNA} of 0.20 reduced the number of called variants from 251 to 79 (31.5% rejected). Since 53 of these 79 variants were supported by WES, the overall precision was improved to 0.6709. The 79 variants include 58 SNVs of which 50 are supported by WES and 21 indels of which only 3 were supported by WES. As such, a precision of 0.8621 was obtained for the detection of SNVs and a precision of 0.1429 for the detection of indels.

Figure 23 depicts the overlap between variants identified in the FF sample and FFPE sample using the MuVaSt variant calling method and the overlap with the Gold Standard variant set. The overlap between these three groups of variants is more pronounced for the SNVs than for the indels. Two SNVs were detected in both sample types but not supported by WES. These 2 SNVs include one A to G and one A to C transition. Since the A to G mutation at location chr1:33270581 was found in both samples, this SNV potentially represents an unknown RNA editing site. For the indels, on the other hand, 38 indels were identified in both samples but not included in the Gold Standard set. These 38 indels consisted of 38 insertions of 1 base pair.

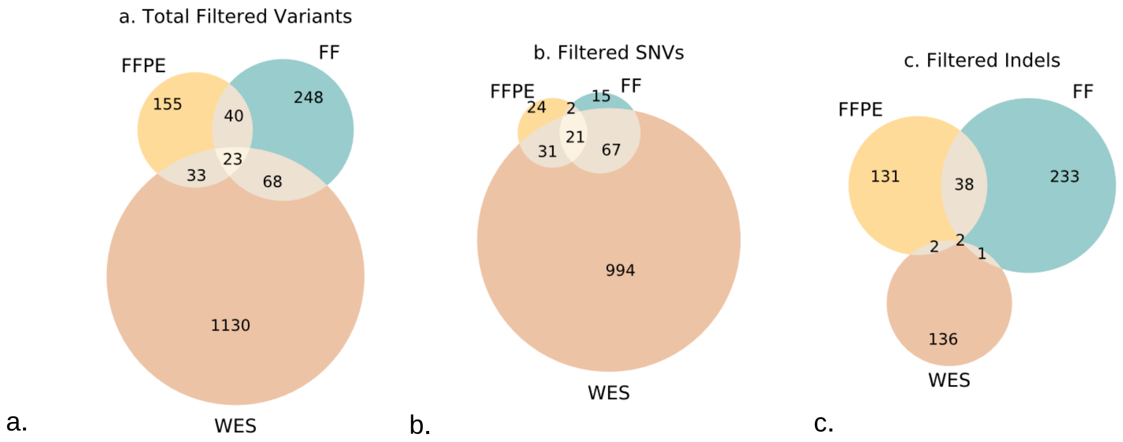


Figure 23. Venn diagrams of the overlap between variants identified using the MuVaSt somatic variant calling method (after GSNAP alignment and pre-processing Protocol 4) and somatic variants supported by WES (Gold Standard). Both sample types were considered. (a) Overlap when all variants were taken into account. (b) Overlap when only SNVs were taken into account. (c) Variants when only indels were taken into account.

Figure 24 depicts the nucleotide substitutions for both sample types. It appears that for the FFPE sample a higher percentage of C to T and G to A substitutions were found. The enrichment of these nucleotide substitutions can probably be assigned to mutational artefacts specific for FFPE sample preservation [49].

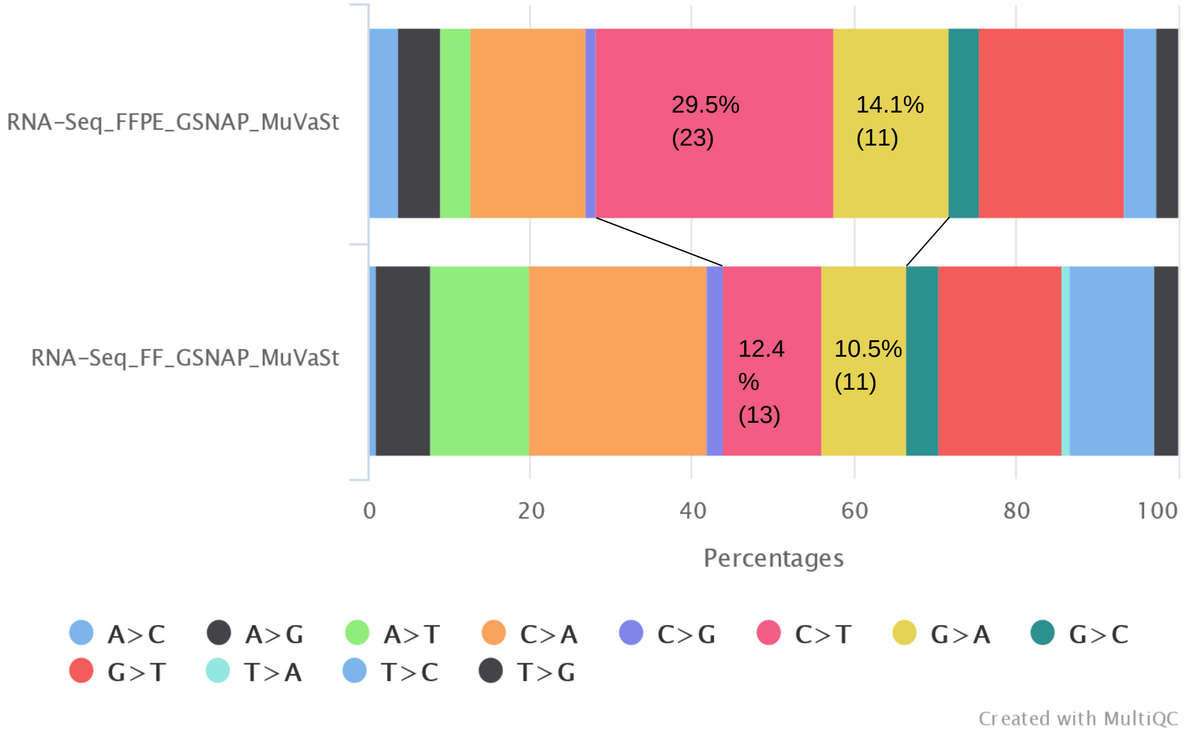


Figure 24. Percentages of identified nucleotide substitutions identified using the MuVaSt somatic variant calling method for both the FFPE and FF sample.

3.4 Validation

To validate this bioinformatics pipeline, tumour tissue originating from a lung adenocarcinoma of a second patient was analysed. RNA sequencing data was obtained from an FFPE sample with an estimated tumour purity of 40%. The FFPE data consisted of 65101981 x 2 reads. The quality of this dataset was assessed using FastQC [109]. Sequencing data was aligned using GSNAP, mapped reads were sorted, duplicate reads were marked, read groups were added, pre-processing Protocol 4 was applied, and the somatic variant calling was performed using the MuVaSt method.

Figure 25a depicts the overlap between variants identified using MuTect2, VarDict and Strelka2. Similar as before, it was observed that Strelka2 called the highest number of variants. Analysis of the overlap between the variant callers revealed 191 variants, including 69 SNVs and 122 indels, that were identified using the MuVaSt somatic variant calling method (Table 14). To assess the precision of the MuVaSt method, the called variants were compared to a Gold Standard variant set that was obtained from WES data of an FF sample with a tumour purity of 30%. This dataset was provided by the CMGG. The Gold Standard set consisted of 267 somatic variants, including 234 SNVs and 33 indels (Table 14). Considering the Gold Standard set, the tumour of patient 2 seemed to contain less mutations than patient 1 (section 2.3.1), this may be attributed to the fact that patient 2 was a non-smoker [13]. Evaluation of the overlap between the MuVaSt variant calls and the Gold Standard set revealed only 10 DNA concordant variants, including 9 SNVs and 1 indel. Therefore, the overall precision of the MuVaSt somatic variant calling method was only 0.0524 for the FFPE sample of patient 2. The precision of SNV detection was higher (0.1304) than the precision of indel detection (0.0082). However, variant calling using the MuVaSt method was more precise than when a single variant caller was applied (Figure 25b).

