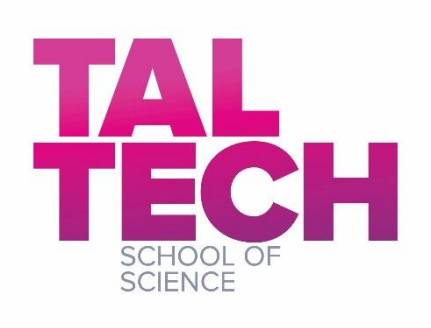


Bioinformatic analysis of tumor whole-exome sequencing data for discovery of clinically relevant mutations in lung cancer

Master's thesis

Author: Laura Luhari
Supervisor: Olli-Pekka Smolander, PhD, TalTech, Associate Professor
Co-Supervisor: Kersti Oselin, PhD, NEMC, Oncologist
Study program: YASM02/18

Tallinn 2022

Kasvaja eksoomi sekveneerimise andmete bioinformaatiline analüüs kopsuvähi kliiniliselt relevantsete mutatsioonide avastamiseks

Magistritöö

Autor: Laura Luhari
Juhendaja: Olli-Pekka Smolander, PhD, Taltech, Kaasprofessor
Kaasjuhendaja: Kersti Oselin, PhD, PERH, Onkoloog
Õppeprogramm: YASM02/18

Tallinn 2022

Declaration

Hereby I declare that I have compiled the paper independently and all works, important standpoints and data by other authors have been properly referenced and the same paper has not been previously been presented for grading.

Author: [Name]
[Signature, date]

The paper conforms to requirements in force.
Supervisor: [name]
[Signature, date]

Permitted to the defence.
Chairman of the Defence Committee: [Name]
[Signature, date

Table of Contents

Abbreviations

ALK – anaplastic lymphoma receptor tyrosine kinase
BAM – binary alignment format
BRAF – B-Raf proto-oncogene
COPD – chronic obstructive pulmonary disease
DRAGEN – Dynamic Read Analysis for GENomics
EGFR – epidermal growth factor receptor
ER – endoplasmic reticulum
FDA – Food & Drug Administration
FF – fresh frozen
FFPE – formalin-fixed paraffin-Embedded
GO – gene ontology
GSEA – gene set enrichment analysis
ORA – over representation analysis
HER2 – erb-b2 receptor tyrosine kinase 2
HTS – high-throughput sequencing
INDEL – insertions and deletions
KRAS – Kirsten rat sarcoma viral oncogene homolog
MAF – minor allele frequency
MEK1 – mitogen-activate protein kinase 1
MET – mesenchymal-to-epithelial transition factor
miRNA – micro-RNA
NEMC – North Estonia Medical Centre
NGS – next-generation sequencing
NMD – nonsense-mediated decay
NSCLC – non-small cell lung cancer
NTRK1 – neurotrophic receptor tyrosine kinase 1
PFS – progression free survival
PIK3CA – phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
QC – quality control
ROS1 – ROS proto-oncogene 1
RTK – receptor tyrosine kinase
SAM – sequence alignment format
SCLC – small cell lung cancer
SNV – single nucleotide variant
SR – sarcoplasmic reticulum
TP53 – p53 tumor suppressor
TSG – tumor suppressor gene
UTR – untranslated region
VCF – variant call format
WES – whole-exome sequencing
WGS – whole-genome sequencing

Introduction

Lung cancer is one of the most often diagnosed tumors and the main cause of death in cancer patients all over the world (World Health Organization 2021). In 2020, over 2,2 million of new lung cancer cases and over 1,7 million deaths were reported globally (Sung et al. 2021). Most of lung cancer patients get their diagnosis in the late phases of the disease leading to poor outcomes (Blandin Knight et al. 2017). Even if the tumor is treated with conventional therapies, cancer is likely to return often more aggressively than before. The risk of developing metastatic recurrence is shown to surpass local recurrence, and its aggressiveness is illustrated by the fact of over half of the patients dying in a one-year timeframe (Consonni et al. 2015). Cancer is characterized by different hallmarks, such as uncontrollable proliferation and resisting cell death among other features, but the foundation of these is set on a molecular level. As cancer stems from accumulating somatic mutations, sequencing cancer genetic material is essential to get a fuller understanding of cancer development and recurrence. Due to rapidly evolving opportunities in bioinformatics, it has become possible to analyze cancer exomes and genomes in a small amount of time. One of the most commonly used methods is whole-exome sequencing (WES), which provides comprehensive information on cancer genetic aberrations in the coding subset of the genome ("Whole Exome Sequencing for Cancer Research: IDT" n.d.). WES is valuable in the discovery of causal tumor variants and gives biological insight into underlying molecular alterations (Rabbani, Tekin, and Mahdieh 2014). WES exhibits its advantages and disadvantages compared to whole-genome sequencing (WGS), but, the easier data analysis and lower cost outweigh the downsides for many researchers. Subsequent steps of the variant detection are equally important: data quality control, alignment, variant calling, variant filtration, variant annotation, and variant prioritization (Ugur Sezerman et al. 2019). Each part of the work holds vast amounts of different tools to be used, however making the choice may not be easy. It has to be reckoned that there are no tools, which are suitable for everything and always. Choosing the most appropriate tools precedes research on the subject and includes weighing the pros and cons for each. As root causes for cancer arise from single nucleotide variants (SNVs) as well as larger structural rearrangements, multiple analysis tools may be needed for comprehensive results. Detection of novel genetic alterations in cancer serves as a starting point for developing new and better diagnostic and prognostic molecular targets, as well as promising options for personalized treatments. This current work focused on small variant (SNPs, INDELs) detection and analysis in tumors that later developed either local or metastatic recurrence. WES data was used to search for novel genetic alterations involved in lung cancer re-growth and provide a workflow to implement the data analysis. Executed work procedures illustrate the possibility of novel variant detection when non-tumor control samples are missing, which is common with formalin-fixed paraffin-embedded (FFPE) samples. Sample preparation and sequencing, as well as initial quality control, alignment, and variant calling were provided by Intermountain Precision Genomics, St. George, Utah, USA. The present thesis encompassed downstream steps of the data analysis, with the data preparation, SNP analysis, INDEL analysis, gene enrichment analysis, and result interpretation being the most fundamental parts of the work. The main emphasis was put on finding differences in groups with local recurrence versus distant recurrence to possibly discover novel genetic aberrations contributing to some patients developing an aggressive relapse leading to shorter survival times. The hypothesis was, that implementation of WES and subsequent data analysis enables the discovery of potential biomarkers capable of

predicting the prognosis and outcome of patients. Importantly, discrepancies between the two groups were present, showing a greater amount of enrichments in many gene set terms within the metastatic group. Multiple potential prognostic markers were detected, with most of them being related to Ca^{2+} transport, mitosis, microtubules, and cell motility.

1 Review of the literature

1.1 Lung cancer

Lung cancer is the leading cause of death out of different cancer types (Figure 1). The survival of lung cancer patients at 5 years after diagnosis is only 10% to 20% in most countries (Sung et al. 2021). In 2020, 2.21 million new lung cancer cases and 1.80 million deaths were reported globally (World Health Organization 2021).

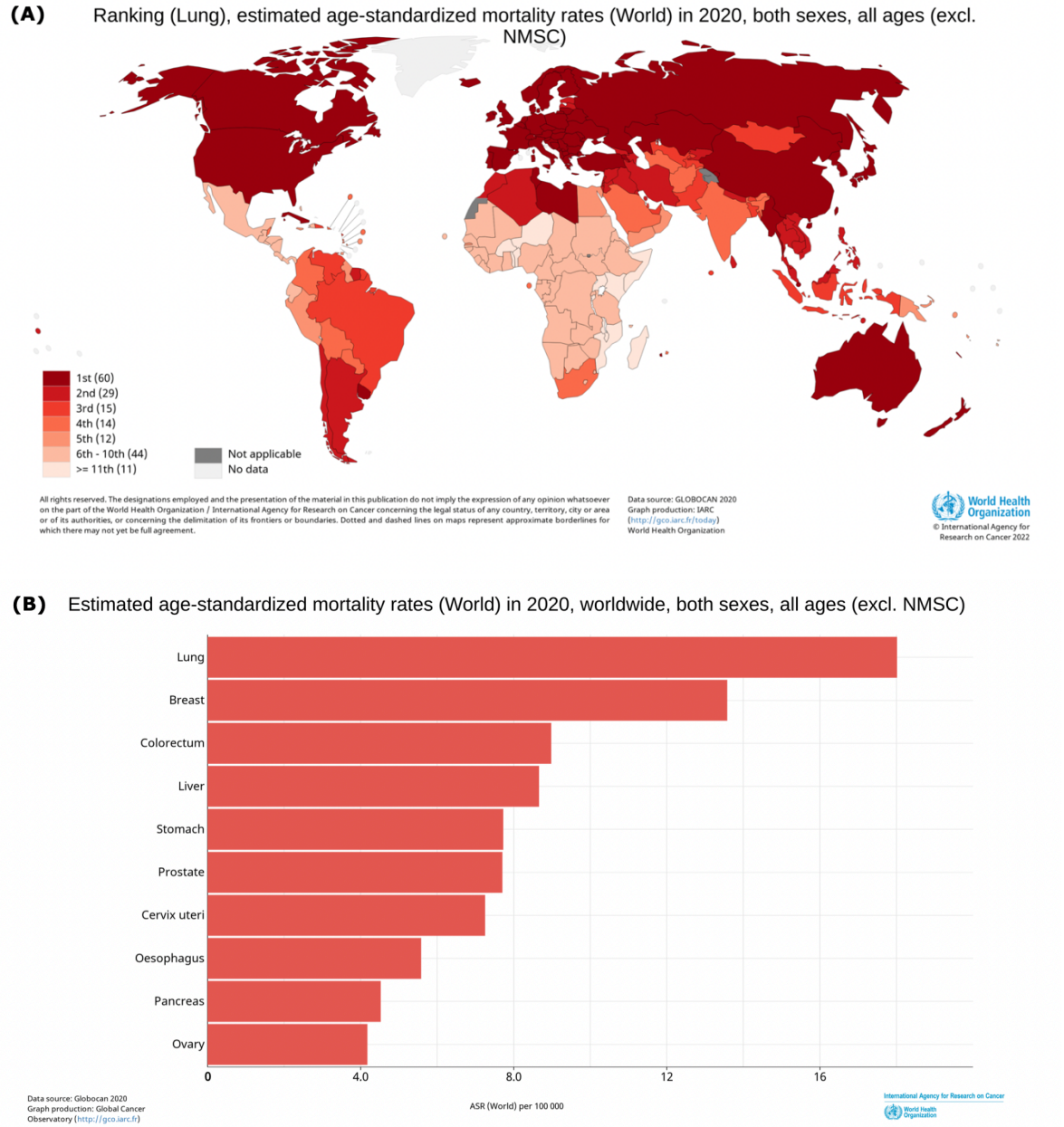


Figure 1. Lung cancer mortality. (A) Lung cancer is the leading cause of death out of all cancer types in many countries. (B) Lung cancer mortality exceeds even breast cancer, which has the highest incidence rate (Global Cancer Observatory n.d.).

Lung cancer is a vastly heterogeneous disease that can emerge in different locations of the bronchial tree, consequently producing highly variable symptoms and signs depending on its anatomic site (Lemjabbar-Alaoui et al. 2015). The most common lung cancer symptoms experienced by patients are cough, dyspnea, hemoptysis, and weight loss. Risk factors include tobacco use or exposure, environmental exposures to radon and asbestos, comorbidities such as HIV, chronic obstructive pulmonary disease (COPD), and family history of the disease (Kelly M. Latimer and Timothy F. Mott 2015). The individual susceptibility to tobacco-evoked lung cancer may rely on competitive gene-enzyme interactions that exert influence on activation or detoxification of procarcinogens as well as determined by the integrity of DNA repair mechanisms (Lemjabbar-Alaoui et al. 2015). Based on histology, lung cancer is classified into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), which comprise 85% and 15% of all, respectively (Figure 2A). SCLC is characterized by high proliferative capacity, a tendency for metastasis generation, and a poor prognosis (Hiddinga et al. 2021). Additionally, NSCLC comprises mainly of adenocarcinomas, followed by squamous cell carcinomas (Figure 2B) (Thai et al. 2021).

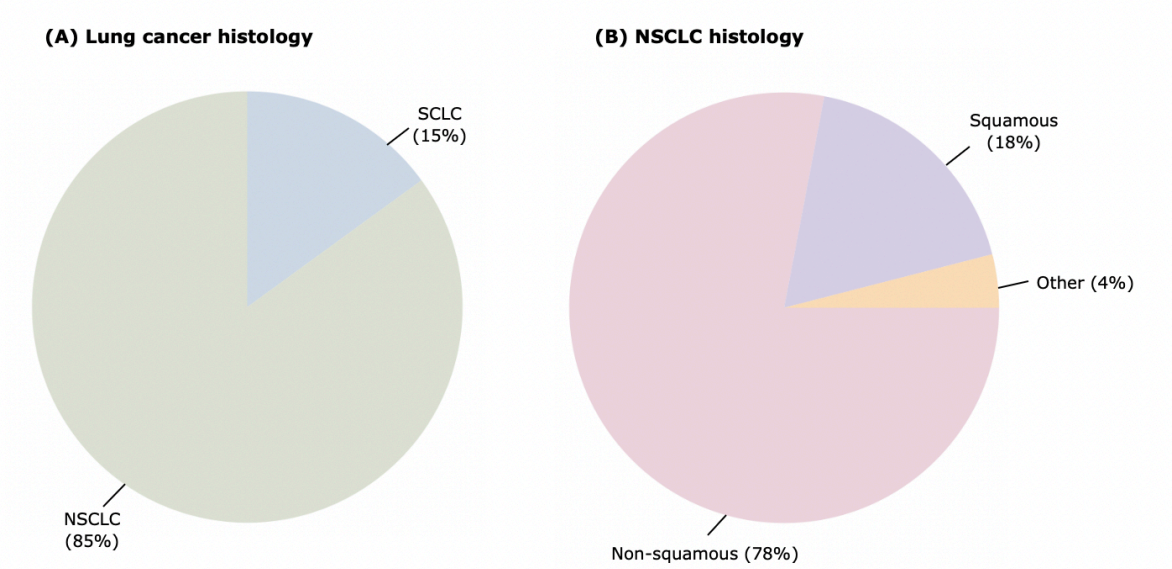


Figure 2. Lung cancer histology. (A) NSCLC constitutes the majority of lung cancer cases with an 85% prevalence, while SCLC makes up the remaining 15%. (B) In the NSCLC cases, non-squamous forms 78% and squamous 18% of the occurrences. Adapted from (Thai et al. 2021).

One of the main reasons for the high mortality rate among lung cancer patients is the high percentage of brain metastases occurrence. Furthermore, 50% out of all brain metastases appear with lung cancer (Yousefi et al. 2017). Poor prognosis is caused by a 30-77% recurrence rate in NSCLC cases (Subotic, Van Schil, and Grigoriu 2016). Furthermore, most of the recurrences are metastatic, whereas patients can die in a one-year timeframe. Recognizing and detecting subsets of patients with a high risk for recurrence and mortality could accelerate interventions that improve the survival (Consonni et al. 2015). Therapeutic approaches include conventional therapies, such as surgery, radiotherapy, and chemotherapy (Gridelli, Rossi, and Maione 2003). Surgery is the recommended treatment for NSCLC patients with a lower stage (I-II) (Vansteenkiste et al. 2014). However, most patients are diagnosed at advanced stages when systemic treatment is needed (Bodor, Kasireddy, and Borghaei 2018). A deeper understanding of molecular mechanisms behind tumorigenesis has allowed the development of precision medicine. Possibly, by targeting those

driver genetic aberrations, clinicians could hit the weak spot of the tumor (Q. Zhang et al. 2020). Most of the therapies are targeted against receptor tyrosine kinases (RTKs) known to be involved in cellular growth and survival (Schrank et al. 2018). Tyrosine kinase inhibitors of the epidermal growth factor receptor (EGFR-TKIs) are the standards of care for first-line treatment in patients with EGFR mutation-positive advanced NSCLC (Alanazi et al. 2021). Patients who have tumors with specific genomic alterations have benefited from targeted therapies. Up to 69% of patients with advanced NSCLC could have a potentially actionable molecular target. Patients lacking targetable mutations receive platinum-based doublet therapy (chemotherapy) (Hirsch et al. 2017; Tsao et al. 2016). As a consequence of routine testing for molecular alterations and the introduction of FDA-approved targeted therapies, mortality from NSCLC and to a lesser extent from SCLC has begun to decline. The mortality from NSCLC has been driven by both decreasing incidence and improving survival. Mortality from SCLC declined almost entirely due to declining incidence, with no improvement in the survival (Howlader et al. 2020). Although major advances have been made with the discovery and use of targeted therapies, their efficacy is limited by cancer drug resistance. The resistance can be intrinsic or acquired, and in the case of the latter, this manifests by the tumor obtaining secondary mutations, using alternative signaling pathways, or changing its phenotype (Lin and Shaw 2016). In recent years, immunotherapy has come to light as a treatment possibility that has led to powerful responses in a subset of patients. These agents hinder key immune checkpoints that normally regulate the immune response but are used by cancer cells to evade the patients' immune system. By blocking these receptor-ligand interactions, a subset of T cells is better activated to recognize and respond to tumor cells (Bodor, Kasireddy, and Borghaei 2018). The main immune checkpoint targets are CTLA-4, PD-1, and PD-L1 (Pérez-Ruiz et al. 2020). Immunotherapy response rates remain highly variable. Prediction of responsiveness is possible due to tumor mutational burden (TMB), which is defined as the total amount of nonsynonymous mutations in the coding area of a tumor genome (Meléndez et al. 2018). The higher the tumor mutational burden, the better the response to the immunotherapy (Pérez-Ruiz et al. 2020).

1.2 Genetic alterations and cancer

Cancer emerges as a result of accumulating changes in the genetic and epigenetic levels. Genetic alterations are provoked by aging, mutagenic chemicals, radiation, ultraviolet light, and oxygen radicals, on the other hand, epigenetic alterations are induced by aging and chronic inflammation (Takeshima and Ushijima 2019). Cancer is characterized by common hallmarks, which serve as a basis for tumor complexity: preserving proliferative signaling, evading growth suppressors and immune destruction, resisting cell death, enabling replicative immortality, promoting angiogenesis, and inducing invasion and metastasis. The foundation for these hallmarks is genomic instability and mutations (Hanahan and Weinberg 2011). All cancers stem from a single cell starting to act in perplexing ways as a result of acquired somatic mutations (Alexandrov and Stratton 2014). Genomic landscapes of cancer usually encompass a small number of frequently mutated genes and a much larger number of rarely altered genes (Vogelstein et al. 2013). Genetic alterations can take place in either coding or non-coding part of the genome. Exome includes many genetic changes leading to altered protein sequences and *de novo* mutations, yet only a minor part of these are disease-causing (Jalali Sefid Dashti and Gamieldien 2017). Some effects of alterations in the coding area of the genome are shown in Figure 3. About 95% of somatic mutations are single-base substitutions, whereas rest are insertions and deletions. 90,7% of the SNPs cause missense changes, 7,6% lead to

truncated protein through stop-gain mutations, and 1,7% result in modification of splice sites or untranslated regions adjacent to start or stop codons (Vogelstein et al. 2013). The least common are frameshift and stop-gain/stop-loss variants, which interestingly give rise to the most damaging effects on the protein level (Seaby, Pengelly, and Ennis 2016). Stop-loss mutations convert a stop codon to a sense codon, hence leading to extended protein translation and may cause tumor suppressor degradation (Dhamija et al. 2020). Missense (non-synonymous) mutations cause changes in the amino acid sequence, and by that, alter protein structure and function. Although synonymous mutations seem benign, as they do not affect the primary protein structure, they can have an indirect impact on protein structure and function (Deng et al. 2017). Synonymous mutations recurrently alter exonic motifs and through that, regulate oncogene splicing. The p53 tumor suppressor (TP53) also has recurrent synonymous mutations adjacent to splice sites, which inactivate them (Supek et al. 2014).