Table 14. Column 2: Number of variants called in the RNA-Seq data of the FFPE sample using the MuVaSt somatic variant calling method. Column 3: Number of variants in the Gold Standard variant set obtained from WES data of the FF sample. Column 4: Number of MuVaSt variants supported by WES (DNA-concordant). Column 5: Resulting precision.

	MuVaSt	**Gold Standard**	**DNA Concordant**	**Precision**
Total Variants	191	267	10	0.0524
SNVs	69	234	9	0.1304
Indels	122	33	1	0.0082

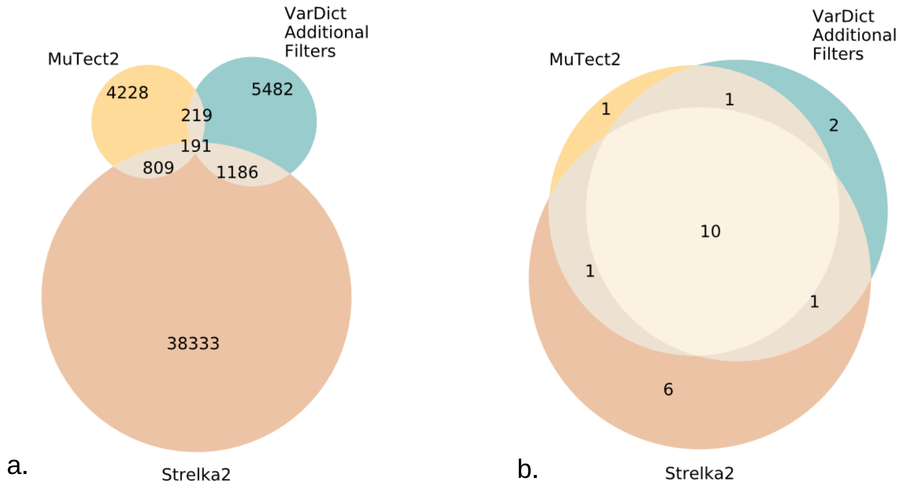


Figure 25. Overview of the overlapping variants between MuTect2, VarDict (with additional filters) and Strelka2 for the FFPE sample when GSNAP alignment and pre-processing Protocol 4 were applied. (a) This Venn diagram only takes into account variants that have passed the built-in quality filters. (b) This Venn diagram only includes variants that were validated by WES (DNA concordant).

It was remarkable that only 5.24% of the RNA variants called by MuVaSt were supported by WES (Gold Standard). Since a large fraction of the Gold Standard variants was missed using RNA-Seq, the allele-expression of the DNA variants was evaluated. It was observed that only 32 of the 225 DNA unique SNVs were supported by at least one RNA read and only 26 had a VAF greater than 0.04. Therefore, at least 193 DNA unique SNVs (85.8%) could not be identified in RNA-Seq. As it was already observed in patient 1, a large fraction of the DNA unique variants is not, or lowly, expressed.

To further assess the limited overlap between the RNA variants and the DNA variants, the VAF was analysed. It was observed that variants identified using WES had a VAF_{DNA} distribution that was mainly below 0.25 (Figure 26b). This may be attributed to the fact that the tumour purity was only 30%, which consequently resulted in a lower VAF_{DNA} for both clonal and subclonal mutations. For the DNA unique variants, a first peak was observed between 0 and 0.1, comprising most likely the subclonal mutations; and a second smaller peak between 0.1 and 0.2, which may be attributed to the clonal mutations. This suggests that the tumour tissue only contained a limited number of clonal mutations and a larger fraction of subclonal mutations. The VAF_{DNA} distribution of the concordant variants shows a peak between 0.1 and 0.2 which indicates that the concordant variants comprise most likely clonal mutations. Indeed, further analysis of the concordant variants revealed a mutation in the KRAS gene (chr12: 21971123, C to A) and in the CDKN2A gene (chr9: 21971123, G to GT). As already described in section 1.1.2, these genes are considered driver genes of lung adenocarcinoma. Since driver gene mutations are crucial in the progression to a neoplastic cancer cell, these mutations are considered clonal mutations that are present in the majority of cancer cells in the tumour. Furthermore, a lot of the discordant variants may be attributed to a low VAF due to subclonal mutations and a low tumour purity (tumour heterogeneity), since these mutations are more difficult to identify correctly (as was already described in section 3.3.5); and to false positive variants caused by artefacts. Moreover, it should be noted that an FFPE sample was used for RNA-Seq and an FF sample for WES, which may contribute to the low overlap between both sequencing techniques. In conclusion, the sample used for the validation of the MuVaSt somatic variant calling method appeared to be more challenging due to a higher percentage of subclonal mutations and a low tumour purity. Nevertheless, MuVaSt variant calling succeeded in identifying two driver gene mutations that were verified by WES.

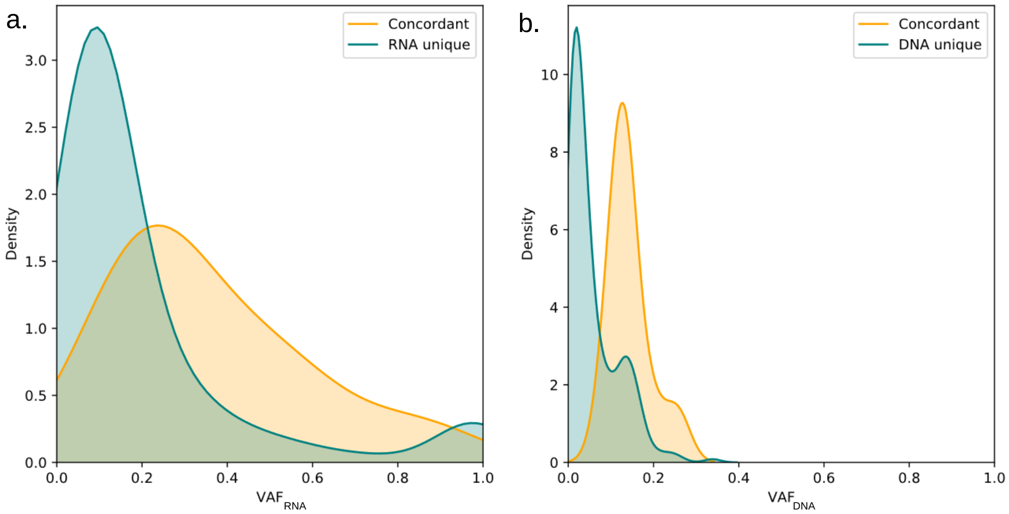


Figure 26. Distribution plots of the variant allele frequency (VAF) for the DNA concordant and discordant fractions for the FFPE sample when the MuVaSt variant calling method was used (GSNAP and Protocol 4). (a) VAF_{RNA} for the RNA variants as calculated in section 2.3.4. The blue graph depicts the RNA variants that were not supported by WES, the yellow graph depicts variants that were supported by WES. (b) VAF_{DNA} for the DNA variants as calculated in section 2.3.4. The blue graph depicts the DNA variants that were not supported by RNA-Seq, the yellow graph depicts variants that were supported by RNA-Seq.

4 DISCUSSION

One crucial step in the development of a personalised therapeutic vaccine is the detection of tumour specific antigens: neoantigens. The detection of neoantigens is a multistep process comprising alignment of sequencing data, pre-processing of alignment files, somatic variant calling, filtering of identified variants, neoepitope prediction and selection of immunogenic neoantigens candidates. Every single step in this pipeline requires extensive optimisation to ensure the selection of reliable neoantigens that can be used for the development of a personalised therapeutic vaccine. In this master thesis the alignment, pre-processing and variant calling were analysed and evaluated in order to pave the way for the development of a fully operational bioinformatics pipeline that allows the accurate identification of somatic variants from transcriptome analysis of both FF and FFPE samples.

4.1 Performance Evaluation of the Bioinformatics Pipeline

The alignment is a crucial step in the detection of somatic mutations from RNA-Seq data. Both GSNAP and STAR are splice-aware aligners that have already proven to perform reliable and accurate alignment of RNA-Seq data [66], [67]. However, these benchmarking studies did not evaluate the somatic variant calling accuracy. Therefore, in this master thesis, the precision of both alignment algorithms was assessed using a Gold Standard variant set containing somatic variants detected using WES data. It could be concluded that GSNAP alignment resulted in a higher overall variant calling precision than STAR alignment for both the FF and the FFPE sample (Figure 13). The difference in overall precision was mainly attributable to a higher precision of SNV detection after GSNAP alignment. The precision of indel detection remained considerably low for both aligners. It is reasonable to assume that RNA-Seq aligners yield nearly identical results when using the same input data. Nevertheless, Figure 14 demonstrates only limited concordance between GSNAP and STAR. It should be noted that these results were obtained using only MuTect2 as variant caller, which means a large fraction of the called variants might be false positive variants. Nonetheless, identical analysis with the optimised variant calling method MuVaSt also resulted in a limited concordance between both aligners (Supplementary Figure 8). Hong *et al*. [116] already revealed such a discrepancy among splice-aware alignment algorithms. The inconsistent identification of somatic variants may be attributed to the algorithmic differences between GSNAP and STAR. To further investigate the RNA somatic variant discrepancy among splice-aware alignment algorithms, a larger number of samples should be examined. However, this was out of scope of this master thesis.

Before proceeding to the actual variant calling, the alignment files were subjected to various pre-processing steps in order to enhance variant calling accuracy. The BAM files were sorted, duplicate reads were marked and an additional pre-processing protocol was applied (Figure 6). The performance evaluation of different pre-processing protocols demonstrated superior performances for Protocol 3 and 4. Protocol 3 consisted of three pre-processing steps: SplitNCigarReads, IndelRealigner and BaseRecalibrator. Protocol 4, on the other hand, only consisted of two pre-processing steps: SplitNCigarReads and BaseRecalibrator. Consecutive application of SplitNCigarReads and BaseRecalibrator had a positive effect on the overall precision of variant calling. IndelRealigner, on the other hand, had only a limited effect, since no vast improvement was observed for the precision of indel detection. This is not surprising since the variant calling algorithms used, include a realignment step and thus do not require additional indel realignment. As expected, Protocol 3 and 4 identified a large fraction of common variants (Figure 12). Among the somatic variants identified by only one pre-processing protocol, only a limited number of DNA concordant variants was observed (Supplementary Figure 6). It can be concluded that GATK's BaseRecalibrator and SplitNCigarReads are the most appropriate pre-processing steps to apply to RNA-Seq alignment files before proceeding to somatic variant calling. This is not surprising since these pre-processing steps are included in the workflow recommended by GATK [117] for somatic variant calling on RNA-Seq data. It is important to note that the vast improvement in the overall precision of somatic variant calling was mainly attributable to an enhanced precision of SNV detection since for each pre-processing protocol the precision of indel detection remained low.

There are plenty of tools available for performing variant calling of SNVs and indels in NGS data [77]. However, for implementation in our bioinformatics pipeline, the variant calling algorithm should meet a small set of criteria. First of all, the variant calling tool should be able to handle transcriptomic data. Second, the variant calling tool should be able to operate in paired tumour-normal-sample mode and accordingly identify somatic variants. Based on these selection criteria and previous benchmarking studies [55], [75] MuTect2, VarDict and Strelka2 were selected. MuTect2 and Strelka2 are more "modern" haplotype-based variant callers while VarDict employs a position-based strategy to identify somatic variants [77]. With respect to the precision, the results in Figure 18 and Figure 19 demonstrated that MuTect2 achieved the highest overall precision and the highest precision of both SNV and indel detection for both sample types. Nevertheless, after additional filtering of VarDict called variants, a higher precision of SNV detection was obtained for the FF sample. Contrarily, this improvement was not as pronounced for the FFPE sample. It is remarkable that no single variant calling algorithm succeeded in identifying indels with a precision greater than 0.01. However, all three variant callers claim to be suitable for the accurate detection of indels [73], [78], [80]. It can be concluded that each variant caller alone called many somatic variants that were not supported by WES. MuTect2 provided the least amount of DNA discordant variant calls compared to the other two variant callers and, therefore, MuTect2 seems the most appropriate variant caller. Nonetheless, only about 5-7% of the variants identified using MuTect2 were supported by WES. Therefore, the implementation of a single variant caller seemed not adequate in discovering somatic variants with a high precision in RNA-Seq data.

One interesting observation is the fact that only a limited number of variants were called by all three variant callers (Figure 20, Figure 22), a phenomenon already observed by Neums *et al.* [56]. Hence, a large portion of the variants identified by only one variant calling tool were most likely false positive variants. Therefore, in order to enhance the precision of variant calling, a new method was employed combining all three aforementioned variant callers: MuVaSt. A vast improvement in the precision of SNV detection was observed (Table 12, Table 13). For the FF sample, 84% of the detected SNVs was supported by WES. Nevertheless, the precision of indel detection remained below 0.03 for both sample types.

It is clear that the accurate identification of indels poses a major bioinformatic challenge because both the alignment and the variant calling are hindered by indels. Furthermore, no vast improvement in the precision of indel detection was observed when different pre-processing steps, different aligners or different variant callers were applied. Correct alignment of indels requires splice-aware aligners since RNA-Seq data is composed of spliced exons. In addition, the alignment algorithm must allow gapped alignment. However, both GSNAP and STAR have already proven to be suitable for indel detection [55]. Sun e*t al.* [55] indeed selected GSNAP and STAR as most optimal alignment tools for accurate indel detection. Nevertheless, when somatic indels identified using RNA-Seq were evaluated in WES data, only a low overlap was demonstrated [55]. A large fraction of the RNA unique indels consisted of insertions of only one base pair. Further in-depth analysis of these insertions revealed a systematic error at repeated nucleotide sites (Appendix 8.15). For example, an insertion of one T was followed by 5 successive T residues (Supplementary Figure 13). This indicates that it is possible that these insertions may be originating from systematic errors introduced during the library preparation. To assess this systematic error, further research is required.

4.2 RNA-Seq versus WES

WES data has been employed extensively for the identification of somatic mutations [7], [35], [36], [39], [40] due to its reliability and relatively low cost. Nevertheless, RNA-Seq is gaining interest because this sequencing technique has the ability to address a multitude of different questions, such as the quantification of gene expression levels, detection of somatic mutations, detection of alternative splicing, allele-specific expression, gene fusions and RNA editing. However, extensive research is required to assess the reliability of somatic variants detected using RNA-Seq data alone. In this master thesis, a combination of three variant callers, MuVaSt, was employed to identify somatic variants in RNA-Seq data with a high precision. The overlap between variants called from WES data and variants called from RNA-Seq data was summarised in Figure 23. Similar to previous studies [33], [54], [55], only a low overlap was observed, both for SNVs and indels. In a recent study, O'Brien *et al.* [54] examined the detection of somatic SNVs from both WES and RNA-Seq data of the same tumour and normal samples, and also revealed only a low overlap. Sun *et al.* [55] confirmed this finding and revealed an even lower overlap for somatic indels. Coudray e*t al.* [33] used a workflow comprising STAR alignment in 2-pass mode, SplitNCigarReads, BaseRecalibrator and MuTect2 somatic variant calling in paired tumour-sample mode. Again, only a small overlap was found between RNA-Seq and WES variants.