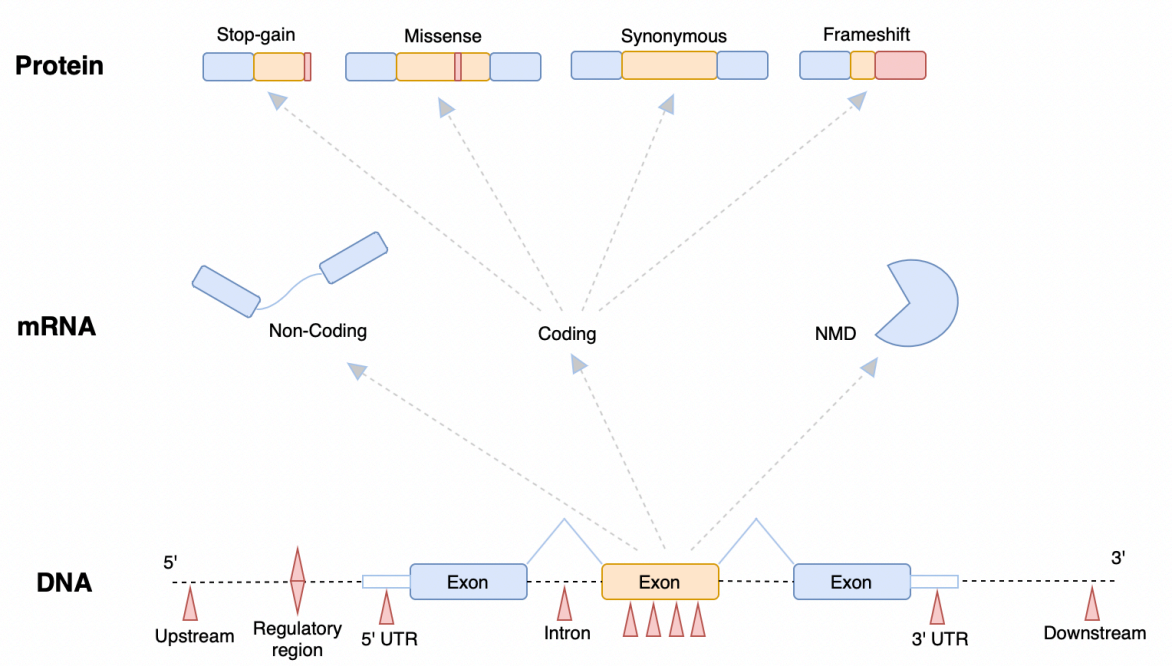


Figure 3. The outcome of genetic alterations. Aberrations in the coding area of the DNA can result in changed mRNA. The cellular process, called nonsense-mediated decay (NMD), is able to remove abnormal mRNAs. A mutation is synonymous if the change in amino acid does not occur. Missense mutations result in one amino acid substitution. The consequence of stop-gain mutation is the occurrence of an early stop codon and therefore, a truncated protein. Base insertions and deletions lead to a frameshift, which alters all the following amino acids. In addition to coding mRNA which is translated to amino acids, there are also non-coding RNAs. Adapted from (Bartha and Győrffy 2019).

Beyond the exome, non-coding mutations are also shown to be involved in cancer development. Those kinds of mutations can be present in the promoter, enhancer, UTR, or miRNA regions (Piraino and Furney 2016). The miRNAs are small non-coding RNAs that are dysregulated in cancer through up- or downregulation, deletion, or epigenetic modifications of miRNA genes. MiRNAs may act as oncogenes or tumor suppressors depending on the conditions and are known to affect common hallmarks of the cancer (Peng and Croce 2016). UTRs mediate post-transcriptional gene regulation and mutations in them are reported as potential drivers of the cancer etiology (Schuster and Hsieh

2019). In addition to smaller alterations, larger chromosomal arrangements are also considered as the primary cause of cancer (Nambiar, Kari, and Raghavan 2008). Major structural variation events are inversion, deletion, duplication, and translocation (van Belzen et al. 2021). Gene fusion events have also been reported in cancers, which can arise from chromosomal rearrangements or transcriptional errors in splicing. In addition, fusion proteins can originate from transcriptional read-throughs (Tuna, Amos, and Mills 2019).

1.2.1 Genetic alterations in lung cancer

Specific tumor types exhibit a larger number of mutations than average, and lung cancer belongs to one of them with ~200 nonsynonymous mutations per tumor. The larger number reflects the involvement of potential mutagen(s), which in the case of lung cancer is cigarette smoke. Therefore, lung cancers from smokers hold 10 times as many somatic mutations as those from the non-smokers' (Vogelstein et al. 2013) (Govindan et al. 2012). Evolvements in sequencing and subsequent data analysis have made it possible to detect mutations, which are the foundation of cancer. Tumor suppressor genes (TSGs) are vital for the regulation of normal cell growth and division. Loss of tumor suppressor gene function is a common mechanism of cancer onset. Alterations inactivating TSGs usually involve two events: deletion of a large chromosomal DNA segment of one allele and a smaller mutational or epigenetic event in the other allele (Osada and Takahashi 2002). In lung cancer, frequently inactivated TSGs are TP53, retinoblastoma 1 (RB1), serine-threonine kinase 11 (STK11), CDKN2A, FHIT, RASSF1A, and PTEN (Cooper et al. 2013). New technologies have facilitated the utilization of targeted therapies. Current oncogenic protein targets are EGFR, ALK, MET, HER2, ROS1, BRAF, RET, NTRK1, MEK1, PIK3CA (Figure 4). Some of the drugs developed against these targets are FDA approved, for example, drugs targeting EGFR, but most of them are going through clinical trials (Hirsch et al. 2017). It is predicted that the discovery of novel mutated genes and molecular pathways lays the way for increasing the number of targeted therapy drugs, reaching combinational use, and better outcome (Sanchez-Vega et al. 2018). Wide-ranging profiling studies implemented with WES or broad targeting panels in NSCLC tumors have exposed multiple non-random patterns of co-occurring or mutually exclusive mutations, which usually vary depending on the specific oncogenic driver mutation. Co-mutations form major determinants of tumor molecular variety and can affect cancer hallmarks, determine prognosis, predict response to systemic therapies and impact mechanisms of resistance (Skoulidis and Heymach 2019). For example, co-existing TP53 and EGFR mutations are associated with noticeably shorter time to progression and shorter overall survival (Yu et al. 2018).

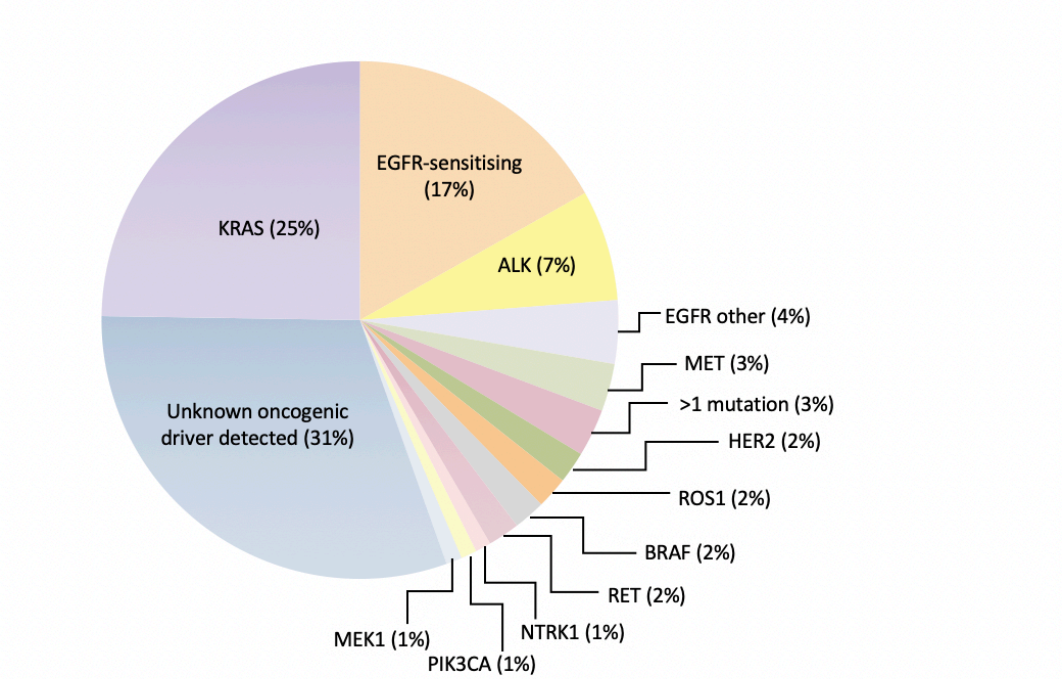


Figure 4. Commonly altered oncogenes in lung adenocarcinoma. The most frequently occurring mutations include unknown oncogenic drivers (31%), KRAS (25%) and EGFR-sensitizing (17%). Following are ALK (7%), EGFR-other (4%), MET (3%), >1 mutation (3%), HER2 (2%), ROS1 (2%), BRAF (2%), RET (2%), NTRK (1%), PIK3CA (1%), and MEK1 (1%). Adapted from (Hirsch et al. 2017).

1.3 Whole-exome sequencing

Whole-exome sequencing (WES) is a next generation sequencing (NGS) technology used to determine the protein coding region of the genome. As most of the disease-causing mutations are known to be located in the coding area, exome sequencing is the most advantageous tool to use making it more cost-efficient compared to WGS (Bartha and Győrffy 2019). Retrospective analysis has shown that mutation calls within the coding regions of WGS and WES data are consistent to a large extent (MC3 Working Group et al. 2020). The main parts of the WES work-flow can be categorized into two actions (Figure 5): the preparation of genomic libraries plus capturing the exome and NGS of eluted target sequences. Subsequently, data analysis pipeline follows (Goh and Choi 2012). Studies have shown that exome analysis is suitable for cancer research, enabling detection of genetic predispositions and new molecular targets (Réda et al. 2020; Mendoza-Alvarez et al. 2019). Abundant use of NGS and novel comprehension of cancer on a molecular level has shifted from cancer type-based approaches to gene- and biomarker-based strategies (Suwinski et al. 2019). A standard pipeline for WES data analysis has not been set and the variability across studies is substantial (Rotunno et al. 2020). Evidence suggests that exome analysis should be conducted using tumor-normal sample pairs if possible, while utilization of tumor-only can lead to false positives as the definitive identification of germline mutations is unachievable (Jones et al. 2015). However, there exist a large collection of tumor-only samples that contain valuable genomic information. WES utilizes hybridization capture to provide comprehensive data, which can be used for tumor profiling. The possibilities of WES include acquiring information of patient's risk of developing specific types of cancer, providing knowledge of genetic changes affecting tumor

progression and helping doctors make decisions of targeted therapies ("Whole Exome Sequencing for Cancer Research: IDT" n.d.).

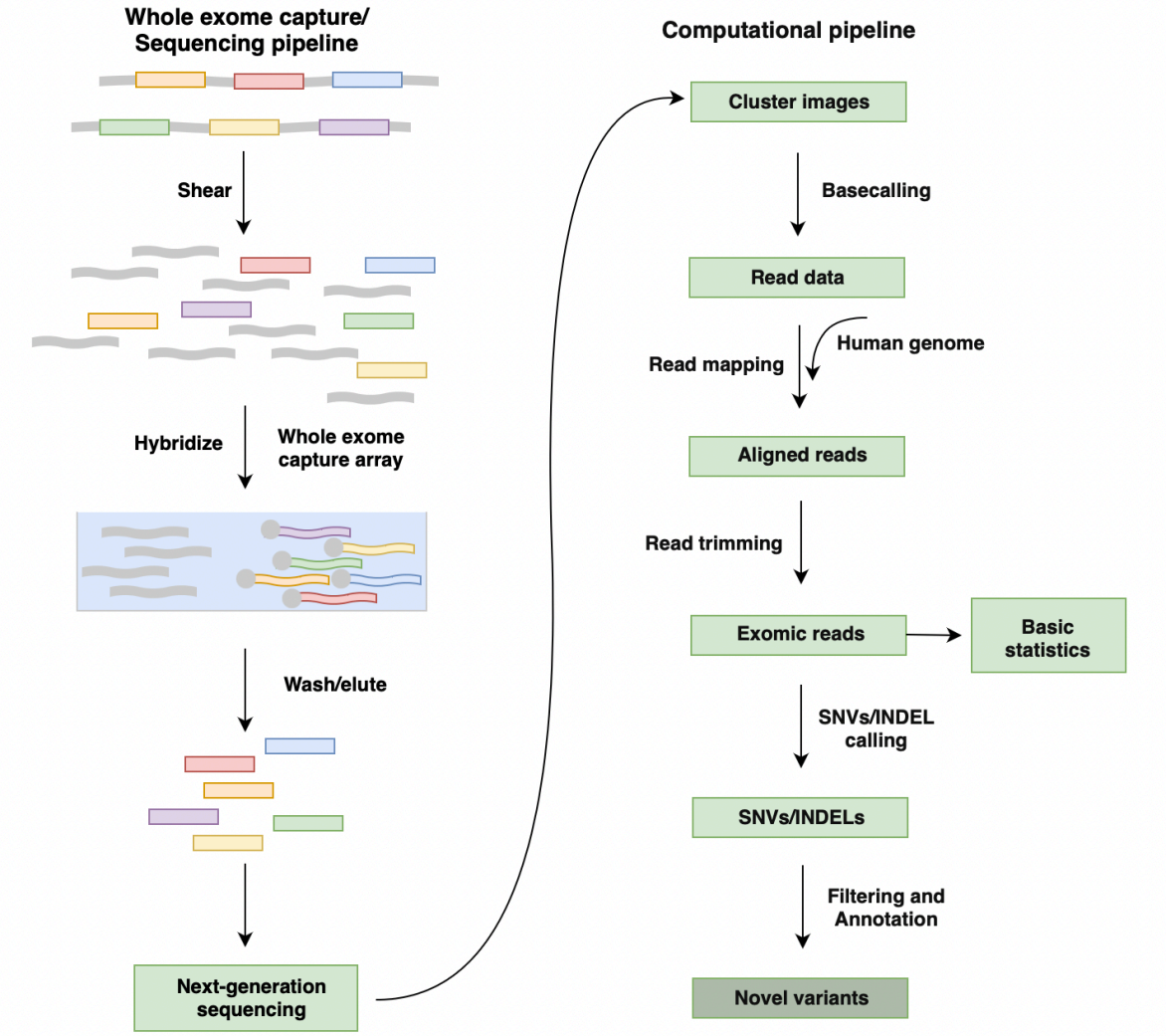


Figure 5. Whole-exome sequencing workflow. The experimental pipeline can be separated into two main parts: 1) Preparation of genomic DNA libraries and hybridization to capture arrays 2) NGS of the eluted target fragments. After generation of short sequencing reads, they are mapped to a reference genome and variant calling is performed. Subsequent filtration and annotation of variants is carried out for comprehensive analysis of their potential biological effect. Adapted from (Goh and Choi 2012).

1.4 Formalin-fixed paraffin-embedded tissues in genomic analysis

Formalin fixation and paraffin embedding of tissues preserve the morphology, therefore it has become the mainly used method for diagnostic surgical pathology. Nucleic acid extraction is shown to be equally successful 1-year or 12-years from preservation (Kokkat et al. 2013). Utilization of fresh frozen (FF) or formalin-fixed paraffin-embedded (FFPE) tissues for NGS has been discussed in several articles. The main downsides of FFPE technique samples are difficulties in extracting high quality DNA and differentiating true variant calls from artifacts, as formalin can induce mutations. Nonetheless FFPE has many advantages over FF sampling, owing to the fact that clinicians do not

have the capability to regularly collect FF tissue and its preservation is complicated due to requirement of liquid nitrogen or dry ice (Mathieson and Thomas 2019). The study comparing FFPE and FF from hepatocellular carcinoma with the aim to produce WES data showed 91% sensitivity of variants detection for FFPE (Ouchi et al. 2013). Gao and colleagues found, that there is a high concordance between FF and FFPE tissue variant detection, nevertheless important differences between tissue types was observed. They also noticed disparities in total coverage, as FF tissue had significantly higher rate (Gao et al. 2020). All SNVs and INDELs taken into account, 90% of cancer-related genes were found altered with higher frequency in the FFPE data set than FF data set. However, investigating the most clinically relevant types of aberrations, such as missense, stop, frameshift, and splicing variants >99% of genes did not show higher rate of mutations in FFPE data sets (on behalf of the 100,000 Genomes Project et al. 2018). Multiple studies conclude, that the use of FFPE samples is feasible in WES variant discovery (Astolfi et al. 2015; Bailey et al. 2018).

1.5 Data analysis from raw data to clinically relevant variants

FASTQ and FASTA are standard formats incorporating biological sequence data. The FASTA format is a text-based depiction of sequenced material starting with the sequence name followed by nucleotides, FASTQ format additionally includes base quality scores for smoother evaluation of sequencing quality (Institute for Systems Genomics 2017). WES data analysis starts with quality control, continues with mapping to reference genome, follows with variant calling, and ends with annotation, filtration, and prioritization (Goh and Choi 2012; Ugur Sezerman et al. 2019). A typical WES data analysis pipeline is seen in Figure 6. WES data analysis has many challenges, as there are vast amounts of tools created for each phase of variant identification. Hence, the advantages and downsides of every tool have to be considered before being put into the application (Ugur Sezerman et al. 2019). WES generates a large amount of data, which has to be aligned to a reference genome in order to allow variant calling for identification of SNVs and INDELs (Suwinski et al. 2019). Identification of single nucleotide variants (SNV) is dependent on the accuracy of the variant calling (Bartha and Győrffy 2019). After variant detection, annotation characteristics such as genomic feature, gene symbol, exonic function, and amino acid alteration can be added to the variant list. Most studies focus on non-synonymous SNPs and INDELs in the coding part of the region, as they account for most of the disease-associated mutations in complex diseases (Bao et al. 2014). Annotation has a substantial effect on the final interpretation of findings, as errors could lead to false negatives or false positives (Goh and Choi 2012). ANNOVAR is one of the most commonly used software tools for annotation of called variants (Rotunno et al. 2020). ANNOVAR is a command-line based tool, which takes text-based input files (e.g. VCF files) and generates output files with annotations for every variant in the input file (H. Yang and Wang 2015). ANNOVAR provides fast and simple gene-based, region-based and filter-based annotations (Wang, Li, and Hakonarson 2010). Additionally, choice of reference transcript set, such as RefSeq or Ensembl, is equally important (McCarthy et al. 2014). The output of high-throughput sequencing (HTS) is about one called variant per 1000 base pairs of sequenced genetic material compared to the reference genome, yielding tens of thousands of sequence variants in WES. Appropriate filtering helps to reduce the excess of variants with the aim to retain potentially pathogenic ones (Caspar et al. 2018). Every filtering process starts with making decisions, which are based on logical suppositions of causal variants, and must be conducted wisely. The fundamental element of filtering is elimination of benign variants (Seaby, Pengelly, and Ennis 2016).

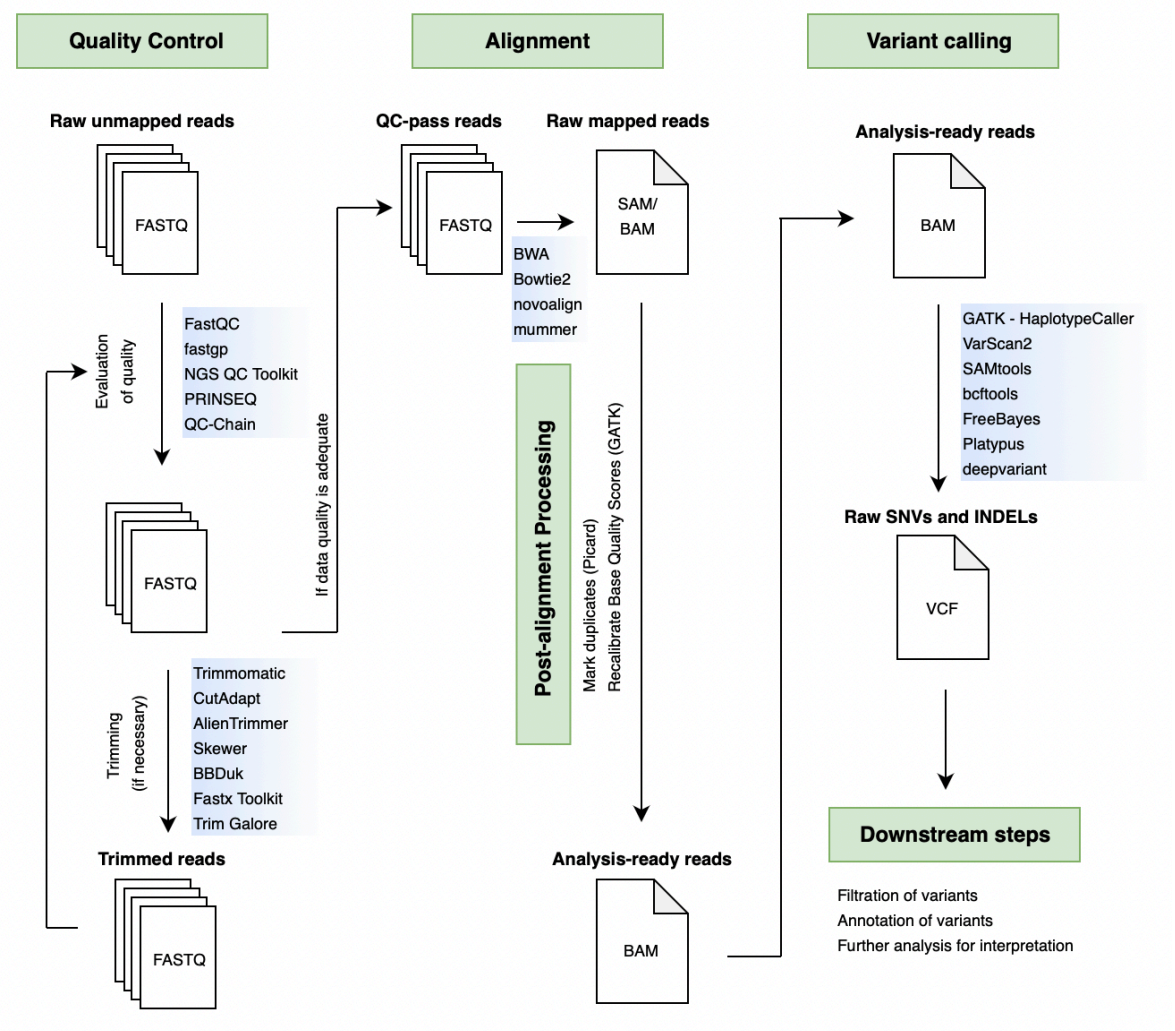


Figure 6. Variant discovery pipeline. Green rectangles represent different steps of the workflow with examples of possible tools for implementation (blue gradient background). White rectangles display produced file formats. The data analysis process starts with quality control: raw unmapped reads (FASTQ) pervade evaluation of the quality and subsequent trimming, producing trimmed reads (FASTQ). High quality as a prerequisite, work proceeds with alignment: QC-pass reads (FASTQ) are aligned to a reference genome and raw mapped reads (SAM/BAM) are created. The next step is post-alignment processing, which leads to analysis-ready reads (BAM). The analysis continues with variant calling, where raw SNVs and INDELs are produced (VCF). The final part comprehends downstream steps, such as filtration, annotation, and further analysis of variants. Adapted from (Ugur Sezerman et al. 2019).