Further analysis of the RNA-Seq unique variants, DNA unique variants and concordant variants revealed some (biological) factors contributing to this limited overlap (Table 15). Expression analysis using RNA-Seq data demonstrated that about 46% of the genes containing WES variants were not, or very lowly, expressed. Therefore, it would be impossible to identify these variants using RNA-Seq only. This was also found by O'Brien *et al*. [54]. Nevertheless, expression of the gene does not automatically imply expression of the alternate (non-reference) allele. There may be several explanations for this. Firstly, the alternate alleles of the DNA unique variants may be located on the untranscribed strand and, therefore, were not transcribed at the mRNA level due to allele-specific expression [54]. Secondly, it is possible that the variant indeed occurred on the transcribed strand but its variant allele frequency was too low to be detected due to tumour heterogeneity. Evaluation of the allele-specific expression of SNVs revealed that a large fraction (about 65%) of SNVs identified by WES was not supported by RNA-Seq data. It can be concluded that analysis of both gene expression and allele-specific expression indicated that a large fraction of the DNA unique variants couldn't be identified using RNA-Seq because the variants were not expressed. It is important to note that for the aim of the therapeutic pipeline only expressed variants are of clinical relevance. Moreover, it has already been observed that gene expression levels and the resulting neoantigens presented on the cell surface are positively correlated, and therefore have a great influence of the immune recognition and the subsequent lysis of the respective cell [90], [91]. As a result, many variants called in WES may not have an impact at the biological level because the variants are located within non-expressed genes or alleles.

One interesting metric to assess the difference between RNA-Seq and WES variants is the variant allele frequency (VAF). It was observed that RNA and DNA unique variants had a lower VAF than concordant variants (Figure 21). Moreover, a large fraction of the RNA unique variants had a VAF_{RNA} below 0.2 (Figure 21a). It is possible that this low VAF_{RNA} was due to subclonal mutations that were only present in a limited fraction of the cancer cells (tumour heterogeneity). Consequently, it is more difficult to detect these variants in WES. Contrarily, a low VAF_{RNA} may also be attributed to a lowly expressed clonal mutation that can be detected using WES. The VAF_{DNA} of DNA variants not supported by RNA-Seq (Figure 21b) showed two peaks. Since somatic mutations with a high VAF_{DNA} are more likely to be clonal [115], the VAF_{DNA} peak of the DNA unique variants between 0.17 and 0.4 may be denoted to clonal mutations that were located in non-expressed genes and, therefore, were not detected in RNA-Seq. The VAF_{DNA} peak below 0.17, on the other hand, can potentially be allocated to subclonal mutations [115] that were only present in a limited fraction of the tumour. As a result, only a limited number of RNA-Seq reads supported these mutations and, therefore, they were not identified as a somatic variant in RNA-Seq data. It can be concluded that it is challenging to correctly identify variants with a low VAF. Both RNA-Seq and WES allowed the detection of variants with a VAF below 0.2, nevertheless, only a limited fraction of these variants was called by both sequencing techniques and therefore considered true positive. It is important to note that a low VAF may also be due to artefacts originating from sequencing errors, misalignments, library preparation artefacts or sample preservation damage.

For the development of a therapeutic vaccine, clonal mutations are attractive immunological targets as they are expressed on all cells within the cancer cell population. Subclonal mutations, on the other hand, are solely expressed on that subclone or subpopulation, hence targeting them would only eliminate a fraction of the tumour cells [118]. As already described before, clonal mutations can be found in the majority of cancer cells in the tumour tissue since these mutations were part of the original set of mutations present when a cell transformed into a neoplastic cancer cell. These clonal mutations have a VAF that is around 0.5, since most somatic mutations are heterozygous. However, somatic mutations in tumour tissue are expected to appear at a lower VAF, due to the normal cell content of a tumour sample [33]. Moreover, copy number variations can lead to gain or loss of chromosomal regions, and duplication or deletion of genes [92]. Indeed, Figure 21b shows a VAF_{DNA} peak between 0.17 and 0.4 which may be allocated to the clonal mutations present in the tumour tissue [115]. Subclonal mutations, on the other hand, are acquired by daughter cells during tumour growth and appear only in a limited fraction of the cancer cells and, as a result, have a VAF below 0.5. Indeed, both RNA-Seq and WES identified variants with a VAF below 0.2, comprising most likely the subclonal mutations [115]. Nevertheless, only a limited fraction of these variants was considered true positive. It should be noted that the VAF is also affected by the tumour purity. A reduced percentage of tumour cells present in the sample will lower the VAF of both clonal and subclonal mutations which subsequently complicates the correct classification of somatic mutations.

Both the fact that subclonal mutations are not of great relevance as vaccine targets [118], and the fact that only a limited precision was obtained for variants with a VAF below 0.2, imply that it would be beneficial to apply a threshold VAF value of 0.2 to filter out false positive variants. Indeed, when this threshold was applied, a higher overall precision was obtained. Nevertheless, this filter was only beneficial for the detection of SNVs. The precision of indel detection remained very low because very few concordant indels were identified. Implementation of this VAF threshold for RNA variants indeed improved the overlap between RNA-Seq and WES variants. Nevertheless, it is possible that some of the RNA unique variants with a low VAF_{RNA} are true positive variants [119]. Moreover, it was already observed by Coudray *et al.* [33] that true variants showing a low VAF_{RNA} but a high coverage in RNA-Seq data are more likely to be missed by WES. Therefore, care should be taken when applying this VAF threshold because it is possible that true variants were excluded. Moreover, it is important to note that the implementation of this VAF threshold value was based on the analysis of only two tumour samples with a tumour purity of around 50%. Since the VAF is highly dependent of the tumour purity, this threshold value of 0.20 is not relevant for every tumour sample. Although this master thesis indicates that the application of a VAF threshold might be beneficial for the accurate identification of somatic variants, this VAF threshold value should be adjusted for different tumour purities. Therefore, a dynamic

VAF threshold would be more appropriate to use. Nevertheless, the establishment of a dynamic VAF threshold would require investigation of a large number of samples comprising different tumour purities.

Evaluation of the RNA unique variants revealed that only a limited fraction was covered by the WES capture regions, this was also found by O’Brien *et al.* [54]. It can be concluded that while RNA-Seq covers the whole transcriptome, WES is limited to detecting variants in known exons and their flanking regions. This is an important aspect to consider, because it means that many potentially important somatic mutations that were not located in the WES target region would be missed if only WES was considered. Another biological explanation why RNA-Seq unique SNVs might be missed by WES is due to RNA editing. As already described before, in humans, RNA editing mostly results in an A:T→G:C mutational signature [62]. Subsequently, WES of tumour tissue might not be adequate to detect all present somatic variants. However, in the tumour samples analysed, no clear enrichment in potential RNA editing sites was observed.

In conclusion, RNA-Seq has proven to be suitable for the detection of somatic variants and, moreover, imposes some additional benefits over WES. The most important advantage is the fact that RNA sequencing only allows the detection of expressed variants and that RNA-Seq data can be used for the quantification of expression. In addition, RNA-Seq has the potential to detect novel transcripts, gene fusions, alternative splicing events and other features without the limitation of required prior knowledge [32], [33], [54], [58]. For this reason, RNA-Seq is considered a powerful sequencing technique for the detection of somatic mutations in cancer transcriptomes that might be missed by WES. Nevertheless, some RNA-Seq variants could be considered questionable, since RNA-Seq data has been shown to be more prone to false positive calls due to technical variation introduced during library preparation, particularly during the RNA to cDNA conversion with reverse transcriptase [60]. Other important sources of false positive calls are sequencing [120] and alignment, especially incorrect alignment at the very ends of a read due to splicing [121]. However, a splice junction filter was already applied to reject variants close to a known exon-exon junction (Table 2, section 2.2.3).

The accurate detection of somatic variants is not only hindered by bioinformatics challenges but also by biological factors. Most importantly, the nature of tumour tissue makes somatic variant calling a challenging task [122]. Tumour heterogeneity, including normal cell content (tumour purity) and subclonal mutations, complicates the somatic variant calling because the heterozygote allele distribution cannot be expected in tumour tissue. Hence, in addition to specificity also sensitivity of a somatic variant caller is of great interest, in order to detect a delicate signal. Especially variants with a low VAF impose a great challenge since it becomes more difficult to determine whether a somatic mutation is a true mutation or an artefact.

Table 15. Summary of factors that may lead to inconsistencies in the detection of somatic variants in WES versus RNA-Seq.

Factors causing RNA-Seq unique variants	Factors causing WES unique variants
Variants outside WES capture regions	Non-expressed variants so no coverage in RNA-Seq
RNA editing: conversion from A:T to G:C	Variants on non-transcribed strand of the gene: allele-specific expression
Low VAF in WES due to tumour heterogeneity (subclonal mutations or normal cell content)	Low VAF in RNA-Seq due to allele-specific expression
Error-prone RNA-Seq data	Errors introduced during WES

4.3 FFPE versus FF Sample Preparation

Apart from the difficulties introduced due to tumour heterogeneity, the type of tissue preservation imposed an additional factor of variability. As already described in section 1.2.2, the use of FFPE samples can introduce artefacts due to the fixation process, and the storage time and conditions [45]–[50]. Indeed, it was observed that the precision for variant calling in the FFPE sample was lower than for the FF sample (Table 12, Table 13). Moreover, in the FFPE sample, a higher percentage of T to C and A to G nucleotide substitutions was observed (Figure 24). These substitutions may potentially be assigned to FFPE preservation artefacts [49] and, therefore, more likely represent false positive variants. In addition, only a limited fraction of the variants was identified in both the FFPE and the FF sample (Figure 23). This limited overlap may be the result of low-quality mRNA that was obtained from the RNA extraction on the FFPE sample (Appendix 8.1). Besides, it should be noted that the tumour sample region was not exactly the same for the FFPE and the FF sample. Hence, it is possible that due to the geographical heterogeneity of tumour tissue, different variants were identified in different samples. It can be concluded that the identification of somatic mutations in FFPE samples is possible although the precision is lowered because of additional artefacts origination from the formalin fixation process.

4.4 Validation

To validate the ability of the MuVaSt somatic variant calling method to identify somatic variants with a high precision, a second FFPE sample originating from a lung adenocarcinoma was used. The Gold Standard variant set consisted of only 267 variants (Table 14). This smaller number of mutations may be attributed to the fact that patient 2 was a non-smoker [13]. The overall precision of MuVaSt variant calling was considerably low (0.0524) because only 10 DNA concordant variants, including 9 SNVs and 1 indel, were identified. Nevertheless, MuVaSt succeeded in identifying two driver gene mutations that were verified by WES. Further investigation of the tumour data revealed a large fraction of subclonal mutations and only a small fraction of clonal mutations (Figure 26b). Moreover, the tumour purity of the FF sample was only 30%, which resulted in a reduced VAF_{DNA} for the clonal (peak between 0.1 and 0.2) and subclonal mutations (peak between 0 and 0.1). This low tumour purity and the high percentage of subclonal mutations might explain why only a limited precision was obtained for the MuVaSt method. As already described before, variants with a low VAF can also represent false positive variants caused by artefacts originating from sequencing errors, misalignments, library preparation artefacts or sample preservation damage. It can be concluded that the sample used for the validation appeared to be more challenging due to the presence of a higher percentage of subclonal mutations and the low tumour purity. Nevertheless, this tumour heterogeneity complicates both WES and RNA-Seq [77]. Therefore, further research is still required to enhance the somatic variant calling in heterogeneous tumour samples.

The bioinformatics workflow evaluated in this thesis evaluates only a small part in the therapeutic pipeline for the development of a personalised therapeutic vaccine. The somatic variants identified in the workflow described in this master thesis have to be further processed to ultimately obtain a set of immunogenic neoantigens that can be used for the development of a dendritic cell vaccine. A few steps that are involved are HLA-typing, and neoepitope prediction and selection. Obviously, also these steps need to be evaluated in order to optimise the detection of neoantigens. The accurate detection is only one of the obstacles that has to be overcome. Subsequently, the implementation of a personalised neoantigen vaccine as standard therapy is still ongoing research and many challenges have to be solved. The optimal combination therapy, the reduction of vaccine production time, the upscaling of manufacturing, and ensuring affordability are only a few factors that require further research. Nevertheless, if one succeeds in producing a properly optimised personalised combination therapy, a new era of cancer treatment will be introduced that allows to treat patients of different cancer types with a high accuracy and efficacy.

5 CONCLUSION

In this master thesis the alignment, pre-processing and variant calling was analysed and evaluated in order to pave the way for the development of a fully operational bioinformatics pipeline that allows the accurate identification of somatic variants (SNVs and indels) from transcriptome analysis of FFPE and FF samples. Careful evaluation of the alignment step revealed a more precise variant calling when the GSNAP alignment algorithm was employed. For the pre-processing of the alignment files, consecutive application of GATK's SplitNCigarReads and BaseRecalibrator enhanced the precision of variant calling, in particular for SNVs. Next, evaluation of the individual somatic variant calling algorithms MuTect2, VarDict and Strelka2 revealed suboptimal performances. Therefore, it was tested if the combination of these three variant callers would provide a higher rate of variant calls supported by WES. Indeed, the new variant calling method MuVaSt succeeded in identifying somatic SNVs with a higher precision. Nevertheless, for the identification of indels only a low precision was achieved.

Although WES has been the mainstay for the identification of somatic variants in cancer genomes, this master thesis suggests that variant calling from RNA-Seq offers a valuable complement. Comparison of somatic variants from WES and RNA-Seq revealed a low overlap. It was observed that some of the DNA variants were missed in RNA-Seq due to low, or no, expression of the gene or the variant allele. Furthermore, some of the variants detected in RNA-Seq were missed in WES because they were located outside the WES capture regions. This means that potential somatic mutations that were not located in the WES target region would be missed if only WES was considered. It was observed that a large fraction of the discordant variants had a low VAF, potentially caused by tumour heterogeneity and/or allele-specific expression; or due to artefacts originating from sequencing errors, misalignments, library preparation artefacts or sample preservation damage. Accordingly, if a somatic mutation only appears in a small fraction of the tumour mass or the variant allele has a low expression, it is more difficult to identify this somatic variant both in WES and RNA-Seq. The validation of the FFPE data from a second patient supported this finding: a low tumour purity and a high fraction of subclonal mutations complicated the detection of somatic variants. Therefore, the bioinformatics pipeline should be optimised in order to allow more precise variant calling of heterogeneous tumour samples.

Application of this optimised workflow revealed a limited overlap between somatic variants identified in the FF sample and the FFPE sample. This low overlap may be allocated to the geographical heterogeneity of the tumour tissue. Moreover, FFPE samples are more likely to contain artefacts originating from the formalin fixation process. Subsequently, these artefacts can be misclassified as somatic variants. As a result, care should be taken when solely FFPE samples are used for somatic variant calling.

In conclusion, the highest precision (0.8381) was obtained for the detection of SNVs in the FF sample of patient 1 when GSNAP alignment, SplitNCigarReads and BaseRecalibrator pre-processing tools, and MuVaSt somatic variant calling were applied. Therefore, this master thesis suggests that the implementation of RNA-Seq for the detection of somatic SNVs might pose a feasible alternative for WES. Nevertheless, for samples with a low tumour purity neoantigen detection using RNA-Seq seems inadequate. Moreover, RNA-Seq did not allow the accurate identification of indels, which indicates that WES will still be required for indel detection. Accordingly, the implementation of a pipeline solely based on RNA-Seq is not yet sufficiently reliable to replace WES for the detection of neoantigens. Further validation involving more samples and optimisation for low purity samples is still required. In addition, neoantigen detection from FFPE tissue seems feasible, although the implementation of FFPE samples in the therapeutic pipeline demands more research as for now, higher precisions are obtained for FF samples.

6 FUTURE IDEAS

The feasibility and precision of a bioinformatics pipeline for the detection of somatic variants in RNA-Seq data was assessed in this master thesis. Although a considerable high precision was obtained for the detection of SNVs, the precision for the detection of indels remained relatively low. Moreover, a low tumour purity and a high fraction of subclonal mutations complicated the precise detection of somatic variants. Therefore, several ways of improving the detection of somatic mutations using RNA-Seq are possible. Variant calling algorithms have been evolving and improving over the past years. For example, some of the original position-based variant callers were upgraded to haplotype-based variant callers (e.g. Strelka to Strelka2 and MuTect to MuTect2) [77]. More recently, the use of deep learning algorithms was introduced for somatic variant calling. For example, NeuSomatic [123] is the first Convolutional Neural Networks-based approach that can effectively leverage signals derived from the alignment of WES data to accurately identify somatic variants. Benchmarking of NeuSomatic revealed superior performances compared to MuTect2, Strelka2 and VarDict [123]. For now, this deep learning approach is limited to the analysis of DNA sequencing data. Implementation of deep learning algorithms in variant callers that specifically handle RNA-Seq data might improve the detection of both SNVs and indels.