Excessive dependence on automation and predictive tools may result in the elimination of clinically relevant variants or provide false positives (Jalali Sefid Dashti and Gamieldien 2017). Many databases can be used for filtration with the purpose of distinguishing novel variants from common polymorphisms. For that, variants are given a score called minor allele frequency (MAF) to retain only rare variants occurring in less than 1% of the population, which are considered most influential in cancer studies (Hintzsche, Robinson, and Tan 2016). Prioritization of detected mutated genes is necessary and can be implemented by looking at individual genes or gene sets. For looking at genes in sets, gene set enrichment analysis (GSEA) is performed directing to variants associated with statistically enriched pathways, functions, and more. Genes not standing out in GSEA can be analyzed separately applying knowledge-driven prioritization (Jalali Sefid Dashti and Gamieldien 2017). In the upcoming years, WES/WGS may start to be used routinely in clinical laboratories for

disease treatments. New standards come with new challenges: vast amounts of variant data has to be integrated with clinical records and patient information in order to allow fast discovery of new variants contributing to disease, obtaining the information and user-friendly decision making for the specialists (Bao et al. 2014).

1.6 Gene enrichment and pathway analysis

Gene set enrichment analysis (GSEA) is applicable to find a connection between the disease phenotype and a group of genes/proteins (Zito et al. 2021). The root of the method's strength is focusing on gene sets (groups of genes) that share similar biological functions, chromosomal locations, or regulation. GSEA shows multiple advantages in comparison with single-gene methods. It eases the interpretation of a large-scale experiment by identifying pathways and processes and additionally, gene sets enable more reproducible results (Subramanian et al. 2005). There are many GSEA methods available with over-representation analysis (ORA) being the most commonly used due to its simplicity and ease of use (Maleki et al. 2020). The gene ontology (GO) represents information about the biological region in relation to three different classes: molecular function shows molecular activities executed by gene products, cellular compartment indicates the localization of main activity in the cell, and biological process gives a broader understanding of programs achieved by multiple molecular activities (Ashburner et al. 2000; "Gene Ontology Overview" n.d.). In addition to detecting over-representation in GO terms, enrichments in pathways also should be inspected, as it provides a compelling prospect capable of enhancing the interpretation of genomic variations and is shown to produce biological insight (Ugur Sezerman et al. 2019). In the process of carcinogenesis aberrations in signal pathways managing the cell cycle, cell death, and cell growth occur, yet the scope, co-existence, and mechanisms of these changes vary depending on the tumor type and individuality (Sanchez-Vega et al. 2018). Pathway enrichment analysis is a statistical method identifying if pathways are noticeably enhanced in a set of genes. The protocol of pathway enrichment analysis comprises three distinct steps: defining the gene list to be analyzed, implementing the pathway enrichment analysis using the chosen tool, and visualizing the results (Reimand et al. 2019). The selection of tools for gene enrichment and pathway analysis is diverse, with all of them with their own unique virtues. The properties of used tools used in present thesis are described as follows. The advantages of g:Profiler are data quality, frequent data updates, technical robustness, as well as availability via many platforms, including web service, Python, R, and Galaxy. The g:Profiler's primary data source is Ensembl, and updates are provided quarterly. However, previous releases are preserved for the reproducibility of the results. The adjusted P-values are calculated using g:SCS method, which is more conservative than Benjamini-Hochberg approach but not as strict as Bonferroni correction (Raudvere et al. 2019; Reimand et al. 2007). Another tool, Enrichr incorporates an impressive amount of resources and covers a vast proportion of data sets. The Enrichr tool's assets are comprehensiveness, result visualization opportunities, and ease of use. For Enrichr tool an adjusted P-value is calculated using the Benjamini-Hochberg correction method (Chen et al. 2013; Kuleshov et al. 2016; "Enrichr Help Center" n.d.). The third tool, Metascape, stands out by its convenient use, ready-to-use visualizations, and easier interpretation. For the functional gene enrichment analysis, Metascape uses the well-known hypergeometric test and Benjamini-Hochberg correction method (Zhou et al. 2019).

2 Aims of the study

The research's importance consists in demonstrating the possible implementation of bioinformatic workflow to find disease-related variants from enormous amounts of data in the case of lacking tumor-normal sample pairs. The purpose of the study is to analyze lung cancer whole-exome sequencing data from primary tumor samples to discover novel mutations involved in cancer recurrence. The present study addresses small genetic alterations, SNPs and INDELs, with the aim of acquiring knowledge of possibly pathogenic variants leading to mutated genes and pathways. The hypothesis is, that undiscovered specific genetic aberrations could affect and stimulate lung cancer metastatic relapse, or the presence of protective mutations could prevent it. New findings in this topic could pave the way for a deeper understanding of the underlying mechanisms, and therefore better prognostic or therapeutic opportunities for patients.

3 Research design and methods

Following chapter describes research setting, analysis workflow and implemented methods. All commands and operations are retained in the data analysis diary for reproducibility of the results. Additionally, same research structure can be applied to any dataset exploring genomic variation.

3.1 Research structure

Initially, the acquired data was prepared for further analysis starting with the data organized into different groups. The first step was done by supervisor Olli-Pekka Smolander prior to this thesis work. Next, SNPs and INDELs were separated to be inspected individually. Data was transformed into suitable format for compatibility with the ANNOVAR tool. ANNOVAR tool was set up in order to utilize it for SNP and INDEL investigation. SNP analysis covered SNP filtering, SNP annotation, and SNP grouping by effect. Using INDEL data, the work proceeded with annotation, filtering and looking at effects by the group. For more comprehensive and reliable data interpretation, gene enrichment analysis was carried out using multiple tools: g:Profiler, EnrichR and Metascape. The overall research structure is elucidated in Figure 7. All data and implemented commands have been documented for reproducibility of the results.

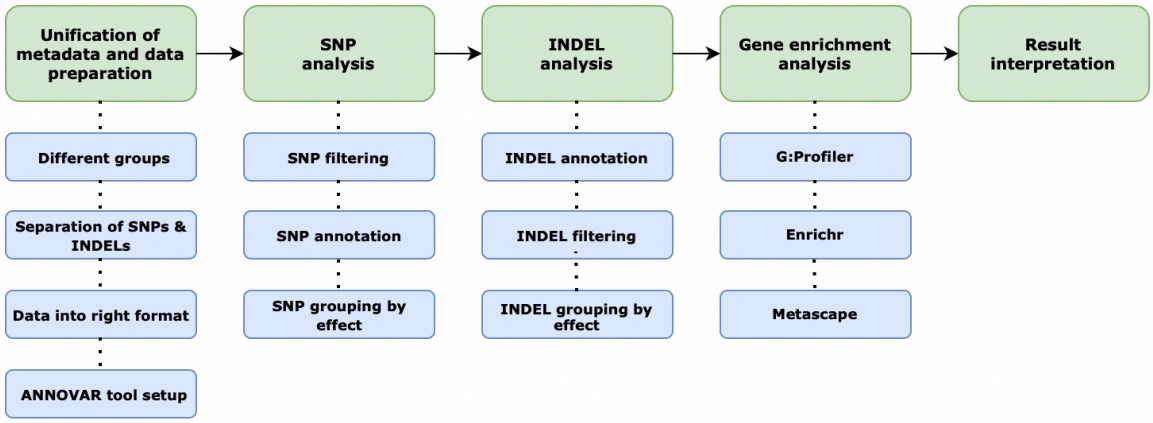


Figure 7. Overall research structure. The work started with preliminary steps: Unification of metadata and data preparation. For first, the samples were distributed into groups by: histology, cancer type, recurrence type, metastases location, extent of metastases and progression-free survival (PFS). Data analysis was started by separating SNPs and INDELs. Next steps were setting data into correct format and setting up ANNOVAR tool. Groundwork finished, next stage was SNP analysis including filtering, annotation and grouping by effect. Next step was investigation of INDELs, covering annotation and grouping by effect. After looking at effects in groups, gene enrichment analysis was performed with both, g:Profiler and EnrichR. Last, but not least important part was result interpretation.

3.2 Background information

3.2.1 Cohort description

Cohort was composed of lung cancer patients with an age between 43-84. The entire number of samples was 94. For three patients, initial and recurrence samples were collected. Additionally, for

three samples clinical information was missing. Histology and cancer type was determined by the North Estonia Medical Centre (NEMC) pathology department. Non-small cell lung cancer (NSCLC) was present in 91% and small cell lung cancer (SCLC) in 9% of the samples. Represented lung cancer subtypes were adenocarcinoma, adenosquamous carcinoma, squamous cell carcinoma, large cell carcinoma or small cell carcinoma. Three samples were evaluated as uncertain by histology. Biopsy samples were preserved by FFPE technique. In addition to current cohort, the analysis of the second batch of data of control group patients has been started, however, it is not included in the scope of present thesis.

3.2.2 Sequencing and variant data

Cancer WES was conducted with Illumina sequencing platform by Intermountain Precision Genomics (USA). Sequencing of 94 samples was attempted, however 12 samples received less than 1 million reads likely due to low-quality libraries. Therefore, data from the remaining 82 samples were applied for further analysis. The average autosomal coverage over the target region was 165x with values ranging from 8.8 to 584.5. Variant calling was implemented by the same company using the Edico DRAGEN (Dynamic Read Analysis for GENomics) (Version 01.011.269.3.2.2-4-g960897cf) tumor-only pipeline. A strand orientation bias filter was used to help with FFPE artifacts. For variant calling, the genome GRGCH38 was used. Unification of metadata and sequencing was implemented by supervisor Olli-Pekka Smolander. For that, NEMC clinical data of patients was combined with SNP data: patient ID-s were added as there were multiple samples from one patient, filenames were matched to sample names, and different groups were formed. Six different groups were produced (Figure 8): cancer type (NSCLC, SCLC), histology (adenocarcinoma, adenosquamous, large cell, small cell, squamous cell), type of recurrence (stage IV, local), metastases grouping (adrenal, brain, hepatic, lymphatic, ossific, other, pleural, pulmonary), extent of metastases (local-metastatic, oligo-metastatic, poly-metastatic) and PFS (less than a year, one to three years, more than three years).

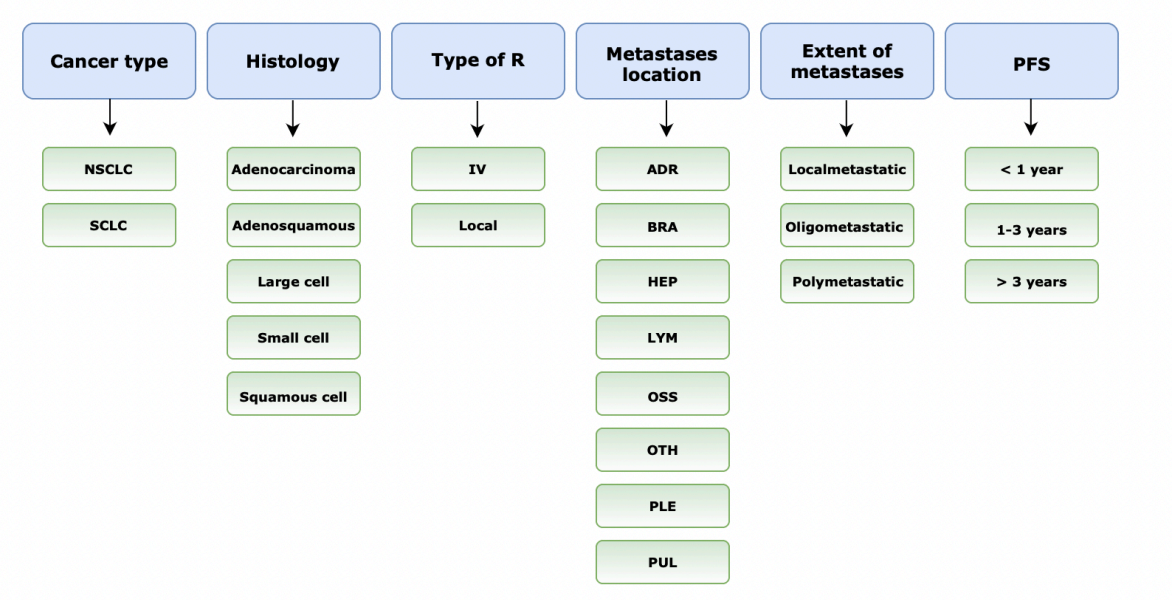


Figure 8. Generated groups for lung cancer samples. Different groups: cancer type (NSCLC, SCLC), histology (adenocarcinoma, adenosquamous, large cell, small cell, squamous cell), type of recurrence (local, stage IV), metastases (adrenal, brain, hepatic, lymphatic, ossific, other, pleural, pulmonary), extent of metastases (local-metastatic, oligo-metastatic, poly-metastatic), PFS (less than a year, one to three years, more than three years). NSCLC refers to non-small cell lung cancer and SCLC to small cell lung cancer. ADR=adrenal, BRA=brain, HEP=hepatic, LYM=lymphatic, OSS=ossific, OTH=other, PLE=pleural, PUL=pulmonary. PFS indicates progression-free survival, which is defined as the time without disease progression or death.

3.3 Data formatting and set-up

The first step of the work was separation of single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs) using VCFtools in Linux command line. Script was written in Nano (command line text editor for Linux operating systems) to allow simple processing of large number of files and after this step, run in the command line. Implementation of the scripts is presented in Figure 9 and denotation of the script contents can be found in Appendix 1.

```
 GNU nano 2.0.9                File: recode_snp_only.run                                Modified

module load HPC1/vcftools-0.1.17
while read line
do

vcftools --gzvcf $(echo $line) --remove-indels --recode --out snp_only.$(echo $line)

done < $1
```

```
 GNU nano 2.3.1                File: recode_indel_only.run                              Modified

module load HPC2/vcftools-0.1.17
while read line
do

vcftools --gzvcf $(echo $line) --keep-only-indels --recode --out indels_only.$(echo $line)

done < $1
```

Figure 9. Separation of SNPs and INDELs. Scripts (recode_snp_only.run and recode_indel_only.run) for separating SNPs and INDELs was written in Nano, made executable and run in the command line.

SNP and INDEL data was converted into a simpler format for ANNOVAR to ensure the compatibility and simpler analysis. A list of files created by VCFtools (contain only SNPs or only INDELs) was made and a script was written in Nano, so that irrelevant columns would be eliminated from each file, retaining only five: chromosome with the SNP/INDEL, SNP/INDEL starting position, SNP/INDEL

ending position, initial nucleotide(s), new nucleotide(s). The implemented commands for this purpose regarding SNP data is shown in Table 1 and denotation of the scripts contents can be found in Appendix 2. For INDELs, python script written by supervisor was used with the aim of retaining only INDELS <50bp, as it is the limit for the use of ANNOVAR.

Table 1. The workflow of putting SNP data into simpler format (e.g., chr1 398 398 A T). A file listing files was generated, a script was written in Nano, the script was made executable and run.

1) Make a list of files	`ls snp_only.*.vcf > snp_only_vcf_files.txt`

2) Write a script in Nano	`while read line`
	`do`
	
	`cat $(echo $line) | grep -v ^# | awk '{print $1,$2,$2,$4,$5}' > snp_positions.$(echo $line).txt`
	
	`done < $1`

3) Make the script executable	`chmod u+x snp_position.run`
4) Run the script	`./snp_position.run snp_only_vcf_files.txt`

3.3.1 ANNOVAR set-up

Next step was installation of ANNOVAR (Wang, Li, and Hakonarson 2010), a software tool for filtering and annotation, along with selected databases: ENSEMBL (Howe et al. 2021), cytoband (BAC Resource Consortium et al. 2001), ExAC03nontcga (Exome Aggregation Consortium et al. 2016), ClinVar (Landrum et al. 2014), COSMIC92-coding (Tate et al. 2019), COSMIC92-noncoding (Tate et al. 2019), gnomAD-211exome (Karczewski et al. 2020), and gnomAD-30genome (Karczewski et al. 2020). The process of downloading databases is shown in Table 2.

Table 2. Downloading ANNOVAR tool and relevant databases. ANNOVAR was downloaded, the contents were extracted, the module was loaded, the software tool was added to PATH, databases were downloaded, and the COSMIC databases were copied from the shared folder.

1) Download ANNOVAR	`wget http://www.openbioinformatics.org/annovar/download/annovar.latest.tar.gz`
2) Extract the contents	`tar xvzf annovar.latest.tar.gz`
3) Load the module	`module load perl5.26.1`
3) Add to PATH	`export PATH=/home/laura.vitsut/software/annovar:$PATH`

4) Download databases	`annotate_variation.pl -buildver hg38 -downdb -webfrom annovar ensGene humandb/`
	
	`annotate_variation.pl -buildver hg38 -downdb cytoBand humandb/`
	
	`annotate_variation.pl -buildver hg38 -downdb -webfrom annovar exac03nontcga humandb/`
	
	`annotate_variation.pl -buildver hg38 -downdb -webfrom annovar clinvar_20210123 humandb/`
	
	`annotate_variation.pl -buildver hg38 -downdb -webfrom annovar gnomad30_genome humandb/`
	
	`annotate_variation.pl -buildver hg38 -downdb -webfrom annovar gnomad211_genome humandb/`

5) Copy COSMIC databases from the shared folder	`cp /home/bioinf-jagatud/hg38_cosmic92_coding.txt .` `cp /home/bioinf-jagatud/hg38_cosmic92_noncoding.txt .`

3.4 SNP analysis

3.4.1 Filtering with population databases

Databases including common population-level variants were applied on the data, as those are not relevant for the discovery of novel cancer-related mutations. Databases used to filter out frequently occurring variants were ExaC03 (Exome Aggregation Consortium) non-TCGA (The Cancer Genome Atlas), gnoMAD (Genome Aggregation Database) v2.1.1 exome, and gnoMAD v3.1 genome. Those databases contain information about the frequency of variants in different populations. TCGA ExAC is derived from matched germline sample of "cancer patients." Non-TCGA excludes these variants and includes only information from healthy samples. In this filtering step, MAF cutoff 0.01 was used as a threshold. This limit is commonly used to filter out frequent variants, which probably are not clinically relevant, as variants with deleterious effects are usually rare. Estonian population variant information was included in gnoMAD v3.1 genome. This was also separately confirmed by contacting Geenivaramu. For first, filtering was conducted with ExaC03. Description of the actions and commands are presented in Figure 10. A file listing all input files, where each file lays on a separate line, was created. Next, a script was written in Nano, named exac03nontcgavol2.run. The purpose of the program was for it to go through the input file line by line until the end. The program was told to do the following for each file: modify the output file name for it not to be too long, filter out variants with MAF >1% using the Exac03-nontcga database. The script was made executable and run, resulting in two separate types of files: dropped and filtered. Dropped files were filtered out and filtered files were used as an input for a next step of the work. Similarly, filtering proceeded with the gnoMAD v2.1.1 exome and gnoMAD v3.1 genome.