To overcome the limitation of accurate detection of indels in RNA-Seq, machine learning algorithms specifically targeting indels were developed recently. For example, RNAIndel [124] employs a machine learning based algorithm for the classification of indels identified in RNA-Seq data into somatic, germline, and artefact by random forest models. Another important factor contributing to the accurate identification of indels is the alignment. Recently, ABRA2 [125] was developed providing an improved realignment in order to enable more accurate variant calling. Despite the fact that more modern variant callers, such as MuTect2 and Strelka2, already include a local realignment step in their algorithm, a more sensitive and precise detection of both SNV and indels was obtained when ABRA2 was used to realign reads initially mapped by STAR [125]. Since it was observed that a large fraction of the indels might potentially be attributed to systematic errors originating from the library preparation, another solution could be to use a Panel Of Normals (PON), an option which is available for e.g. MuTect2. A PON consists of a large set of VCF files of normal samples, that underwent the same processing i.e. library preparation and sequencing as the study sample, and its primary purpose is to eliminate false positive variant calls that are due to systematic technical errors [126].

In this master thesis, real-life sequencing data was used. To calculate the precision and sensitivity of variant calling, a Gold Standard variant set containing somatic variants detected in WES data of the tumour tissue was employed. Nevertheless, as already described before, this Gold Standard variant set is not optimal for several reasons. First, it may contain variants that are not expressed in the tumour tissue. Second, RNA-Seq has the ability to call variants that cannot be detected in WES. Since it is impossible to verify whether the Gold Standard set of variants includes all true variants, the resulting precision and sensitivity might be biased. Despite these limitations, to date, no good validation datasets are available for independent and unbiased benchmarking of somatic variant callers [77]. Other recent benchmarking studies have used artificial spiked in datasets, RNA-Seq simulations programs such as BEERS [127], datasets obtained from cell lines, and real tumour-normal pairs. However, none of these are perfect validation datasets. Therefore, the development of a benchmarking dataset composed of a collection of real cancer transcriptomes that are deep sequenced to generate high-confidence and validated somatic variants would be a great asset for the research community.

7 REFERENCES

[1] J. Ferlay *et al.*, “Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.,” *Int. J. cancer*, vol. 136, no. 5, pp. E359-86, Mar. 2015.

[2] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[3] P. A. Monach, S. C. Meredith, C. T. Siegel, and H. Schreiber, “A unique tumor antigen produced by a single amino acid substitution.,” *Immunity*, vol. 2, no. 1, pp. 45–59, Jan. 1995.

[4] G. M. Cooper, *The cell : a molecular approach*. ASM Press, 2000.

[5] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, “Cancer genome landscapes.,” *Science*, vol. 339, no. 6127, pp. 1546–58, Mar. 2013.

[6] J. Yokota, “Tumor progression and metastasis,” *Carcinogenesis*, vol. 21, no. 3, pp. 497–503, Mar. 2000.

[7] U. Sahin and Ö. Türeci, “Personalized vaccines for cancer immunotherapy.,” *Science*, vol. 359, no. 6382, pp. 1355–1360, Mar. 2018.

[8] “The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI).” [Online]. Available: https://www.genome.gov/sequencingcosts/. [Accessed: 09-Oct-2018].

[9] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013.

[10] W. D. Travis *et al.*, “The 2015 World Health Organization Classification of Lung Tumors,” 2015.

[11] S. Viswanath, A. Pathak, A. Kapoor, A. Rathore, and B. N. Kapur, “Changing paradigm in treatment of lung cancer,” *J. Cancer Metastasis Treat.*, vol. 2, no. 6, p. 214, Jun. 2016.

[12] E. A. Collisson *et al.*, “Comprehensive molecular profiling of lung adenocarcinoma,” *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014.

[13] M. Imielinski *et al.*, “Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing,” *Cell*, vol. 150, pp. 1107–1120, 2012.

[14] G. Veronesi, F. Bianchi, and K. Inamura, “Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification,” *Front. Oncol*, vol. 7, p. 193, 2017.

[15] W. Pao and N. Girard, “New driver mutations in non-small-cell lung cancer,” *Lancet Oncol.*, vol. 12, no. 2, pp. 175–180, Feb. 2011.

[16] T. Cancer Genome Atlas Research Network, “Comprehensive genomic characterization of squamous cell lung cancers,” 2012.

[17] R. D. Schreiber, L. J. Old, and M. J. Smyth, “Cancer Immunoediting: Integrating Immunity’s Roles in Cancer Suppression and Promotion,” *Science (80-.).*, vol. 331, no. 6024, pp. 1565–1570, Mar. 2011.

[18] D. S. Chen and I. Mellman, “Oncology Meets Immunology: The Cancer-Immunity Cycle (Review),” *Immunity*, vol. 39, pp. 1–10, 2013.

[19] J. Charles A Janeway, P. Travers, M. Walport, and M. J. Shlomchik, “The major histocompatibility complex and its functions,” 2001.

[20] J. Neefjes, M. L. M. Jongsma, P. Paul, and O. Bakke, “Towards a systems understanding of MHC class I and MHC class II antigen presentation,” *Nat. Rev. Immunol.*, vol. 11, no. 12, pp. 823–836, Dec. 2011.

[21] K. S. Kobayashi and P. J. van den Elsen, “NLRC5: a key regulator of MHC class I-dependent immune responses,” *Nat. Rev. Immunol.*, vol. 12, no. 12, pp. 813–820, Dec. 2012.

[22] U. Sahin *et al.*, “Human neoplasms elicit multiple specific immune responses in the autologous host.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 25, pp. 11810–3, Dec. 1995.

[23] F. S. Hodi *et al.*, “Improved survival with ipilimumab in patients with metastatic melanoma.,” *N. Engl. J. Med.*, vol. 363, no. 8, pp. 711–23, Aug. 2010.

[24] C. Linnemann *et al.*, “High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma,” *Nat. Med.*, vol. 21, no. 1, pp. 81–85, Jan. 2015.

[25] P. F. Robbins *et al.*, “Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells.,” *Nat. Med.*, vol. 19, no. 6, pp. 747–52, Jun. 2013.

[26] S. Champiat, C. Ferté, S. Lebel-Binay, A. Eggermont, and J. C. Soria, “Exomics and immunogenics: Bridging mutational load and immune checkpoints efficacy.,” *Oncoimmunology*, vol. 3, no. 1, p. e27817, Jan. 2014.

[27] N. A. Rizvi *et al.*, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.”

[28] G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, and R. D. Schreiber, “Cancer immunoediting: from immunosurveillance to tumor escape,” *Nat. Immunol.*, vol. 3, no. 11, pp. 991–998, Nov. 2002.

[29] H. Matsushita *et al.*, “Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting.,” *Nature*, vol. 482, no. 7385, pp. 400–4, Feb. 2012.

[30] M. Wang *et al.*, “Role of tumor microenvironment in tumorigenesis,” *J. Cancer*, vol. 8, no. 5, pp. 761–773, 2017.

[31] T. L. Whiteside, “The tumor microenvironment and its role in promoting tumor growth,” 2008.

[32] R. Piskol, G. Ramaswami, and J. B. Li, “Reliable Identification of Genomic Variants from RNA-Seq Data,” *Am. J. Hum. Genet.*, vol. 93, pp. 641–651, 2013.

[33] A. Coudray, A. M. Battenhouse, P. Bucher, and V. R. Iyer, “Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data.”

[34] T. Karasaki *et al.*, “Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing.,” *Cancer Sci.*, vol. 108, no. 2, pp. 170–177, Feb. 2017.