Description of the actions	Code
1) Makes a file listing input files	ls *.recode.vcf.txt > snp_position_files.txt
2) Makes a script in nano	script in nano (**exac03nontcgavol2.run**):
Goes through the input file line by line until the end (while read line) Tells what to do each time we go through file (do-done pair)	while read line do
Changes output file name for it not to be too long (OUTFILE variable)	OUTFILE=$(echo $line | sed 's/.*position.//g' | sed 's/_PASS.*//g')
Specifies the action done for each file - filtering, database to use, with hg38 Specifies the location of input files, specify the output file names Specifies the location of database, keep files with MAF <0.01	/home/laura.vitsut/software/annovar/annotate_variation.pl -filter -dbtype exac03nontcga -buildver hg38 /home/laura.vitsut/SNP_folder/SNP_position_folder/$(echo $line) -out $OUTFILE /home/laura.vitsut/software/annovar/humandb/ -otherinfo -score_threshold 0.01
Specifies that the input is a file given when the script is run (< $1)	done < $1
3) Makes the script executable for running	chmod u+x exac03nontcgavol2.run
4) Runs the script	./exac03nontcgavol2.run SNP_position_files.txt
5) Makes a file with the number of SNPs left for each file	wc -l *.hg38_exac03nontcga_filtered > SNP_left_exac03nontcgavol2.csv

Figure 10. Filtering SNPs with ExaC03 population database. Right column presents the commands used to implement the filtering. Left column describes the meaning and contents of each command.

3.4.2 Filtering with cancer databases

The next step of the work used filtered files from gnoMAD v3.1 genome filtering as an input. Commands and working procedures were similar to previous steps, except for the use of argument for allele frequency specification. Firstly used database, ClinVar, aggregates information about genomic variation and its relationship to human health. ClinVar includes both germline and somatic variants, but somatic variants are not well-represented. It evaluates the clinical significance of each

mutation. Again, dropped and filtered files were produced. Filtered files include variants not found in the ClinVar database for the potentiality to discover novel mutations. COSMIC92 coding and non-coding databases were used to filter out already discovered cancer-involved mutations. COSMIC includes somatic mutations reported in the literature in various types of cancers. Using ANNOVAR, a file with mutations is scanned against the database. This provides information concerning previous knowledge about the mutations, also their appearance and incidence in different cancer types. Filtered files from the previous step were used as an input and implementation was similar to filtering with ClinVar.

3.4.3 Gathering distinct SNPs

Before proceeding to annotation, distinct SNPs were gathered to one list and all samples where particular SNP occurred were listed for that SNP (overall procedure is shown in Table 3). For that purpose two files were produced: one listing previously filtered files and second listing only sample names. Both of those files were merged into one. The Python script was generated by supervisor Olli-Pekka Smolander.

Table 3. Gathering distinct SNPs and collecting information about samples in which they were present. Right column depicts the commands used to implement the process. Left column describes the aim of the command.

1) Make a file listing previously filtered files	`ls simple.*noncoding*.csv > csv_filenames.txt`
2) Make a file listing only sample names	`ls simple.*noncoding*.csv | sed 's/.*only.//g' | sed 's/.exac.*//g' > samplenames.txt`
3) Merge those two files into one	`paste csv_filenames.txt samplenames.txt > filenames_with_samplenames.txt`
4) Copy the python script from the shared folder	`cp /home/bioinf-jagatud/variants_with_samplenames.py/ /home/laura.vitsut/SNP_folder/SNP_simple_annotation`
5) Load the module	`module load python-3.8.7+jupyter`
6) Run the python script	`python variants_with_samples.py filenames_with_samplenames.txt variants_withsamplenames.csv`

3.4.4 Annotation and gathering SNPs by distinct effects

Next, the SNPs were annotated with ANNOVAR and grouped by distinct effects on particular genes. The process started with annotation of variants and associated sample names. The sample names were put into one file and the header ("Samples") was added. The two files, sample names and annotations, were merged into one using paste command. The columns were extracted, sorted and the duplicates were removed. The python script was constructed by supervisor Olli-Pekka Smolander. Used commands are seen in Table 4.

Table 4. Annotation and gathering SNPs by distinct effects. Left column shows the purpose of each step and command. Right column demonstrates used commands.

1) Annotation of variants and associated sample names	`/home/laura.vitsut/software/annovar/table_annovar.pl variants_with_samplenames.csv -out variants_with_samplenames_annotated /home/laura.vitsut/software/annovar/humandb/ -buildver hg38 -remove -protocol ensGene,refGene,dbnsfp41a,cytoBand -operation g,g,f,r -nastring . -csvout -polish`
2) Put sample names in one file and add header 'Samples')	`echo Samples | cat - variants_with_samplenames.csv | cut -d' ' -f 6 > variant_samplenames.csv`
3) Merge two files (sample names + annotated) into one	`paste -d',' variants_with_samplenames_annotated.hg38_multianno.csv variant_samplenames.csv > variants_annotated_with_samplenames.csv`
4) Extract the columns, sort, remove duplicates	`cat variants_annotated_with_samplenames.csv | cut -d',' -f 6,7,9 | sort | uniq > unique_effects_in_variants_with_annotations_and_samplenames.csv`
5) Copy the python script from the shared folder	`cp /home/bioinf-jagatud/final_variants_by_effect.py /home/laura.vitsut/SNP_folder/SNP_simple_annotation`
6) Load the module	`module load python-3.8.7+jupyter`
7) Run the python script	`python final_variants_by_effect.py unique_effects_in_variants_with_annotations_and_samplenames.csv variants_annotated_with_samplenames.csv variants_grouped_by_effects.csv`

3.4.5 Looking at effects by group in SNP data

The subsequent step of the work was to separate the effects by group: cancer type (NSCLC, SCLC), histology (adenocarcinoma, adenosquamous, large cell, small cell, squamous cell), type of recurrence (stage IV, local), metastases grouping (adrenal, brain, hepatic, lymphatic, ossific, other, pleural, pulmonary), extent of metastases (local-metastatic, oligo-metastatic, poly-metastatic) and PFS (less than a year, one to three years, more than three years). A file listing all samples within a specific group (e.g adenocarcinoma) was created. Files without clinical information were excluded. Anaconda and Scipy (Scientific Computing Library) were installed and added to PATH. Subsequently, a script was composed with Nano listing commands for each group in order to provide group-specific SNP information. The python script was generated by supervisor Olli-Pekka Smolander. The P-value for enrichment of given effect in a group under observation was calculated using hypergeometric distribution.

Table 5. Dividing variants into specific groups. The left column shows the aims of the commands and the right column depicts implemented commands.

1) Make files listing all samples within a specific group (e.g. adenocarcinoma)	`Nano Adenocarcinoma_samples.txt`
2) Install Anaconda	`cp /home/bioinf-jagatud/Anaconda3-2020.11-Linux-x86_64.sh /home/laura.vitsut bash ~/Anaconda3-2020.11-Linux-x86_64.sh`
3) Install Scipy	`conda install -c anaconda scipy`
4) Add it to PATH	`export PATH=/home/laura.vitsut/anaconda3/bin:$PATH`
5) Create a script in nano listing commands for each group	`nano different_groups_SNPs.run (example of one line: python effect_in_group.py Adenocarcinoma_samples.txt variants_grouped_by_effects.csv > Adenocarcinoma_variants.txt)`
6) Make the script executable	`chmod u+x different_groups_SNPs.run`
7) Run the script	`./different_groups_SNPs.run`

3.5 INDEL analysis

3.5.1 Gathering distinct INDELs

INDEL analysis started with gathering distinct INDELs to one list and collecting samples where they were present (Table 6). For that, two files were created, one listing all indel files and other listing sample names. Subsequently, those two files were merged into one. The python script originated from supervisor Olli-Pekka Smolander. Finally, the module was loaded and the script was run.

Table 6. Gathering distinct INDELs. The first column represents the purpose of each command. The second column depicts used commands in Linux system.

1) Make a file listing all indel files	`ls *.bed >filenames.list`
2) Make a file listing sample names	`cat filenames.list | sed 's/.*only.//g' | sed 's/_PASS.*//g' > samplenames.list`
3) Merge those two files into one	`paste filenames.list samplenames.list > filenames_w_samplenames.list`
4) Copy the python script from the shared folder	`cp /beegfs/home/bioinf-jagatud/indels_with_samples.py`
5) Load the module	`module load python-3.8.7+jupyter`
6) Run the python script	`python3 indels_with_samples.py filenames_w_samplenames.list indels_with_samples.csv`

3.5.2 Annotation of INDELs

Indels were annotated with ANNOVAR using the Ensembl database. Sample names were added to one file and a header 'Samples' was attached. The file with annotations and the file with sample names were merged together using the paste command. Commands utilized to implement these steps are shown in Table 7.

Table 7. Annotation of INDELs. The left column demonstrates conducted steps and right column shows implemented commands.

1) Load the module	`module load perl5.26.1`
2) Annotate the INDELs	`/ceph/home/laura.vitsut/software/annovar/table_annovar.pl indels_with_samples.csv -out indels_with_samplenames_annotated /ceph/home/laura.vitsut/software/annovar/humandb/ -buildver hg38 -remove -protocol ensGene,refGene,dbnsfp41a,cytoBand -operation g,g,f,r -nastring . -csvout -polish`
3) Put sample names in one file and add header 'Samples'	`echo Samples | cat indels_with_samples.csv | -d' ' -f6 > indel_samplenames.csv`
4) Merge two files (sample names + annotated) into one	`paste -d',' indels_with_samplenames_annotated.hg38_multianno.csv indel_samplenames.csv > indels_annotated_with_samplenames.csv`

3.5.3 Filtering of INDELs

Similarly to SNP filtering, INDELS were first filtered with population databases using ExaC03 (Exome Aggregation Consortium) non-TCGA, gnoMAD (Genome Aggregation Database) v2.1.1 exome, and gnoMAD v3.1 genome applying MAF cutoff <0.01. Following the elimination of frequently occurring INDELs in the population, removal of already published mutations was conducted using ClinVar and COSMIC92 coding as well as non-coding.

3.5.4 Gathering INDELs by distinct effects

For the purpose of collecting all INDELs into one file by their effects on particular genes, the lastly produced file with the remaining INDELs needed conversion from tab-separated values (TSV) to comma-separated values (CSV). The columns were extracted, sorted and duplicates were removed. Subsequently, the python script provided by the supervisor was run. Used commands are shown in Table 8.

Table 8. Gathering INDELs by their distinct effects on particular genes. The left column shows the aims of used commands and the right column indicates the used commands.

Convert from TSV to CSV	`cat indels_filtered_with_cosmic92noncoding | sed 's/\t/,/g' > indels_filtered_annotated_with_samplenames.csv`
Extract the columns, sort, remove duplicates	`cat indels_filtered_annotated_with_samplenames.csv | cut -d',' -f 6,7,9 | sort | uniq > unique_effects_in_indels_with_annotations_and_samplenames_vol2.csv`
Copy the python script	`cp /gpfs/mariana/home/laura.vitsut/SNP_folder/ SNP_simple_annotation/final_variants_by_effect.py .`
Load the module	`module load python-3.8.7+jupyter`
Run the python script	`python3 final_variants_by_effect.py unique_effects_in_indels_with_annotations_and_samplenames_vol2.csv indels_filtered_annotated_with_samplenames.csv indels_grouped_by_effects_vol2.csv`

3.5.5 Looking at effects by group in INDEL data

This step of the work was carried out similarly to SNPs. Firstly, files listing all samples within specific groups were copied from SNP folder data. A script was written in Nano listing commands for each group. Within the script, the python script was applied (provided by supervisor). Anaconda was added to path and the module was loaded. After that, the script was made executable and run.

3.6 Gene enrichment analysis

Gene enrichment analysis was conducted separately for each group using generated gene lists in order to possibly view if sets of mutated genes are overrepresented in GO terms or pathways. Work started with setup for gene enrichment – only mutations with P-value under 0.05 (<5%) were retained. A strict group of variants was formed, including exonic non-synonymous, exonic start-loss, exonic stop-loss, exonic stop-gain variants, plus frameshift variants in the case of INDELs. The gene lists were manually curated to prevent false-positive overrepresentation, hence some double readthrough genes were removed. All of the gene lists were downloaded and run with gene enrichment tools. For enrichment analysis, INDELs and SNPs were combined into one list and run conjointly as well as investigating SNPs and INDELs separately. Gene enrichment analysis was performed with freely available tools: g:Profiler, Enrichr. The g:Profiler allowed the possibility to insert gene lists in an ordered way, Enrichr did not have this opportunity. For evaluation of protein-protein interactions Metascape tool was used.

4 Results

4.1 SNP analysis results

The total number of SNPs before the filtering step was 2 362 746 with an average of 28 814 per sample. After applying population databases with MAF <0.01 cutoff, the majority of SNPs (~83%) were eliminated retaining 399 756 SNPs (~17%), while the mean number of SNPs in this phase was 4875 per sample. Further filtering was exerted using known cancer databases, including germline and somatic mutations, in order to eliminate previously known and published SNPs. Subsequent to the last filtering step, the number of SNPs was down to 346 280 preserving ~15% of the initial quantity of SNPs with an average of 4223 per sample. The numbers indicating SNP filtration results for each sample are shown below in Figure 11. The median numbers of SNPs in local and metastatic groups were 30 164 and 29 272 before the filtering and 3521 and 4094 after the filtering, respectively. The Mann-Whitney U test was used to determine whether the observed differences were statistically significant. The corresponding P-values were 0.81034 and 0.08544. Therefore, the differences were not statistically significant. Despite the fact that the P-value after the filtering does not reach the significance level of 0.05, a difference in median values is much larger after the filtering. With larger number of samples in local group, statistical significance would probably have been reached even with this non-parametric test. As a result of SNP annotation, the captured 24 distinct elements and effects were: UTR5, UTR3, UTR5-UTR3, upstream, upstream-downstream, downstream, splicing, ncRNA-splicing, ncRNA-intronic, ncRNA-exonic, ncRNA-exonic-splicing, intronic, intergenic, exonic-splicing nonsynonymous SNV, exonic-splicing unknown, exonic-splicing synonymous, exonic-splicing stoploss, exonic-splicing stopgain, exonic-nonsynonymous SNV, exonic-synonymous SNV, exonic-stopgain, exonic-stoploss, exonic-startloss, exonic-unknown (Figure 12).

4.2 INDEL analysis results

Annotation detected 27 following distinct elements and effects: UTR5, UTR3, UTR5-UTR3, upstream, upstream-downstream, downstream, splicing, ncRNA-splicing, ncRNA-intronic, ncRNA-exonic, ncRNA-exonic-splicing, intronic, intergenic, exonic-frameshift substitution, exonic-nonframeshift substitution, exonic nonsynonymous, exonic synonymous, exonic-stopgain, exonic-stoploss, exonic-startloss, exonic-unknown, exonic-splicing frameshift substitution, exonic-splicing nonframeshift substitution, exonic-splicing nonsynonymous, exonic-splicing synonymous, exonic-splicing stopgain, exonic-splicing unknown (Figure 13). The top three most occurring elements were intronic, exonic frameshift substitution, and exonic nonframeshift substitutions. The total number of INDELs in all samples was 376 180 before filtering, with an average of 4588 per sample. With the implementation of databases incorporating frequent population-derived INDELs, the amount of INDELs went down to 213 792, comprising 57% of the initial number. The quantity of INDELs did not further decline by the exploitation of cancer databases with known genetic alterations (Figure 14). By the end of the last filtering, the mean number of INDELs was 2607 per sample. The median numbers for local and metastatic groups were 3677 and 4860,5 before the filtering and 1539 and 2572 after the filtering. Differences in both cases were statistically significant with P-values 0.0466 and 0.01788, respectively.

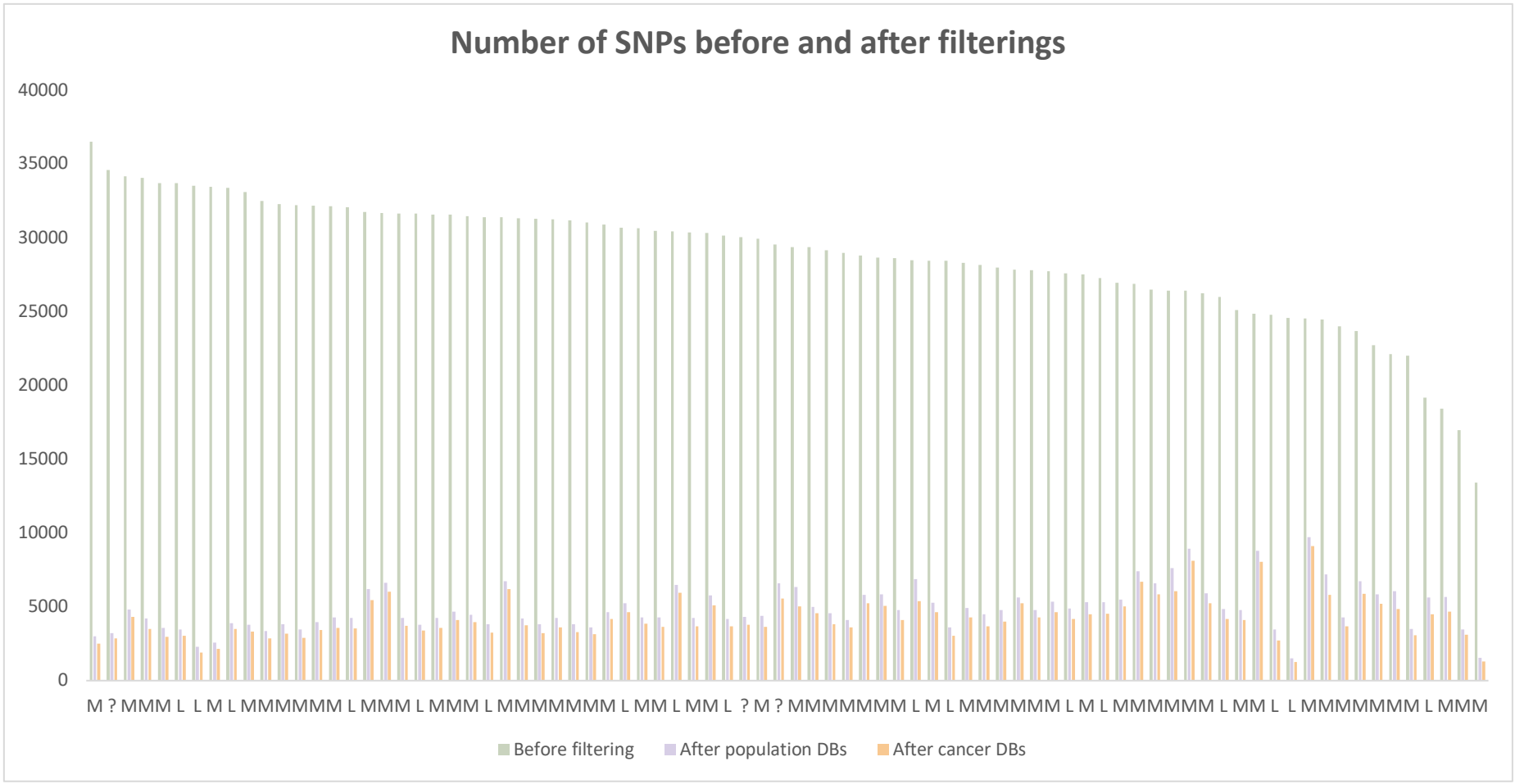


Figure 11. SNP amount before and after filtering. The graph depicts the quantity of SNPs for each sample before filtering, after filtering using population databases with MAF cutoff <0.01, and after filtering with cancer databases. Samples are labeled as "L" or "M", which stands for belonging into local or metastatic recurrence group, respectively. Samples labeled with question marks missed clinical information. Samples are ordered based on the initial number of SNPs. The amount of SNPs declined extensively after applying filtering criteria with an aim to eliminate frequently occurring variants from the dataset. With the starting number of variants being fairly fluctuating and diverse, the final set's differences were more conformed. Differences between groups were not statistically significant (P=0.81034 before filtering and P=0.08544 after filtering).

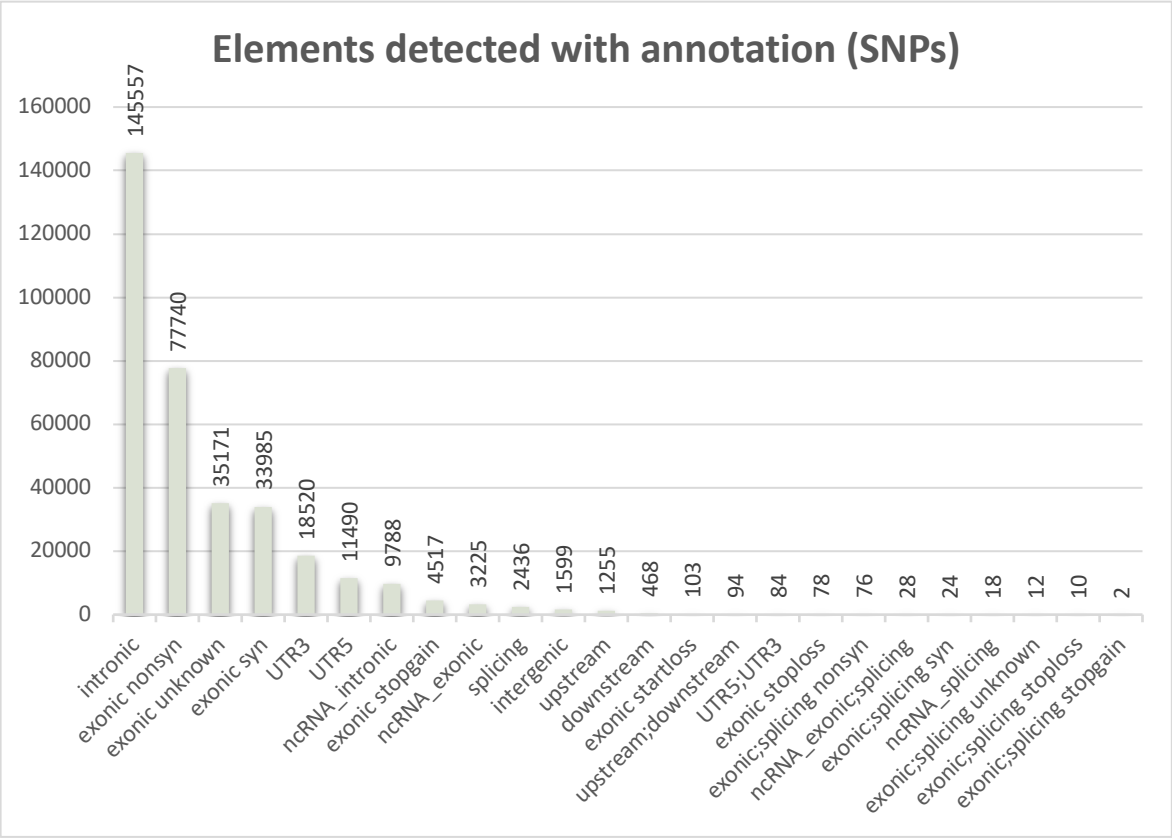


Figure 12. The number and percentage of annotated SNPs with regard to different elements and effects. The top three most abundant elements were intronic, exonic nonsynonymous, and exonic unknown.

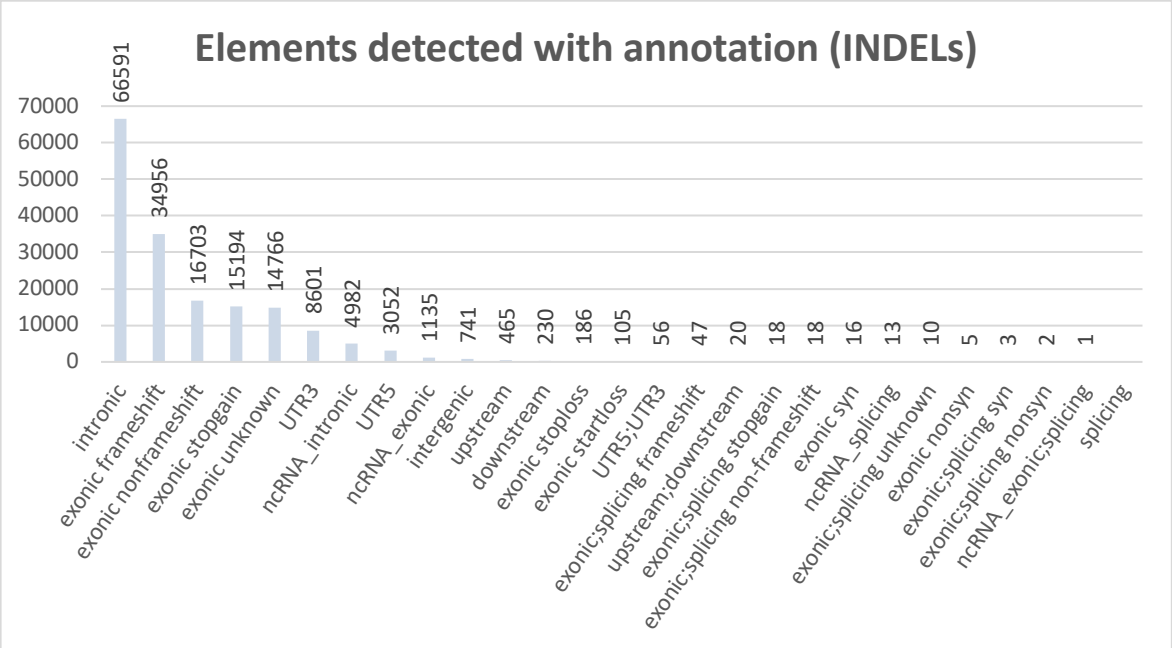


Figure 13. Annotated INDELs with all detected elements and effects. Intronic, exonic frameshift substitutions and exonic non-frameshit substitutions belonged to the top of the table with the highest incidence.

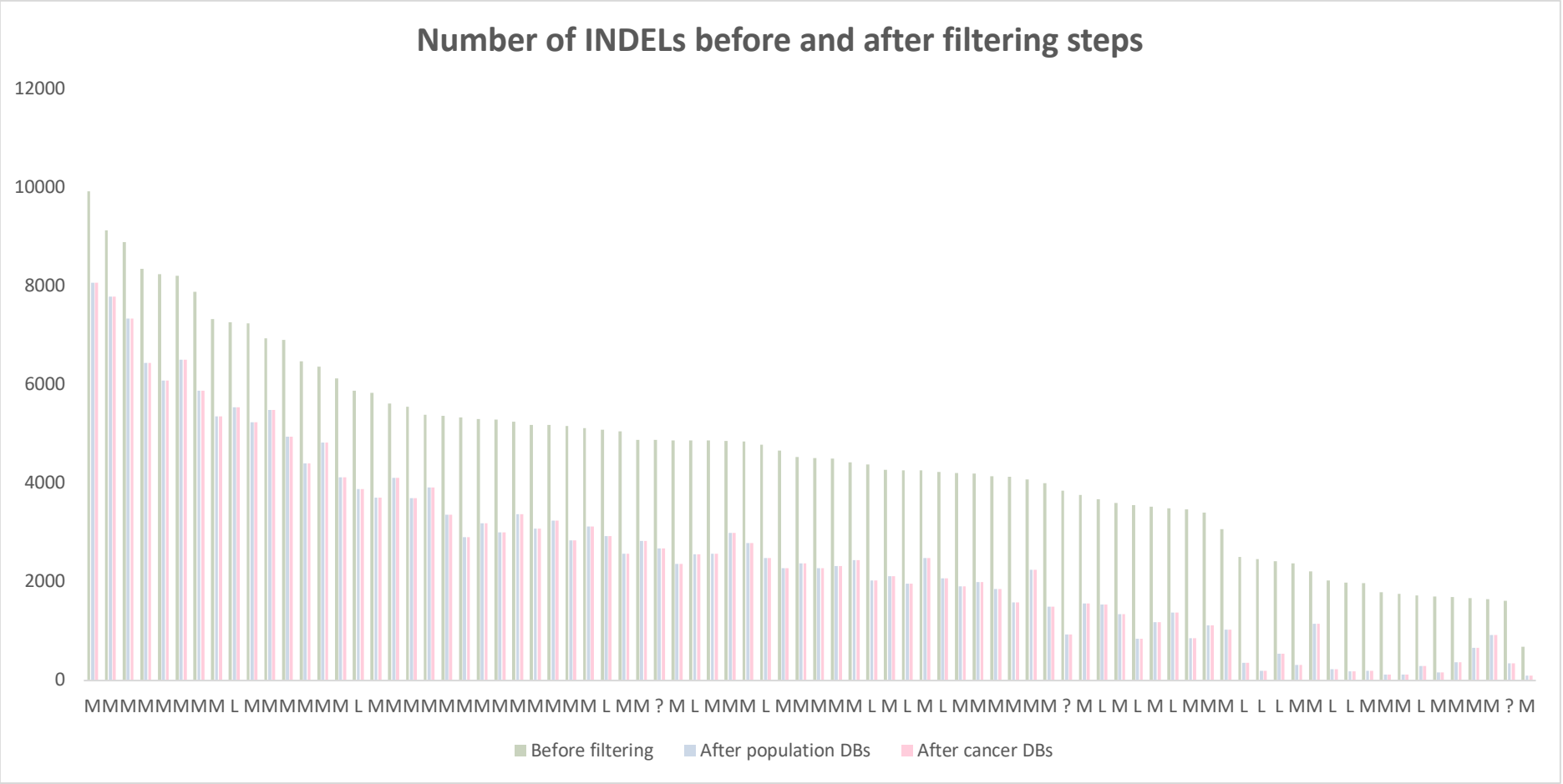


Figure 14. INDEL amount before and after filtering. The bars shown on the graph implicate the volume of INDELs in all 82 samples before filtering and after filtering with population and cancer databases. Samples are labeled as "L" or "M", referring to them belonging to local or metastatic recurrence group, respectively. Samples labeled with question marks missed clinical information. The application of population databases dropped a proportion of INDELs, however applying cancer databases filtration did not give any extra effect. Differences between the groups were statistically significant (P=0.0466 before filtering and P=0.01788 after filtering).

4.3 Genetic characteristics

The number of significantly mutated genes containing INDELs was 38 in the aggressive recurrence group and 98 in the local recurrence group. The number of significantly (P-value < 0.05) enriched genetically altered genes containing single-nucleotide changes was 63 and 179 for the metastatic recurrence and local recurrence, respectively. Investigation of gene lists revealed multiple overlapping genes exhibiting both SNPs and INDELs, with DMXL2 occurring in the metastatic recurrence group exclusively. DMXL2 possessed non-synonymous SNVs with a P-value of 0.002 and frameshift substitutions with a P-value of 0.017. In the aggressive recurrence group, ~32% and ~22% of samples contained the named effect produced by SNPs and INDELs, respectively. The local recurrence group displayed four co-existent variations, such as CCDC181, HNRNPU, NBPF26, and ZNF311. However, those mutations were only borderline significant or could be detected in some of the metastatic recurrence samples as well. Gene CCDC168 was found mutated in high number of samples (50) and had high number of mutations (86). It was found in 67% of the metastatic group samples. However, as the same mutation was found also in several local recurrence group samples (8), the over-representation P-value was only borderline significant (0.049966). Hence, CCDC168 mutation seems to lack a role as a factor predicting distant recurrence, but its overall function in lung cancer recurrence, whether distant or local, is yet to be investigated. The altered gene with the smallest P-value (0.002) occurring merely in the aggressive relapse group was ABCC9 with 13 frameshift substitutions, appearing in ~32% of the metastatic recurrence 2group's samples.

4.4 Gene enrichment analysis

4.4.1 SNPs only

Inspecting gene enrichment analysis results for SNPs only, the discrepancy between metastatic (Figure 15) and local (Figure 16) groups was notably clear, with only single enrichment occurring in the local group. The only over-represented term was related to transcription factor SPI1 possible binding motif with an adjusted P-value of 2.421×10^{-2}. In the aggressive recurrence group, multiple enriched GO classes and pathways were present, with most being associated with binding, organelle localization, cell projection, calcium channel activity, and ion homeostasis. Terms with the most prominent P-value were found in GO Cellular Component 'plasma membrane-bounded cell projection' with an adjusted P-value of 1.106×10^{-5} (g:Profiler) and 'cell projection' with an adjusted P-value of 2.441×10^{-5} (g:Profiler). Participants detected in those two terms were C2CD3, GNAS, PRR12, SEZ6, CLASP2, DOCK7, KIF21B, NPC1L1, CABYR, DIAPH1, ESPN, PTPRH, MAP4, UNC13B, SACS, MTOR, EXOC8, MTSS1, MYO6, MYO7A, which constitute about one-third of the group's genes. Such findings could indicate affected genes' role in cell migration, which is the foundation of metastasis and cancer progression. Many terms related to Ca^{2+} transport were salient, such as GO Molecular Function 'calcium-release channel activity', 'ligand-gated calcium channel activity', as well as GO Biological Process 'release of sequestered calcium ion into cytosol', 'negative regulation of sequestering of calcium ion', 'regulation of sequestering of calcium ion', and 'sequestring of calcium ion' with an adjusted P-values varying between 4.482×10^{-2} and 1.094×10^{-2} (g:Profiler). For the GO Molecular Function terms the related genes were ITPR3 and RYR3, while for the GO Biological Process the affected genes were ITPR3, RYR3, CHD7, and DIAPH1. The GO

Cellular Component term 'sarcoplasmatic reticulum' was also present with an adjusted P-value of 7.783×10^{-3} (g:Profiler) or 0.045 (Enrichr), while genes involved were ITPR3, RYR3, and STIM1.

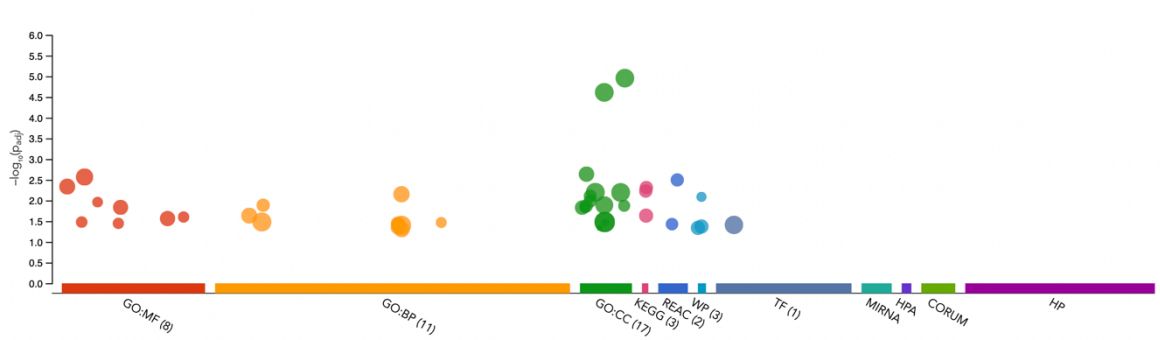


Figure 15. Enriched terms generated by g:Profiler tool in the aggressive recurrence group for SNPs. More over-representations occur when compared to local group, with the most enrichments falling into GO Molecular Function, GO Biological Process, GO Cellular Component terms.

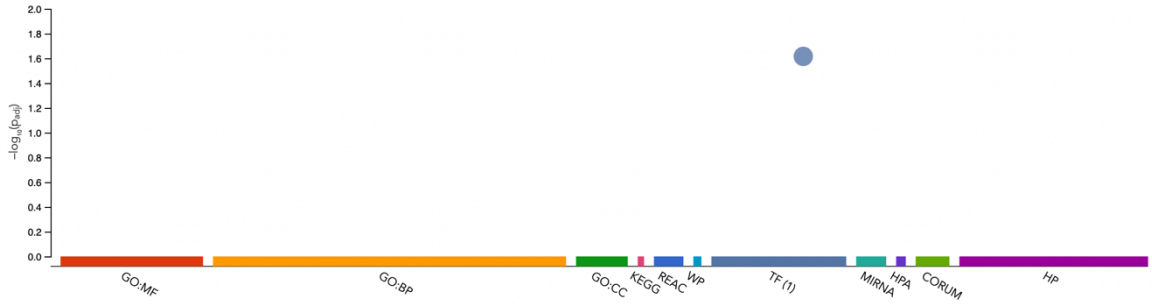


Figure 16. Gene enrichment analysis provided by g:Profiler tool in the local recurrence group for SNPs. Only one enrichment showed up in the TF (transcription factor) term.

4.4.2 INDELs only

Compared to SNPs, gene enrichment analysis of INDELs resulted in less overall enrichments. Nevertheless, more over-representation was detectable in the distant recurrence group conferred to the local recurrence group. Metastatic recurrence group enrichment patterns are shown in Figure 17 and local group enrichments in Figure 18. In the aggressive relapse group, microtubule and cytoskeleton-related terms were mainly prevalent. Terms with the highest significance were GO Cellular Compartment 'kinetochore microtubule' and 'microtubule cytoskeleton', with an adjusted P-values of 5.753×10^{-7} (g:Profiler) and 6.028×10^{-5} (g:Profiler), respectively. First enrichment was produced by genes KNTC1, CLASP1, and CENPE and second by genes KNTC1, CLASP1, MAP1A, CENPE, LYST, NR3C1, RTTN, RIF1, CEP97, EML6, KIAA0586, and DNAH9. In this group, three genes occurring in many distinct terms were KNTC1, CLASP1, and CENPE, showing their role in multiple activities. Enrichments in the local recurrence group were characterized by the presence of regulator activity classes, such as GO Molecular Function 'GTPase regulator activity' and GO Biological Process 'regulation of molecular function' with an adjusted P-value of 2.346×10^{-3} (g:Profiler) and 3.276×10^{-4} (g:Profiler).

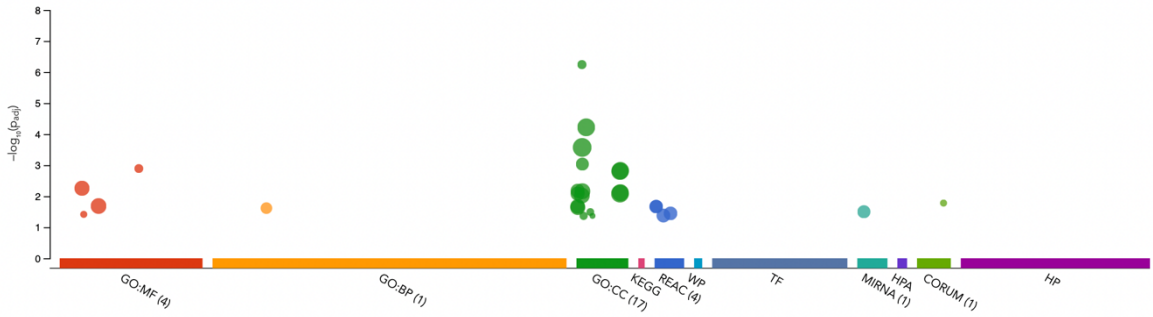


Figure 17. Gene enrichment analysis results provided by g:Profiler in the distant recurrence group for INDELs. Enrichment occurred in GO Molecular Function, GO Biological process, GO Cellular Component, and Reactome terms.

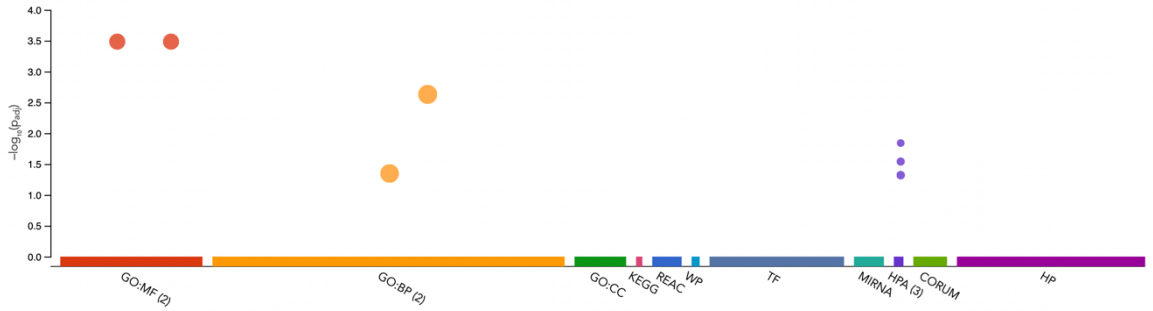


Figure 18. Gene enrichment analysis results generated by g:Profiler in the local recurrence group for INDELs. GO Biological process and GO Molecular function contained some enriched terms.

4.4.3 SNPs and INDELs combined

Enrichment with G:Profiler revealed significantly more over-representation of genes in numerous GO terms, KEGG, WikiPathway, and Reactome pathways within metastatic recurrence group compared to local recurrence group (Figure 19 and 20).