[35] P. A. Ott *et al.*, “An immunogenic personal neoantigen vaccine for patients with melanoma,” *Nature*, vol. 547, no. 7662, pp. 217–221, Jul. 2017.

[36] U. Sahin *et al.*, “Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer,” *Nature*, vol. 547, no. 7662, pp. 222–226, Jul. 2017.

[37] F. Duan *et al.*, “Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.,” *J. Exp. Med.*, vol. 211, no. 11, pp. 2231–48, Oct. 2014.

[38] S. Kreiter *et al.*, “Mutant MHC class II epitopes drive therapeutic immune responses to cancer.,” *Nature*, vol. 520, no. 7549, pp. 692–6, Apr. 2015.

[39] B. M. Carreno *et al.*, “Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells.,” *Science*, vol. 348, no. 6236, pp. 803–8, May 2015.

[40] T. N. Schumacher and R. D. Schreiber, “Neoantigens in cancer immunotherapy.,” *Science*, vol. 348, no. 6230, pp. 69–74, Apr. 2015.

[41] S. Turajlic *et al.*, “Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis,” *Lancet Oncol.*, vol. 18, no. 8, pp. 1009–1021, Aug. 2017.

[42] “FFPE vs Frozen Tissue Samples - BioChain Institute Inc.” [Online]. Available: https://www.biochain.com/general/ffpe-vs-frozen-tissue-samples/. [Accessed: 12-Feb-2019].

[43] “What Is FFPE Tissue And What Are Its Uses - BioChain Institute Inc.” [Online]. Available: https://www.biochain.com/general/what-is-ffpe-tissue/. [Accessed: 19-Feb-2019].

[44] S. von Ahlfen, A. Missel, K. Bendrat, and M. Schlumpberger, “Determinants of RNA quality from FFPE samples.,” *PLoS One*, vol. 2, no. 12, p. e1261, Dec. 2007.

[45] J. Hedegaard, K. Thorsen, M. K. Lund, A.-M. K. Hein, and S. J. Hamilton-Dutoit, “Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer and Normal Tissue,” *PLoS One*, vol. 9, no. 5, p. 98187, 2014.

[46] J.-Y. Chung, T. Braunschweig, and S. M. Hewitt, “Optimization of Recovery of RNA From Formalin-fixed, Paraffin-embedded Tissue,” 2006.

[47] N. Masuda, T. Ohnishi, S. Kawamoto, M. Monden, and K. Okubo, “Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples.,” *Nucleic Acids Res.*, vol. 27, no. 22, pp. 4436–43, Nov. 1999.

[48] A. Esteve-Codina *et al.*, “A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples,” *PLoS One*, vol. 12, no. 1, p. e0170632, Jan. 2017.

[49] S. Graw *et al.*, “Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples,” *Nat. Publ. Gr.*, 2015.

[50] S. von Ahlfen, A. Missel, K. Bendrat, and M. Schlumpberger, “Determinants of RNA quality from FFPE samples.,” *PLoS One*, vol. 2, no. 12, p. e1261, Dec. 2007.

[51] P. Zhang, B. D. Lehmann, Y. Shyr, and Y. Guo, “The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies.,” *Int. J. Genomics*, vol. 2017, p. 1926304, 2017.

[52] S. H. Kresse *et al.*, “Evaluation of commercial DNA and RNA extraction methods for high-throughput sequencing of FFPE samples,” *PLoS One*, vol. 13, no. 5, p. e0197456, May 2018.

[53] B. Heemskerk, P. Kvistborg, and T. N. M. Schumacher, “The cancer antigenome.,” *EMBO J.*, vol. 32, no. 2, pp. 194–203, Jan. 2013.

[54] T. D. O’Brien *et al.*, “Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer,” *Methods*, vol. 83, pp. 118–127, 2015.

[55] Z. Sun, A. Bhagwate, N. Prodduturi, P. Yang, and J.-P. A. Kocher, “Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations,” *Brief. Bioinform.*, vol. 18, no. 6, p. bbw069, Jul. 2016.

[56] L. Neums *et al.*, “VaDiR: an integrated approach to Variant Detection in RNA,” vol. 7, pp. 1–13, 2017.

[57] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, “Ten years of next-generation sequencing technology,” *Trends Genet.*, vol. 30, no. 9, pp. 418–426, Sep. 2014.

[58] I. Chepelev, G. Wei, Q. Tang, and K. Zhao, “Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq,” *Nucleic Acids Res.*, vol. 37, no. 16, pp. e106–e106, Sep. 2009.

[59] B. Heemskerk, P. Kvistborg, and N. M. Schumacher, “The cancer antigenome,” *EMBO J.*, vol. 32, pp. 194–203, 2013.

[60] A. G. Williams, S. Thomas, S. K. Wyman, and A. K. Holloway, “RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis; Richard et,” *Curr Protoc Hum Genet*, vol. 83.

[61] S.-J. Cho, V. Blanc, and N. O. Davidson, “Mouse Models as Tools to Explore Cytidine-to-Uridine RNA Editing,” *Methods Enzymol.*, vol. 424, pp. 417–435, Jan. 2007.

[62] E. Picardi, C. Manzari, F. Mastropasqua, I. Aiello, A. M. D’Erchia, and G. Pesole, “Profiling RNA editing in human tissues: towards the inosinome Atlas,” *Sci. Rep.*, vol. 5, p. 14941, Oct. 2015.

[63] M. Quail *et al.*, “A table of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers,” *BMC Genomics*, vol. 13, no. 1, p. 341, Jul. 2012.

[64] Y. Chu and D. R. Corey, “RNA sequencing: platform selection, experimental design, and data interpretation.,” *Nucleic Acid Ther.*, vol. 22, no. 4, pp. 271–4, Aug. 2012.

[65] J. C. Castle *et al.*, “Exploiting the mutanome for tumor vaccination.,” *Cancer Res.*, vol. 72, no. 5, pp. 1081–91, Mar. 2012.

[66] G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. Fitzgerald, and G. R. Grant, “Simulation-based comprehensive benchmarking of RNA-seq aligners HHS Public Access,” *Nat Methods*, vol. 14, no. 2, pp. 135–139, 2017.

[67] P. G. Engström *et al.*, “Systematic evaluation of spliced alignment programs for RNA-seq data.,” *Nat. Methods*, vol. 10, no. 12, pp. 1185–91, Dec. 2013.

[68] T. D. Wu and S. Nacu, “Fast and SNP-tolerant detection of complex variants and splicing in short reads.,” *Bioinformatics*, vol. 26, no. 7, pp. 873–81, Apr. 2010.

[69] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner.,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.

[70] Ö. Türeci, M. Vormehr, M. Diken, S. Kreiter, C. Huber, and U. Sahin, “Targeting the Heterogeneity of Cancer with Individualized Neoepitope Vaccines.,” *Clin. Cancer Res.*, vol. 22, no. 8, pp. 1885–96, Apr. 2016.

[71] H. Yang, Y. Zhong, C. Peng, J.-Q. Chen, and D. Tian, “Important role of indels in somatic mutations of human cancer genes.,” *BMC Med. Genet.*, vol. 11, p. 128, Sep. 2010.

[72] A. McKenna *et al.*, “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.,” *Genome Res.*, vol. 20, no. 9, pp. 1297–303, Sep. 2010.

[73] K. Cibulskis *et al.*, “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples,” *Nat. Biotechnol.*, vol. 31, no. 3, pp. 213–219, Mar. 2013.

[74] C. Wang *et al.*, “RVboost: RNA-seq variants prioritization using a boosting method.,” *Bioinformatics*, vol. 30, no. 23, pp. 3414–6, Dec. 2014.

[75] S. Sandmann *et al.*, “Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.,” *Sci. Rep.*, vol. 7, p. 43169, 2017.

[76] “GATK | Tool Documentation Index | MuTect2.” [Online]. Available: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php. [Accessed: 23-Apr-2019].

[77] C. Xu, “A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data,” *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 15–24, Jan. 2018.