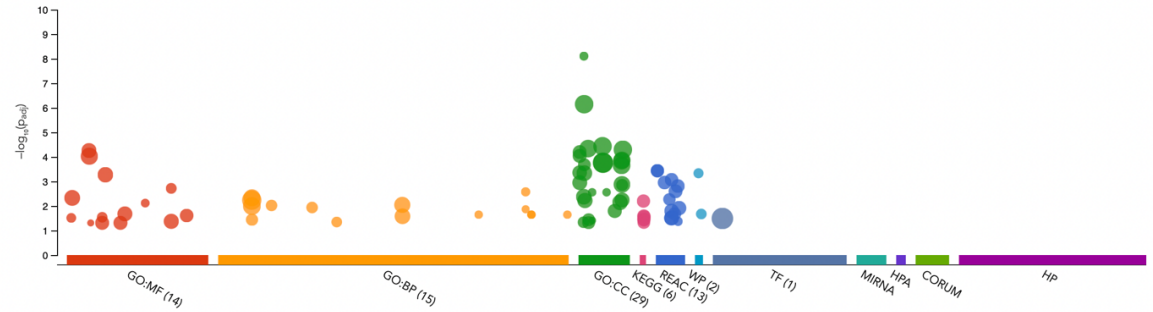


Figure 19. Gene enrichment results for the metastatic recurrence group integrating SNPs and INDELs generated by the G:Profiler. Genes are enriched in multiple GO terms, such as molecular function, biological process, cellular compartment and Reactome pathways.

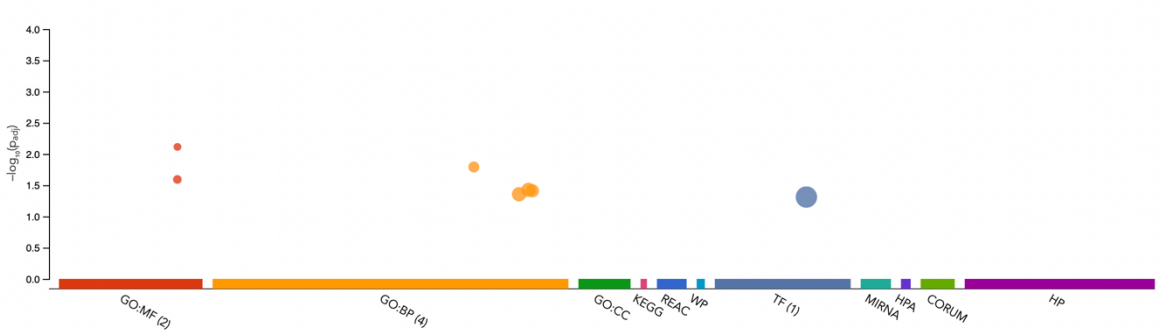


Figure 20. Gene enrichment results for the local recurrence group integrating SNPs and INDELs generated by the G:Profiler. Local, non-aggressive recurrence group exhibits much less enrichments with only few belonging to GO molecular function or biological process group.

Gene enrichment analysis was conducted inspecting SNPs and INDELs separately, as well as combined, and over-representation was even more enhanced in the case of consolidation. Gene enrichment analysis of the local recurrence group revealed much fewer enrichments, with most of them kindred to calcium ion transport. Following terms were enriched in the local group with adjusted P-values ranging from 4.402 x 10^{-2} to 7.690x10^{-3}: GO Molecular Function 'voltage-gated calcium channel activity involved in AV node cell action potential', 'voltage-gated calcium channel activity involved in cardiac muscle cell action potential' with involved genes CACNB2 and CACNA1G, as well as GO Biological Process 'positive regulation of cation transmembrane transport' with genes JPH2, WNK2, ABL1, CACNB2, ATP2A1, 'positive regulation of calcium ion transmembrane transport' with genes JPH2, ABL1, CACNB2, ATP2A1 and 'regulation of calcium ion transmembrane transport' with genes JPH2, ABL1, CACNB2, PIK3CG, ATP2A1. In the distant recurrence group, many enriched GO terms were related to microtubules, binding, mitosis, and calcium channel activity. The most frequently enriched pathways were associated with kinetochores and mitosis. Cancer-inherent signal was prominent in WP 'MFAP5-mediated ovarian cancer cell motility and invasiveness' with an adjusted P-value of 4.642 x 10^{-4} (g:Profiler) or 0.0029 (Enrichr) with involved genes CREB1, ITPR3 and RYR3. Therefore, mutations within these genes could indicate their possible collective contribution to invasiveness in lung cancer. The enriched term with the highest significance was GO Cellular Compartment 'kinetochore microtubule' occurring with an adjusted P-value of 7.855 x 10^{-9} (g:Profiler) or 0.0000024 (Enrichr), encompassing genes KNTC1, CLASP1, CLASP2, and CENPE. Aforementioned findings are illustrated by Volcano plot in Figure 21.

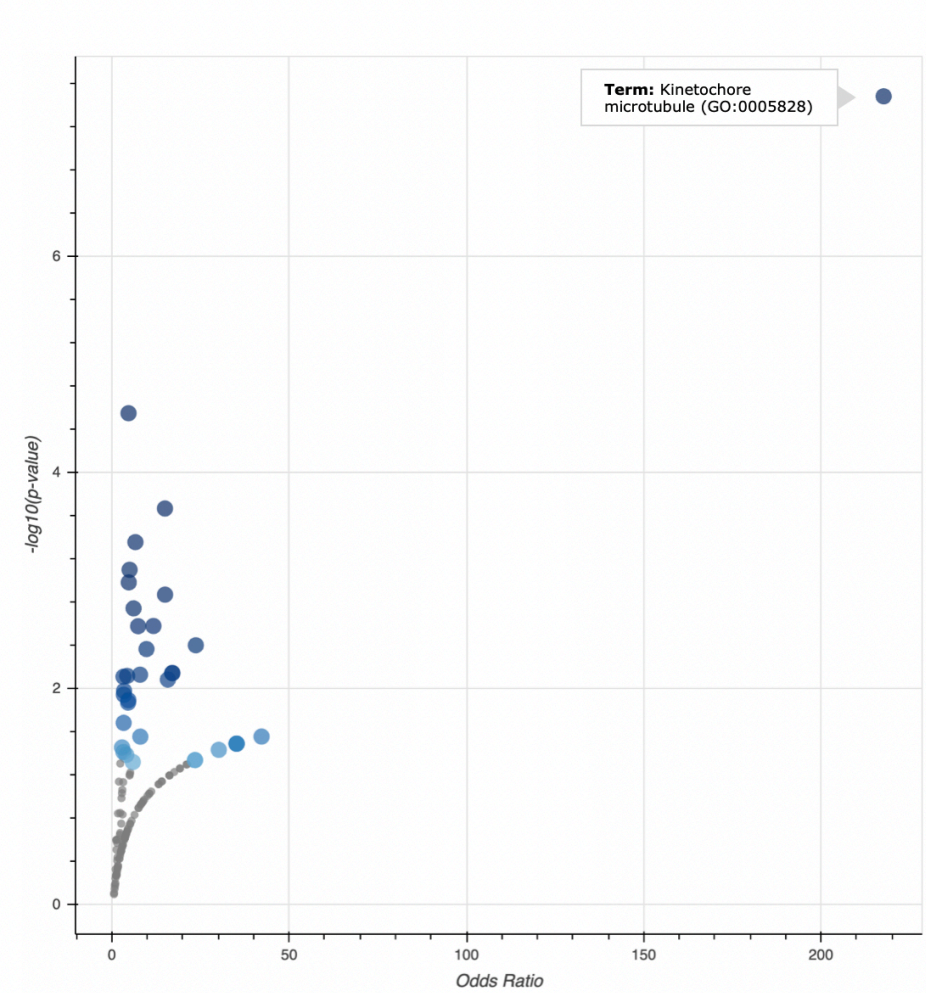


Figure 21. Volcano plot representing enriched gene sets in GO Cellular Component term generated with Enrichr. One term, 'kinetochore microtubule' stood out from others with high significance, comprising of CENPE, KNTC1, CLASP1, CLASP2.

Investigating GO Biological Process term, 'exit from mitosis' outstood with a possible relation to invasiveness, exhibiting an adjusted P-value of 9.528×10^{-3} (g:Profiler) and passing genes KNTC1, CLASP1, and CLASP2. Thus, genetic aberrations in those genes could endorse the process of cells being unable to exit mitosis and start proliferating uncontrollably. GO Molecular Function was also examined, which showed salient enrichment in 'microtubule binding' with an adjusted P-value of 5.527×10^{-5} (g:Profiler) or 0.0000269 (Enrichr) as well as similar 'microtubule plus-end binding' with an adjusted P-value of 1.921×10^{-3} (g:Profiler) or 0.00575 (Enrichr). The latter findings are illustrated by Manhattan plot in Figure 22. In the first term, the appearing genes were CLASP1, MAP1A, CENPE, CLASP2, KIF21B, NEFM, STIM1 detected by g:Profiler, while Enrichr identified additionally CAMSAP3, MAP4, and EML6. The second term comprised a fewer number of genes, such as STIM1, CLASP1, and CLASP2 coinciding by both tools. As microtubules are known to have a function in cell movement, motility, and division, genetic alterations in genes related to microtubules are intriguing to behold. The hypothesis is, that various genes detected by our analysis could conjointly affect cellular processes and pathways giving rise to more aggressive properties of cancer, possibly leading to metastatic nature.

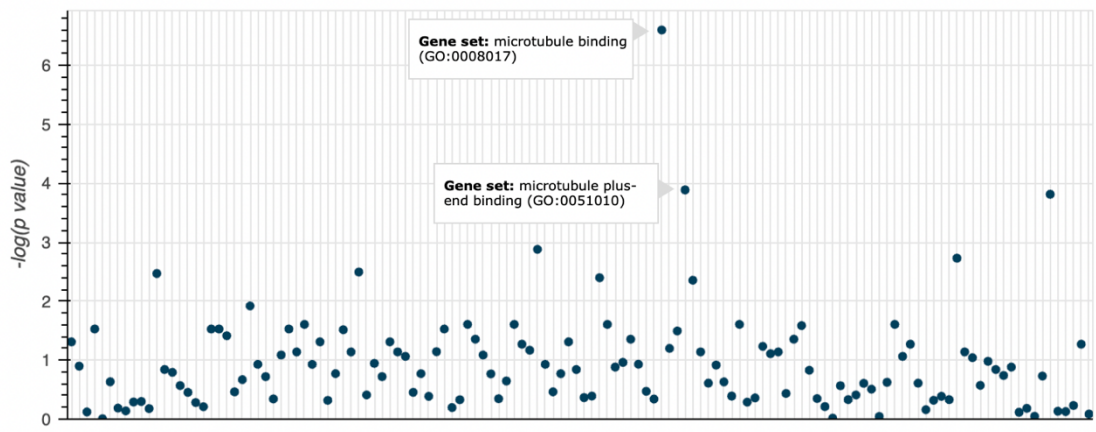


Figure 22. Manhattan plot depicting enriched gene sets in GO Molecular Function term generated by Enrichr. 'Microtubule binding' term distinguished compared to others comprising of CAMSAP3, CENPE, EML6, STIM1, MAP1A, NEFM, KIF21B, MAP4, CLASP1, CLASP2.

Evaluation of protein-protein interactions in the aggressive relapse group detected the associations between CENPE, KNTC1, CLASP1, and CLASP2, reinforcing their possible part in collectively affecting microtubule changes (Figure 23). Overall protein interaction networks were enriched in the following GO terms: 'organelle localization' (P-value $10^{-9.7}$), 'establishment of organelle localization' (P-value $10^{-9.7}$), and 'cell division' (P-value $10^{-7.2}$). More specifically, aforementioned four genes were enriched in following terms: 'amplification of signal from unattached kinetochores via a MAD2 inhibitory signal' (P-value $10^{-10.0}$), 'amplification of signal from kinetochores' (P-value $10^{-10.0}$), and 'mitotic spindle checkpoint' (P-value $10^{-9.7}$). Enrichments in the local group regarding protein interactions were also detected, however the over-representation P-values were larger (Figure 24). Detection of enriched terms exhibited following findings: 'T-cell activation' (P-value $10^{-6.0}$), 'membrane trafficking' (P-value $10^{-5.4}$), and 'lymphocyte proliferation' (P-value $10^{-5.4}$).

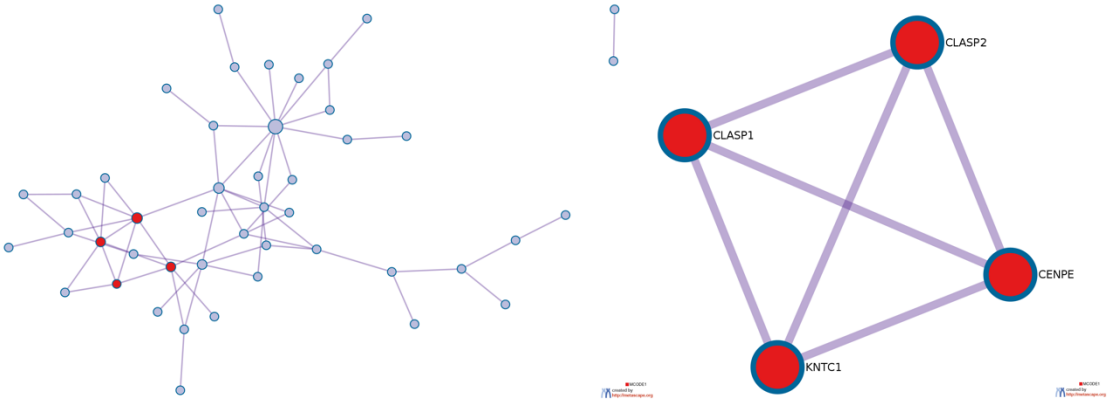


Figure 23. Protein interactions generated with Metascape tool. Detection of physical protein-protein interactions in the metastatic recurrence group examining SNPs and INDELs combined.

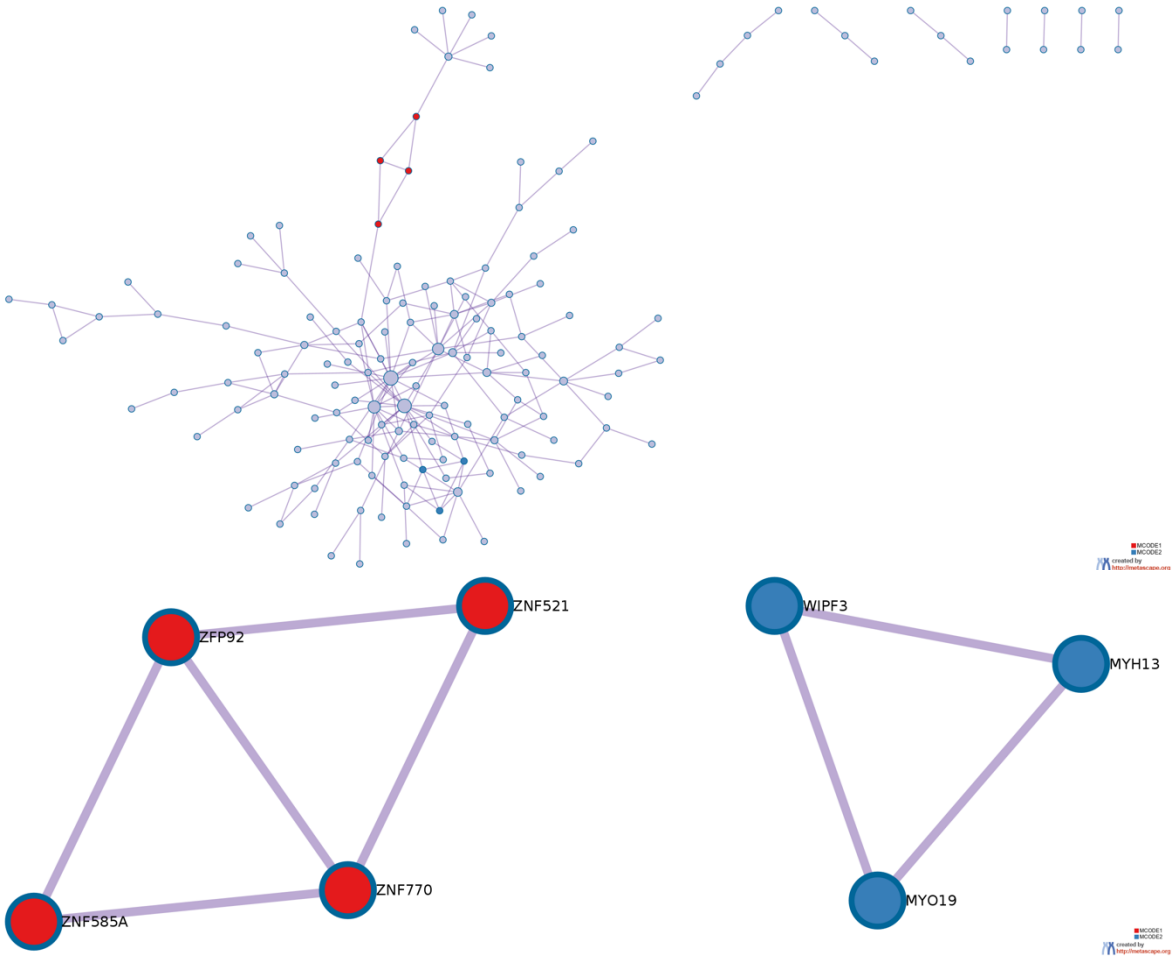


Figure 24. Protein interactions generated with Metascape tool. Detection of physical protein-protein interactions in the local recurrence group examining SNPs and INDELs combined.

5 Discussion and analysis

The present study showed the possible implementation of bioinformatic analysis starting with variant call data with the aim of possibly discovering novel prognostic lung cancer recurrence markers in the absence of normal tissue non-tumor controls, as for the FFPE samples those simply are not available. The main focus was put on comparing distant versus local datasets to examine potential differences between the groups. It has to be considered, that analyzed samples originate from primary tumors which have later on produced local or metastatic recurrence. Many earlier discovered biomarkers are thought to be the main contributors to underlying disease and targeted therapies are used, yet no single suitable target is eligible in all cases. The most known lung cancer marker genes are EGFR, ALK, KRAS, ROS1, HER2, RET, MET, BRAF, PIK3CA, NTRK1, FGFR, DDR2 (Villalobos and Wistuba 2017). Regarding the prediction of recurrence and its extent, multiple circulating proteins, circulating nucleic acids, and circulating tumor microemboli have been reported as promising markers (Crosbie et al. 2013). Nevertheless, the research on the topic continues with attempts to find the best and most trustworthy targets or their combinations. Investigating gene lists acquired as a result of data analysis, the invasive group's genes' relation to cancer or metastasis was explored from previous publications. Additionally, the genes' function was taken into account. The gene DMXL2 from the aggressive relapse group exhibited mutations produced by SNPs and INDELs both. In addition, it appeared in the start of the table with a P-value of 0.002 in the SNPs list and 0.017 in the INDEL list, showing its statistical significance. In breast cancer, DMXL2 is reported as an initiator of epithelial to mesenchymal progression, where cells acquire migratory and invasive characteristics (Faronato et al. 2015). Association between DMXL2 and lung cancer has not been yet discovered nor published to the author's knowledge. From the current study, DMXL2 shows as a promising novel marker potentially driving metastasis in lung cancer patients. Another intriguing unique gene from the metastatic group SNP list was ABCC9, with a P-value of 0.002. Examination of prior research on the gene showed its potential to be a diagnostic and prognostic marker in the triple-negative breast cancer (X. Zhang et al. 2020). Furthermore, breast cancer progression is hindered and tumor resistance to doxorubicin is reduced by inhibition of ABCC9 (Li et al. 2022). Conversely, comparing ABCC9 expression in lung adenocarcinomas and nontumorous tissues, ABCC9 was found to be highly expressed in the last one (L. Zhang et al. 2021). These findings suggest, that more studies have to be employed on its function in lung cancer progression and recurrence. Subsequently, gene enrichment analysis detected a greater extent of over-representation in the metastatic recurrence group in proportion to the local recurrence group, which indicates more randomness in the latter. In other words, genes mutated in the aggressive relapse group exhibit more systematic patterns. Furthermore, enriched terms were often related to mitosis, microtubules, binding, cell projection and motility, and kinetochores. These findings are compelling, as they occurred in distant recurrence group exclusively. In addition, all these terms can serve in one way or another as a basis for cells gaining migratory attributes. In many cancer types, microtubule instability has been reported, giving cancer cells survival advantages. Moreover, it is associated with a poor prognosis (Parker, Kavallaris, and McCarroll 2014). Kinetochores are large protein complexes at the edge of the centromere connecting DNA and microtubules present during mitosis. Disturbance of kinetochore-microtubule associations is found to initiate chromosome instability and cancer evolution (Herman et al. 2015). Importantly, the most significantly enriched term was detected investigating SNPs and INDELs

conjointly with an adjusted P-value of 7.855×10^{-9} (g:Profiler). The term was GO Cellular Component 'kinetochore microtubule' with affected genes KNTC1, CLASP1, CLASP2, and CENPE. In the local group, the enrichments' adjusted P-values did not reach up to this extent. Notably, the same four genes of interest were detected by Metascape analysis forming physical protein-protein interactions. Current data further suggest their possible role in microtubule disturbance and as a result, chromosome instability and cancer progression. Additionally, several Ca^{2+} related terms were enriched both in metastatic and local recurrence groups. Interestingly, detailed investigation of involved genes revealed actions of opposite directions. For example, ATP2A1 gene from local relapse group encodes SERCA1 enzyme with a function of transporting calcium from cytosol to sarcoplasmic reticulum (SR) or endoplasmic reticulum (ER) ("ATP2A1 Gene: MedlinePlus Genetics" 2020, 1). ITPR3 from the metastatic recurrence group however transports calcium out of the ER into the cytosol and RYR3 releases calcium from SR to the cytoplasm ("RYR3" n.d., 3; "Reactome: ITPR3 Transports Ca2+ from the Endoplasmic Reticulum to the Cytosol" n.d.). Hence, a preliminary hypothesis for further studies is that calcium influx/efflux may be altered differently in local and metastatic recurrence. Such assumption is supported by previous evidence in associations between cancer and calcium regulation. For example, deregulation of calcium signaling is related to cancer hallmarks and aberration of Ca^{2+} transporter protein expression is related to some cancer types. However, there is still poor understanding about the essence of changed calcium signaling (Stewart, Yapa, and Monteith 2015). In the case of breast cancer, both increases and decreases in the cellular Ca^{2+} level indicate the malignant potential of the cell and exhibits prognostic significance (O'Grady and Morgan 2021). Furthermore, ITPR3, also known as Type 3 Inositol 1,4,5-Triphosphate Receptor, is been shown to have an anti-apoptotic and proliferative role in tumor cells (Rezuchova et al. 2019) and facilitate tumor growth and metastasis in urinary bladder carcinoma (M. Zhang et al. 2021). From a more broad view, calcium as a cell signal molecule affects cell motility, division and apoptosis, as well as cancer progression (Monteith, Prevarskaya, and Roberts-Thomson 2017). In the metastatic recurrence group, STIM1 was also found over-represented in GO Cellular Compartment 'sarcoplasmic reticulum' along with ITPR3 and RYR3. The STIM1 localizes in ER and acts as a Ca^{2+} sensor, detecting depletion of calcium ions inside ER, and subsequently interacting with ORAI1 for influx of calcium ions into the cell ("STIM1 Gene: MedlinePlus Genetics" 2020, 1). It has been reported already in 2009, that STIM1 along with its interaction partner ORAI1 are in critical importance of breast cancer cell migration and metastasis (S. Yang, Zhang, and Huang 2009, 1). Multiple participants in calcium transport within a cell are depicted in Figure 25, with mutated genes detected from current analysis SERCA (coded by ATP2A1), ITPR3, RYR3, and STIM1 all localizing in the SR/ER. Presumption is, that mutations in those proteins are not random and presents a highly intriguing hypothesis of lung cancer acquiring invasive behavior as a result of accumulating aberrations in calcium transporters. In the scope of this research, preliminary genetic alterations have been discovered. However, narrowing down and finding the most promising markers has to be conducted in the future. The initial results of the present study need validation and further investigation to make solid conclusions.

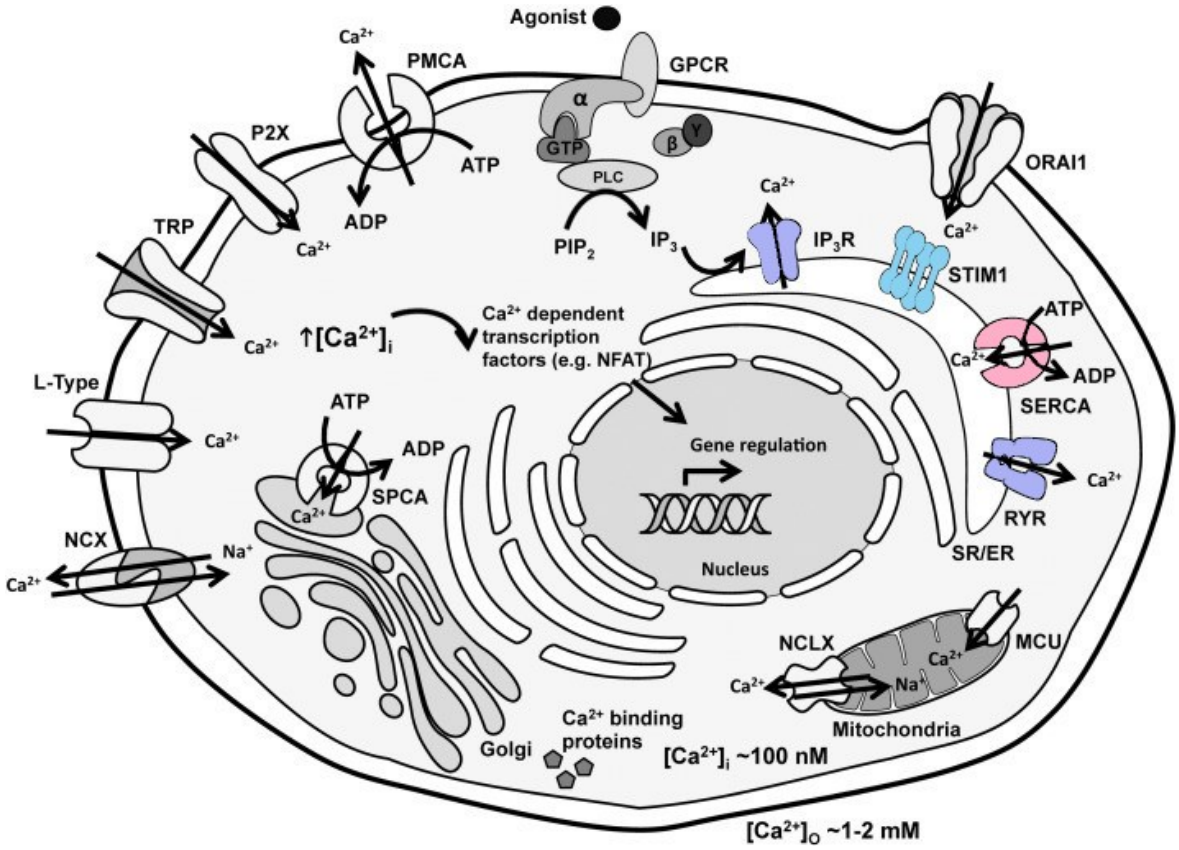


Figure 25. Ca^{2+} pathways within a cell. The activation of GPCR in plasma membrane activates IP3 pathway, where IP3 activates ITPR3 and calcium is released from ER/SR. STIM1 monitors calcium levels in the cell, and in the case of depletion, activates influx of calcium ions from extracellular environment. Similarly to ITPR3, RYR3 transports Ca^{2+} out of the SR/ER into cytosol. Conversly, SERCA is a Ca^{2+} dependent ATPase transporting calcium ions into ER/SR. Purple and blue colored proteins were mutated in the metastatic recurrence group, pink colored protein was found mutated in local recurrence group. Adapted from (Stewart, Yapa, and Monteith 2015).

6 Conclusions

The amount of detected SNPs was noticeably higher than the number of INDELs differing more than over six times (2 362 746 vs 376 180). Application of population databases with the aim to filter out benign variants removed a vast majority of detected SNPs and INDELs, 83% and 43%, respectively. Differences of INDEL numbers between local and metastatic groups were statistically significant, with more INDELs occurring in the latter. Therefore, higher initial INDEL load may be characteristic of tumors that produce metastasizing recurrence. Annotation resulted in 27 different effects produced by INDELs, with intronic, frameshift, and non-frameshift substitutions being the most common ones. Annotation of SNPs lead to the capture of 24 distinct effects, of which intronic, exonic non-synonymous, and exonic unknown appeared most frequently. In the aggressive recurrence group one gene, DMXL2, exhibited SNPs and INDELs both leading to abnormal translation of the gene. In the local recurrence group, four genes contained both small variations. However, their P-value was only borderline significant or they did not belong to the local group exclusively. Significantly more enrichments were present in the metastatic recurrence group compared to the local recurrence group. This effect occurred persistently regardless of investigating SNPs and INDELs separately or conjointly, although the combination of variants yielded the most magnified effects. Conferring two groups' enriched terms, more migration-related and invasiveness-promoting classes arose in the aggressive recurrence group. In addition, lower P-values in the distant relapse group reinforce the higher significance of detected results, with the gap reaching up to a million (10^{-9} vs 10^{-3}) when looking at all variants in conjunction. Some genes, such as CLASP1, CLASP2, KNTC1, and CENPE were showing up in different terms more abundantly than others implicating their role in various functions and processes. The same genes were also over-represented in 'kinetochore microtubule' exhibiting the smallest adjusted P-value over all terms (7.855×10^{-9}). Furthermore, physical protein-protein interactions were detected between the four genes by Metascape, which further reinforces their possible collective role in microtubule disturbance. Additionally, both local and metastatic recurrence group exhibited mutations in Ca^{2+}-transport related terms. Interestingly, the involved genes ATP2A1 in the local and ITPR3 and RYR3 in the aggressive recurrence group, appear to be associated in actions of opposite directions of calcium influx/efflux. Possibly, those findings could indicate different calcium movement in the case of local and invasive recurrence. In conclusion, overall terms being characteristic of an invasive recurrence group solely were associated with cell projection, binding, mitosis, microtubules, and cells gaining migratory properties. In this research, possible genes facilitating recurrent tumor progression and metastasis have been captured, yet the confirmation of the results remains to be investigated with further studies. Hopefully, these findings contribute to ongoing efforts regarding developing novel prognostic biomarkers, capable of predicting patients outcome.

Abstract

Lung cancer is the top one cause for cancer mortality and the second most commonly diagnosed tumor. Lung cancer is characterized by substantial recurrence rate, with most of them falling into metastatic category, leading to poor outcomes. Although multiple lung cancer driver oncogenes have been reported, the precise underlying mechanisms of tumor aggressive relapse remain still unknown. Whole-exome sequencing provides the possibility to decipher cancer-inherent genomic alterations stemming from the coding part of the genome with a fairly small time and manageable cost. The aim of current thesis was to depict possible implementation of bioinformatic data analysis in order to detect novel clinically relevant small mutations, such as SNPs and INDELs contributing to lung cancer recurrence. The main emphasis was put on comparing local and metastatic recurrence patients variant data which originated from primary tumors. The hypothesis was, that there are genetic aberrations with potential to predict aggressive lung cancer re-growth and aid the prognosis of patients. As a result, much more enrichments occurred in the metastatic recurrence group compared to local recurrence group. The enrichments in the aggressive recurrence group were more systematic, being present in terms mostly related to mitosis, microtubules, Ca^{2+} transport, as well as cell projection and motility. As cancer cell migration is the foundation of metastasis, multiple genes detected by present analysis could possibly affect tumor switching to more invasive form. In the future, confirmation of the results is needed to make conclusive inferences. Idea of using control group of patients lacking lung cancer recurrence after primary tumor treatment has been started to ensure the reliability of current results.

Kokkuvõte

Kopsuvähi uute diagnooside arv igal aastal kuulub tabeli tippude hulka, kusjuures suremus antud vähitüüpi on kõrgeim võrreldes teiste kasvajatega. Kopsuvähki iseloomustab kõrge retsidiivide hulk ning enamikel juhtudel on tegu agressiivse progressiooniga, mis on levinud siiretega üle keha erinevatesse organitesse. Selle tulemusena ei ole patsientide prognoos soodne ning suremus haigusesse on kõrge. Kasvaja eksoomi sekveneerimine võimaldab detekteerida mutatsioone võrdlemisi väikese aja- ja rahakuludega. Käesoleva magistritöö eesmärgiks oli uute väikeste mutatsioonide avastamine, mis panevad aluse kopsuvähi taas-tekkele. Põhirõhk antud töös asetati lokaalse ja invasiivse kopsuvähi retsidiivi uurimisele, kusjuures proovid pärinesid primaarse kasvajaga patsientidelt. Hüpoteesiks oli, et aset leiavad geneetilised muudatused, mis on võimelised ennustama retsidiivi ulatuse ja raskusastme teket ning mille abil saaks prognoosida haiguse kulgu. Vaadates tulemusi, olid mitmed bioloogiliste funktsioonid, protsessid ja raku osad ulatuslikult rohkem üle-esindatud metastaatilise kuluga retsidiivi grupis. Enim rikastunud olid Ca^{2+} transpordi, mikrotuubulite, mitoosi, kui ka raku liikumise ja migratsiooni seotud klassid. Kasvaja siirete tekkimise aluseks on rakkude liikumine ja migratsioon, seetõttu mitmed käesolevast analüüsist detekteeritud geneetilised mutatsioonid võivad potentsiaalselt mõjutada haiguse lülitumist agressiivsesse vormi. Tulevikus on vajalikud lisauuringud antud leidude kinnitamiseks, mis võimaldaks teha põhjapanevaid järeldusi. Plaanis on kontroll-grupi kaasamine, kus patsientidel esialgse vähi ravimise järgselt retsidiivi ei tekkinud, võimaldades tõsta käesoleva uurimustöö usaldusväärsust.

Acknowledgements

The present thesis would have not been completed without the guidance and contribution of many people. I acknowledge my supervisor, Olli-Pekka Smolander for his inspiring encouragement, mentoring, and willingness to explain complicate topics. Many appreciations to the NEMC oncologists Kersti Oselin, Ann Valter, and Anu Planken for pleasant collaboration and support. Additionally, I would like to thank Intermountain Precision Genomics (USA) for their stake of sequencing and initial data analysis. And finally, I will be always grateful for the support of my family.

References

Alanazi, Abdullah, Ismaeel Yunusa, Khaled Elenizi, and Abdulaziz I Alzarea. 2021. "Efficacy and Safety of Tyrosine Kinase Inhibitors in Advanced Non-Small-Cell Lung Cancer Harboring Epidermal Growth Factor Receptor Mutation: A Network Meta-Analysis." *Lung Cancer Management* 10 (1): LMT43. https://doi.org/10.2217/lmt-2020-0011.

Alexandrov, Ludmil B, and Michael R Stratton. 2014. "Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes." *Current Opinion in Genetics & Development* 24 (February): 52–60. https://doi.org/10.1016/j.gde.2013.11.014.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Astolfi, Annalisa, Milena Urbini, Valentina Indio, Margherita Nannini, Chiara Giusy Genovese, Donatella Santini, Maristella Saponara, et al. 2015. "Whole Exome Sequencing (WES) on Formalin-Fixed, Paraffin-Embedded (FFPE) Tumor Tissue in Gastrointestinal Stromal Tumors (GIST)." *BMC Genomics* 16 (1): 892. https://doi.org/10.1186/s12864-015-1982-6.

"ATP2A1 Gene: MedlinePlus Genetics." 2020. MedlinePlus. https://medlineplus.gov/genetics/gene/atp2a1/.

BAC Resource Consortium, The, V. G. Cheung, N. Nowak, W. Jang, I. R. Kirsch, S. Zhao, X.-N. Chen, et al. 2001. "Integration of Cytogenetic Landmarks into the Draft Sequence of the Human Genome." *Nature* 409 (6822): 953–58. https://doi.org/10.1038/35057192.

Bailey, Shannon Terrell, Jim Lund, Hao Wang, Hao Chen, Hongye Sun, and Jeffery Gulcher. 2018. "High-Quality Whole-Genome Sequencing of FFPE Samples." *Journal of Clinical Oncology* 36 (15_suppl): e13500–e13500. https://doi.org/10.1200/JCO.2018.36.15_suppl.e13500.

Bao, Riyue, Lei Huang, Jorge Andrade, Wei Tan, Warren A. Kibbe, Hongmei Jiang, and Gang Feng. 2014. "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing." *Cancer Informatics* 13s2 (January): CIN.S13779. https://doi.org/10.4137/CIN.S13779.

Bartha and Győrffy. 2019. "Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology." *Cancers* 11 (11): 1725. https://doi.org/10.3390/cancers11111725.

Belzen, Ianthe A. E. M. van, Alexander Schönhuth, Patrick Kemmeren, and Jayne Y. Hehir-Kwa. 2021. "Structural Variant Detection in Cancer Genomes: Computational Challenges and Perspectives for Precision Oncology." *Npj Precision Oncology* 5 (1): 15. https://doi.org/10.1038/s41698-021-00155-6.

Blandin Knight, Sean, Phil A. Crosbie, Haval Balata, Jakub Chudziak, Tracy Hussell, and Caroline Dive. 2017. "Progress and Prospects of Early Detection in Lung Cancer." *Open Biology* 7 (9): 170070. https://doi.org/10.1098/rsob.170070.

Bodor, J. Nicholas, Vineela Kasireddy, and Hossein Borghaei. 2018. "First-Line Therapies for Metastatic Lung Adenocarcinoma Without a Driver Mutation." *Journal of Oncology Practice* 14 (9): 529–35. https://doi.org/10.1200/JOP.18.00250.

Caspar, S.M., N. Dubacher, A.M. Kopps, J. Meienberg, C. Henggeler, and G. Matyas. 2018. "Clinical Sequencing: From Raw Data to Diagnosis with Lifetime Value." *Clinical Genetics* 93 (3): 508–19. https://doi.org/10.1111/cge.13190.

Chen, Edward Y, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. 2013. "Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool." *BMC Bioinformatics* 14 (1): 128. https://doi.org/10.1186/1471-2105-14-128.

Consonni, Dario, Mariaelena Pierobon, Mitchell H. Gail, Maurizia Rubagotti, Melissa Rotunno, Alisa Goldstein, Lynn Goldin, et al. 2015. "Lung Cancer Prognosis Before and After Recurrence in a Population-Based Setting." *JNCI: Journal of the National Cancer Institute* 107 (6). https://doi.org/10.1093/jnci/djv059.

Cooper, Wendy A., David C. L. Lam, Sandra A. O'Toole, and John D. Minna. 2013. "Molecular Biology of Lung Cancer." *Journal of Thoracic Disease* 5 Suppl 5 (October): S479-490. https://doi.org/10.3978/j.issn.2072-1439.2013.08.03.

Crosbie, Philip A. J., Rajesh Shah, Yvonne Summers, Caroline Dive, and Fiona Blackhall. 2013. "Prognostic and Predictive Biomarkers in Early Stage NSCLC: CTCs and Serum/Plasma Markers." *Translational Lung Cancer Research* 2 (5): 382–97. https://doi.org/10.3978/j.issn.2218-6751.2013.09.02.

Deng, Na, Heng Zhou, Hua Fan, and Yuan Yuan. 2017. "Single Nucleotide Polymorphisms and Cancer Susceptibility." *Oncotarget* 8 (66): 110635–49. https://doi.org/10.18632/oncotarget.22372.

Dhamija, Sonam, Chul Min Yang, Jeanette Seiler, Ksenia Myacheva, Maiwen Caudron-Herger, Angela Wieland, Mahmoud Abdelkarim, et al. 2020. "A Pan-Cancer Analysis Reveals Nonstop Extension Mutations Causing SMAD4 Tumour Suppressor Degradation." *Nature Cell Biology* 22 (8): 999–1010. https://doi.org/10.1038/s41556-020-0551-7.

"Enrichr Help Center." n.d. Ma'ayan Laboratory - Computational Systems Biology - Icahn School of Medicine at Mount Sinai. Accessed April 23, 2022. https://maayanlab.cloud/Enrichr/help#background&q=6.

Exome Aggregation Consortium, Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91. https://doi.org/10.1038/nature19057.

Faronato, Monica, Van T. M. Nguyen, Darren K. Patten, Ylenia Lombardo, Jennifer H. Steel, Naina Patel, Laura Woodley, et al. 2015. "DMXL2 Drives Epithelial to Mesenchymal Transition in Hormonal Therapy Resistant Breast Cancer through Notch Hyper-Activation." *Oncotarget* 6 (26): 22467–79. https://doi.org/10.18632/oncotarget.4164.

Gao, Xian Hua, Juan Li, Hai Feng Gong, Guan Yu Yu, Peng Liu, Li Qiang Hao, Lian Jie Liu, Chen Guang Bai, and Wei Zhang. 2020. "Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer." *Frontiers in Oncology* 10 (March): 310. https://doi.org/10.3389/fonc.2020.00310.

"Gene Ontology Overview." n.d. Gene Ontology Resource. Accessed April 23, 2022. http://geneontology.org/docs/ontology-documentation/.

Global Cancer Observatory, n.d. n.d. "Cancer Today." Accessed April 23, 2022. https://gco.iarc.fr/today/home.

Goh, Gerald, and Murim Choi. 2012. "Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases." *Genomics & Informatics* 10 (4): 214. https://doi.org/10.5808/GI.2012.10.4.214.