[78] Z. Lai *et al.*, “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.,” *Nucleic Acids Res.*, vol. 44, no. 11, p. e108, 2016.

[79] “Developing low frequency filters for cancer variant calling using VarDict – Blue Collar Bioinformatics.” [Online]. Available: http://bcb.io/2016/04/04/vardict-filtering/. [Accessed: 11-Oct-2018].

[80] S. Kim *et al.*, “Strelka2: fast and accurate calling of germline and somatic variants,” *Nat. Methods*, vol. 15, no. 8, pp. 591–594, Aug. 2018.

[81] W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Dec. 2016.

[82] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, pp. e164–e164, Sep. 2010.

[83] A. H. Ramos *et al.*, “Oncotator: Cancer Variant Annotation Tool,” *Hum. Mutat.*, vol. 36, no. 4, pp. E2423–E2429, Apr. 2015.

[84] P. Cingolani *et al.*, “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff,” *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, Apr. 2012.

[85] G. Ramaswami and J. B. Li, “RADAR: a rigorously annotated database of A-to-I RNA editing,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D109–D113, Jan. 2014.

[86] E. Picardi, A. M. D’Erchia, C. Lo Giudice, and G. Pesole, “REDIportal: a comprehensive database of A-to-I RNA editing events in humans.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D750–D757, 2017.

[87] C. A. Brennick, M. M. George, W. L. Corwin, P. K. Srivastava, and H. Ebrahimi-Nik, “Neoepitopes as cancer immunotherapy targets: key challenges and opportunities,” *Immunotherapy*, vol. 9, no. 4, pp. 361–371, Mar. 2017.

[88] M. Andreatta and M. Nielsen, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system,” *Bioinformatics*, vol. 32, no. 4, pp. 511–517, Feb. 2016.

[89] M. Nielsen *et al.*, “Reliable prediction of T-cell epitopes using neural networks with novel sequence representations,” *Protein Sci.*, vol. 12, no. 5, pp. 1007–1017, May 2003.

[90] X. S. Liu and E. R. Mardis, “Applications of Immunogenomics to Cancer.,” *Cell*, vol. 168, no. 4, pp. 600–612, 2017.

[91] M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, and M. Mann, “Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.,” *Mol. Cell. Proteomics*, vol. 14, no. 3, pp. 658–73, Mar. 2015.

[92] D. Yin *et al.*, “High-Resolution Genomic Copy Number Profiling of Glioblastoma Multiforme by Single Nucleotide Polymorphism DNA Microarray,” *Mol. Cancer Res.*, vol. 7, no. 5, pp. 665–677, May 2009.

[93] A. Durgeau, Y. Virk, S. Corgnac, and F. Mami-Chouaib, “Recent Advances in Targeting CD8 T-Cell Immunity for More Effective Cancer Immunotherapy,” *Front. Immunol.*, vol. 9, p. 14, Jan. 2018.

[94] E. Brabants *et al.*, “An accelerated, clinical-grade protocol to generate high yields of type 1-polarizing messenger RNA-loaded dendritic cells for cancer vaccination.,” *Cytotherapy*, vol. 20, no. 9, pp. 1164–1181, Sep. 2018.

[95] Qiagen, “RNeasy ® FFPE Handbook Sample & Assay Technologies QIAGEN Sample and Assay Technologies,” *Handbook*, no. September, 2014.

[96] Promega Corporation, “Maxwell® RSC simplyRNA Cells Kit and Maxwell® RSC simplyRNA Tissue Kit,” 2014.

[97] Illumina, “TruSeq RNA Exome Reference Guide (1000000039582),” 2018.

[98] N. Nijs, “MethodSOP _ RNA Capture Library Prep,” vol. 036, no. version 6, pp. 1–31, 2019.

[99] “Picard Tools - By Broad Institute.” [Online]. Available: https://broadinstitute.github.io/picard/. [Accessed: 19-Feb-2019].

[100] G. Tischler and S. Leonard, “biobambam: tools for read pair collation based algorithms on BAM files,” *Source Code Biol. Med.*, vol. 9, no. 1, p. 13, Dec. 2014.

[101] R Core Team, “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria, 2018.

[102] G. Van Rossum, “The Python Language Reference Manual,” 2011.

[103] J. G. Cleary *et al.*, “Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines,” *bioRxiv*, p. 023754, Aug. 2015.

[104] Peter Krusche, “Haplotype Comparison Tools.” [Online]. Available:

https://github.com/Illumina/hap.py.

[105] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

[106] “VCF Annotation Tools (VAtools) — VCF Annotation Tools (VAtools) 3.1.0 documentation.” [Online]. Available: https://vatools.readthedocs.io/en/latest/index.html. [Accessed: 08-May-2019].

[107] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[108] H. Li, “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data,” *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, Nov. 2011.

[109] “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.” [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 06-Nov-2018].

[110] “GATK | Tool Documentation Index | SplitNCigarReads.” [Online]. Available: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_rnaseq_SplitNCigarReads.php. [Accessed: 24-Apr-2019].

[111] “GATK | Tool Documentation Index | BaseRecalibrator.” [Online]. Available: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_bqsr_BaseRecalibrator.php. [Accessed: 24-Apr-2019].

[112] “GATK | Tool Documentation Index | IndelRealigner.” [Online]. Available: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php. [Accessed: 24-Apr-2019].

[113] “GATK | Tool Documentation Index | ASEReadCounter.” [Online]. Available: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_rnaseq_ASEReadCounter.php. [Accessed: 02-May-2019].

[114] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, “A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data.,” *Bioinformatics*, vol. 30, no. 12, pp. i78-86, Jun. 2014.

[115] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva, “Identification of neutral tumor evolution across cancer types.,” *Nat. Genet.*, vol. 48, no. 3, pp. 238–244, Mar. 2016.

[116] J. H. Hong, Y. H. Ko, and K. Kang, “RNA variant identification discrepancy among splice-aware alignment algorithms,” *PLoS One*, vol. 13, no. 8, p. e0201822, Aug. 2018.

[117] “Calling variants in RNAseq — GATK-Forum.” [Online]. Available: https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq. [Accessed: 02-May-2019].

[118] Z. Hu, P. A. Ott, and C. J. Wu, “Towards personalized, tumour-specific, therapeutic vaccines for cancer,” *Nature Reviews Immunology*, vol. 18, no. 3. Nature Publishing Group, pp. 168–182, 01-Mar-2018.

[119] H.-T. Shin *et al.*, “Prevalence and detection of low-allele-fraction variants in clinical cancer samples.,” *Nat. Commun.*, vol. 8, no. 1, p. 1377, 2017.

[120] K. Nakamura *et al.*, “Sequence-specific error profile of Illumina sequencers.,” *Nucleic Acids Res.*, vol. 39, no. 13, p. e90, Jul. 2011.

[121] E. T. Cirulli *et al.*, “Screening the human exome: a comparison of whole genome and whole transcriptome sequencing.,” *Genome Biol.*, vol. 11, no. 5, p. R57, 2010.

[122] A. B. Krøigård, M. Thomassen, A.-V. Lænkholm, T. A. Kruse, and M. J. Larsen, “Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data,” *PLoS One*, vol. 11, no. 3, p. e0151664, Mar. 2016.

[123] S. Mohammad *et al.*, “Deep convolutional neural networks for accurate somatic mutation detection.”

[124] K. Hagiwara *et al.*, “RNAIndel: a machine-learning framework for discovery of somatic coding indels using tumor RNA-Seq data,” *bioRxiv*, p. 512749, Jan. 2019.

[125] L. E. Mose, C. M. Perou, and J. S. Parker, “Improved indel detection in DNA and RNA via realignment with ABRA2,” *Bioinformatics*, Jan. 2019.

[126] “PON in Mutect2 — GATK-Forum.” [Online]. Available:

https://gatkforums.broadinstitute.org/gatk/discussion/7796/pon-in-mutect2. [Accessed: 28-May-2019].

[127] G. R. Grant *et al.*, “Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM).,” *Bioinformatics*, vol. 27, no. 18, pp. 2518–28, Sep. 2011.