Govindan, Ramaswamy, Li Ding, Malachi Griffith, Janakiraman Subramanian, Nathan D. Dees, Krishna L. Kanchi, Christopher A. Maher, et al. 2012. "Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers." *Cell* 150 (6): 1121–34. https://doi.org/10.1016/j.cell.2012.08.024.

Gridelli, Cesare, Antonio Rossi, and Paolo Maione. 2003. "Treatment of Non-Small-Cell Lung Cancer: State of the Art and Development of New Biologic Agents." *Oncogene* 22 (42): 6629–38. https://doi.org/10.1038/sj.onc.1206957.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Herman, Jacob A., Chad M. Toledo, James M. Olson, Jennifer G. DeLuca, and Patrick J. Paddison. 2015. "Molecular Pathways: Regulation and Targeting of Kinetochore–Microtubule Attachment in Cancer." *Clinical Cancer Research* 21 (2): 233–39. https://doi.org/10.1158/1078-0432.CCR-13-0645.

Hiddinga, Birgitta I., Jo Raskin, Annelies Janssens, Patrick Pauwels, and Jan P. Van Meerbeeck. 2021. "Recent Developments in the Treatment of Small Cell Lung Cancer." *European Respiratory Review* 30 (161): 210079. https://doi.org/10.1183/16000617.0079-2021.

Hintzsche, Jennifer D., William A. Robinson, and Aik Choon Tan. 2016. "A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data." *International Journal of Genomics* 2016: 1–16. https://doi.org/10.1155/2016/7983236.

Hirsch, Fred R, Giorgio V Scagliotti, James L Mulshine, Regina Kwon, Walter J Curran, Yi-Long Wu, and Luis Paz-Ares. 2017. "Lung Cancer: Current Therapies and New Targeted Treatments." *The Lancet* 389 (10066): 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8.

Howe, Kevin L, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, et al. 2021. “Ensembl 2021.” *Nucleic Acids Research* 49 (D1): D884–91. https://doi.org/10.1093/nar/gkaa942.

Howlader, Nadia, Gonçalo Forjaz, Meghan J. Mooradian, Rafael Meza, Chung Yin Kong, Kathleen A. Cronin, Angela B. Mariotto, Douglas R. Lowy, and Eric J. Feuer. 2020. “The Effect of Advances in Lung-Cancer Treatment on Population Mortality.” *New England Journal of Medicine* 383 (7): 640–49. https://doi.org/10.1056/NEJMoa1916623.

Institute for Systems Genomics. 2017. “File Formats Tutorial.” Computational Biology Core. University of Connecticut. https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/#.

Jalali Sefid Dashti, Mahjoubeh, and Junaid Gamieldien. 2017. “A Practical Guide to Filtering and Prioritizing Genetic Variants.” *BioTechniques* 62 (1). https://doi.org/10.2144/000114492.

Jones, Siân, Valsamo Anagnostou, Karli Lytle, Sonya Parpart-Li, Monica Nesselbush, David R. Riley, Manish Shukla, et al. 2015. “Personalized Genomic Analyses for Cancer Mutation Discovery and Interpretation.” *Science Translational Medicine* 7 (283): 283ra53-283ra53. https://doi.org/10.1126/scitranslmed.aaa7161.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Kelly M. Latimer and Timothy F. Mott. 2015. “Lung Cancer: Diagnosis, Treatment Principles, and Screening.” *American Family Physician* 91 (February). https://www.aafp.org/afp/2015/0215/afp20150215p250.pdf.

Kokkat, Theresa J., Miral S. Patel, Diane McGarvey, Virginia A. LiVolsi, and Zubair W. Baloch. 2013. “Archived Formalin-Fixed Paraffin-Embedded (FFPE) Blocks: A Valuable Underexploited Resource for Extraction of DNA, RNA, and Protein.” *Biopreservation and Biobanking* 11 (2): 101–6. https://doi.org/10.1089/bio.2012.0052.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. “Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.” *Nucleic Acids Research* 44 (W1): W90–97. https://doi.org/10.1093/nar/gkw377.

Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. “ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype.” *Nucleic Acids Research* 42 (D1): D980–85. https://doi.org/10.1093/nar/gkt1113.

Lemjabbar-Alaoui, Hassan, Omer UI Hassan, Yi-Wei Yang, and Petra Buchanan. 2015. “Lung Cancer: Biology and Treatment Options.” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1856 (2): 189–210. https://doi.org/10.1016/j.bbcan.2015.08.002.

Li, Yang, Cui Jiang, Xiaoxue Zhang, Zhixuan Liao, Long Chen, Shuang Li, Shunxiong Tang, Zhe Fan, and Qiang Zhang. 2022. “Inhibition of ABCC9 by Zinc Oxide Nanoparticles Induces Ferroptosis and Inhibits Progression, Attenuates Doxorubicin Resistance in Breast Cancer.” *Cancer Nanotechnology* 13 (1): 3. https://doi.org/10.1186/s12645-021-00109-4.

Lin, Jessica J., and Alice T. Shaw. 2016. “Resisting Resistance: Targeted Therapies in Lung Cancer.” *Trends in Cancer* 2 (7): 350–64. https://doi.org/10.1016/j.trecan.2016.05.010.

Maleki, Farhad, Katie Ovens, Daniel J. Hogan, and Anthony J. Kusalik. 2020. “Gene Set Analysis: Challenges, Opportunities, and Future Research.” *Frontiers in Genetics* 11 (June): 654. https://doi.org/10.3389/fgene.2020.00654.

Mathieson, William, and Gerry Thomas. 2019. “Using FFPE Tissue in Genomic Analyses: Advantages, Disadvantages and the Role of Biospecimen Science.” *Current Pathobiology Reports* 7 (3): 35–40. https://doi.org/10.1007/s40139-019-00194-6.

MC3 Working Group, PCAWG novel somatic mutation calling methods working group, PCAWG Consortium, Matthew H. Bailey, William U. Meyerson, Lewis Jonathan Dursi, Liang-Bo Wang, et al. 2020. “Retrospective Evaluation of Whole Exome and Genome Mutation Calls in 746 Cancer Samples.” *Nature Communications* 11 (1): 4748. https://doi.org/10.1038/s41467-020-18151-y.

McCarthy, Davis J, Peter Humburg, Alexander Kanapin, Manuel A Rivas, Kyle Gaulton, asds, Jean-Baptiste Cazier, and Peter Donnelly. 2014. “Choice of Transcripts and Software Has a Large Effect on Variant Annotation.” *Genome Medicine* 6 (3): 26. https://doi.org/10.1186/gm543.

Meléndez, Bárbara, Claude Van Campenhout, Sandrine Rorive, Myriam Remmelink, Isabelle Salmon, and Nicky D’Haene. 2018. “Methods of Measurement for Tumor Mutational Burden in Tumor Tissue.” *Translational Lung Cancer Research* 7 (5): 661–67. https://doi.org/10.21037/tlcr.2018.08.02.

Mendoza-Alvarez, Alejandro, Beatriz Guillen-Guio, Adrian Baez-Ortega, Carolina Hernandez-Perez, Sita Lakhwani-Lakhwani, Maria-del-Carmen Maeso, Jose M. Lorenzo-Salazar, Manuel Morales, and Carlos Flores. 2019. “Whole-Exome Sequencing Identifies Somatic Mutations Associated With Mortality in Metastatic Clear Cell Kidney Carcinoma.” *Frontiers in Genetics* 10 (May): 439. https://doi.org/10.3389/fgene.2019.00439.

Monteith, Gregory R., Natalia Prevarskaya, and Sarah J. Roberts-Thomson. 2017. “The Calcium–Cancer Signalling Nexus.” *Nature Reviews Cancer* 17 (6): 373–80. https://doi.org/10.1038/nrc.2017.18.

Nambiar, Mridula, Vijayalakshmi Kari, and Sathees C. Raghavan. 2008. “Chromosomal Translocations in Cancer.” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1786 (2): 139–52. https://doi.org/10.1016/j.bbcan.2008.07.005.

O’Grady, Shane, and Maria P. Morgan. 2021. “Calcium Transport and Signalling in Breast Cancer: Functional and Prognostic Significance.” *Seminars in Cancer Biology* 72 (July): 19–26. https://doi.org/10.1016/j.semcancer.2019.12.006.

on behalf of the 100,000 Genomes Project, Pauline Robbe, Niko Popitsch, Samantha J L Knight, Pavlos Antoniou, Jennifer Becq, Miao He, et al. 2018. “Clinical Whole-Genome Sequencing from Routine Formalin-Fixed, Paraffin-Embedded Specimens: Pilot Study for the 100,000 Genomes Project.” *Genetics in Medicine* 20 (10): 1196–1205. https://doi.org/10.1038/gim.2017.241.

Osada, Hirotaka, and Takashi Takahashi. 2002. “Genetic Alterations of Multiple Tumor Suppressors and Oncogenes in the Carcinogenesis and Progression of Lung Cancer.” *Oncogene* 21 (48): 7421–34. https://doi.org/10.1038/sj.onc.1205802.

Ouchi, K., S. Takahashi, K. Tatsuno, A. Hayashi, S. Yamamoto, H. Ueda, M. Inoue, H. Nakano, H. Aburatani, and C. Ishioka. 2013. “Whole-Exome Sequencing (WES) Using Formalin-Fixed Paraffin Embedded (FFPE) Tissue.” *Annals of Oncology* 24 (November): ix93. https://doi.org/10.1093/annonc/mdt460.132.

Parker, Amelia L., Maria Kavallaris, and Joshua A. McCarroll. 2014. “Microtubules and Their Role in Cellular Stress in Cancer.” *Frontiers in Oncology* 4: 153. https://doi.org/10.3389/fonc.2014.00153.

Peng, Yong, and Carlo M Croce. 2016. “The Role of MicroRNAs in Human Cancer.” *Signal Transduction and Targeted Therapy* 1 (1): 15004. https://doi.org/10.1038/sigtrans.2015.4.

Pérez-Ruiz, Elisabeth, Ignacio Melero, Joanna Kopecka, Ana Bela Sarmento-Ribeiro, Marilina García-Aranda, and Javier De Las Rivas. 2020. “Cancer Immunotherapy Resistance Based on Immune Checkpoints Inhibitors: Targets, Biomarkers, and Remedies.” *Drug Resistance Updates* 53 (December): 100718. https://doi.org/10.1016/j.drup.2020.100718.

Piraino, S.W., and S.J. Furney. 2016. “Beyond the Exome: The Role of Non-Coding Somatic Mutations in Cancer.” *Annals of Oncology* 27 (2): 240–48. https://doi.org/10.1093/annonc/mdv561.

Rabbani, Bahareh, Mustafa Tekin, and Nejat Mahdieh. 2014. “The Promise of Whole-Exome Sequencing in Medical Genetics.” *Journal of Human Genetics* 59 (1): 5–15. https://doi.org/10.1038/jhg.2013.114.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. “G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47 (W1): W191–98. https://doi.org/10.1093/nar/gkz369.

“Reactome: ITPR3 Transports Ca2+ from the Endoplasmic Reticulum to the Cytosol.” n.d. Reactome Pathway Database. Accessed April 24, 2022. https://reactome.org/content/detail/R-HSA-9717215.

Réda, Manon, Corentin Richard, Aurelie Bertaut, Julie Niogret, Thomas Collot, Jean David Fumet, Julie Blanc, et al. 2020. “Implementation and Use of Whole Exome Sequencing for Metastatic Solid Cancer.” *EBioMedicine* 51 (January): 102624. https://doi.org/10.1016/j.ebiom.2019.102624.

Reimand, Jüri, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, et al. 2019. “Pathway Enrichment Analysis and Visualization of Omics Data Using

g:Profiler, GSEA, Cytoscape and EnrichmentMap." *Nature Protocols* 14 (2): 482–517. https://doi.org/10.1038/s41596-018-0103-9.

Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. 2007. "G:Profiler—a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments." *Nucleic Acids Research* 35 (suppl_2): W193–200. https://doi.org/10.1093/nar/gkm226.

Rezuchova, Ingeborg, Sona Hudecova, Andrea Soltysova, Miroslava Matuskova, Erika Durinikova, Barbora Chovancova, Michal Zuzcak, et al. 2019. "Type 3 Inositol 1,4,5-Trisphosphate Receptor Has Antiapoptotic and Proliferative Role in Cancer Cells." *Cell Death & Disease* 10 (3): 186. https://doi.org/10.1038/s41419-019-1433-4.

Rotunno, Melissa, Rolando Barajas, Mindy Clyne, Elise Hoover, Naoko I. Simonds, Tram Kim Lam, Leah E. Mechanic, Alisa M. Goldstein, and Elizabeth M. Gillanders. 2020. "A Systematic Literature Review of Whole Exome and Genome Sequencing Population Studies of Genetic Susceptibility to Cancer." *Cancer Epidemiology Biomarkers & Prevention* 29 (8): 1519–34. https://doi.org/10.1158/1055-9965.EPI-19-1551.

"RYR3." n.d. GeneCards Is a Searchable, Integrative Database That Provides Comprehensive, User-Friendly Information on All Annotated and Predicted Human Genes. Accessed April 24, 2022. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RYR3.

Sanchez-Vega, Francisco, Marco Mina, Joshua Armenia, Walid K. Chatila, Augustin Luna, Konnor C. La, Sofia Dimitriadoy, et al. 2018. "Oncogenic Signaling Pathways in The Cancer Genome Atlas." *Cell* 173 (2): 321-337.e10. https://doi.org/10.1016/j.cell.2018.03.035.

Schrank, Zachary, Gagan Chhabra, Leo Lin, Tsatsral Iderzorig, Chike Osude, Nabiha Khan, Adijan Kuckovic, Sanjana Singh, Rachel Miller, and Neelu Puri. 2018. "Current Molecular-Targeted Therapies in NSCLC and Their Mechanism of Resistance." *Cancers* 10 (7): 224. https://doi.org/10.3390/cancers10070224.

Schuster, Samantha L., and Andrew C. Hsieh. 2019. "The Untranslated Regions of MRNAs in Cancer." *Trends in Cancer* 5 (4): 245–62. https://doi.org/10.1016/j.trecan.2019.02.011.

Seaby, Eleanor G., Reuben J. Pengelly, and Sarah Ennis. 2016. "Exome Sequencing Explained: A Practical Guide to Its Clinical Application." *Briefings in Functional Genomics* 15 (5): 374–84. https://doi.org/10.1093/bfgp/elv054.

Skoulidis, Ferdinandos, and John V. Heymach. 2019. "Co-Occurring Genomic Alterations in Non-Small-Cell Lung Cancer Biology and Therapy." *Nature Reviews Cancer* 19 (9): 495–509. https://doi.org/10.1038/s41568-019-0179-8.

Stewart, Teneale A., Kunsala T.D.S. Yapa, and Gregory R. Monteith. 2015. "Altered Calcium Signaling in Cancer Cells." *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1848 (10): 2502–11. https://doi.org/10.1016/j.bbamem.2014.08.016.

"STIM1 Gene: MedlinePlus Genetics." 2020. MedlinePlus. https://medlineplus.gov/genetics/gene/stim1/.

Subotic, Dragan, Paul Van Schil, and Bogdan Grigoriu. 2016. "Optimising Treatment for Post-Operative Lung Cancer Recurrence." *European Respiratory Journal* 47 (2): 374–78. https://doi.org/10.1183/13993003.01490-2015.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*, February, caac.21660. https://doi.org/10.3322/caac.21660.

Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. 2014. "Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers." *Cell* 156 (6): 1324–35. https://doi.org/10.1016/j.cell.2014.01.051.

Suwinski, Pawel, ChuangKee Ong, Maurice H. T. Ling, Yang Ming Poh, Asif M. Khan, and Hui San Ong. 2019. "Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big

Data Analytics." *Frontiers in Genetics* 10 (February): 49. https://doi.org/10.3389/fgene.2019.00049.

Takeshima, Hideyuki, and Toshikazu Ushijima. 2019. "Accumulation of Genetic and Epigenetic Alterations in Normal Cells and Cancer Risk." *Npj Precision Oncology* 3 (1): 7. https://doi.org/10.1038/s41698-019-0079-0.

Tate, John G, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47. https://doi.org/10.1093/nar/gky1015.

Thai, Alesha A, Benjamin J Solomon, Lecia V Sequist, Justin F Gainor, and Rebecca S Heist. 2021. "Lung Cancer." *The Lancet* 398 (10299): 535–54. https://doi.org/10.1016/S0140-6736(21)00312-3.

Tsao, Anne S., Giorgio V. Scagliotti, Paul A. Bunn, David P. Carbone, Graham W. Warren, Chunxue Bai, Harry J. de Koning, et al. 2016. "Scientific Advances in Lung Cancer 2015." *Journal of Thoracic Oncology* 11 (5): 613–38. https://doi.org/10.1016/j.jtho.2016.03.012.

Tuna, Musaffe, Christopher I. Amos, and Gordon B. Mills. 2019. "Molecular Mechanisms and Pathobiology of Oncogenic Fusion Transcripts in Epithelial Tumors." *Oncotarget* 10 (21): 2095–2111. https://doi.org/10.18632/oncotarget.26777.

Ugur Sezerman, Osman, Ege Ulgen, Nogayhan Seymen, and Ilknur Melis Durasi. 2019. "Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization." In *Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations*, edited by Ali Samadikuchaksaraei and Morteza Seifi. IntechOpen. https://doi.org/10.5772/intechopen.85524.

Vansteenkiste, J., L. Crinò, C. Dooms, J.Y. Douillard, C. Faivre-Finn, E. Lim, G. Rocco, et al. 2014. "2nd ESMO Consensus Conference on Lung Cancer: Early-Stage Non-Small-Cell Lung Cancer Consensus on Diagnosis, Treatment and Follow-Up." *Annals of Oncology* 25 (8): 1462–74. https://doi.org/10.1093/annonc/mdu089.

Villalobos, Pamela, and Ignacio I. Wistuba. 2017. "Lung Cancer Biomarkers." *Hematology/Oncology Clinics of North America* 31 (1): 13–29. https://doi.org/10.1016/j.hoc.2016.08.006.

Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58. https://doi.org/10.1126/science.1235122.

Wang, K., M. Li, and H. Hakonarson. 2010. "ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data." *Nucleic Acids Research* 38 (16): e164–e164. https://doi.org/10.1093/nar/gkq603.

"Whole Exome Sequencing for Cancer Research: IDT." n.d. Integrated DNA Technologies. Accessed April 23, 2022. https://eu.idtdna.com/pages/research-area/cancer/cancer-research/cancer-exome-sequencing.

World Health Organization. 2021. "Cancer." https://www.who.int/news-room/fact-sheets/detail/cancer.

Yang, Hui, and Kai Wang. 2015. "Genomic Variant Annotation and Prioritization with ANNOVAR and WANNOVAR." *Nature Protocols* 10 (10): 1556–66. https://doi.org/10.1038/nprot.2015.105.

Yang, Shengyu, J. Jillian Zhang, and Xin-Yun Huang. 2009. "Orai1 and STIM1 Are Critical for Breast Tumor Cell Migration and Metastasis." *Cancer Cell* 15 (2): 124–34. https://doi.org/10.1016/j.ccr.2008.12.019.

Yousefi, Meysam, Tayyeb Bahrami, Arash Salmaninejad, Rahim Nosrati, Parisa Ghaffari, and Seyed H. Ghaffari. 2017. "Lung Cancer-Associated Brain Metastasis: Molecular Mechanisms and Therapeutic Options." *Cellular Oncology* 40 (5): 419–41. https://doi.org/10.1007/s13402-017-0345-5.

Yu, Helena A., Ken Suzawa, Emmet Jordan, Ahmet Zehir, Ai Ni, Ryan Kim, Mark G. Kris, et al. 2018. "Concurrent Alterations in EGFR-Mutant Lung Cancers Associated with Resistance to EGFR Kinase Inhibitors and Characterization of MTOR as a Mediator of Resistance." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 24 (13): 3108–18. https://doi.org/10.1158/1078-0432.CCR-17-2961.

Zhang, Linbo, Ping Huang, Chunxia Huang, Lingmei Jiang, Zhijie Lu, and Peng Wang. 2021. "Varied Clinical Significance of ATP-Binding Cassette C Sub-Family Members for Lung Adenocarcinoma." *Medicine* 100 (16): e25246. https://doi.org/10.1097/MD.0000000000025246.

Zhang, Mengzhao, Lu Wang, Yangyang Yue, Lu Zhang, Tianjie Liu, Minxuan Jing, Xiao Liang, et al. 2021. "ITPR3 Facilitates Tumor Growth, Metastasis and Stemness by Inducing the NF-κB/CD44 Pathway

in Urinary Bladder Carcinoma." *Journal of Experimental & Clinical Cancer Research* 40 (1): 65. https://doi.org/10.1186/s13046-021-01866-1.

Zhang, Qi, Qihan Fu, Xueli Bai, and Tingbo Liang. 2020. "Molecular Profiling–Based Precision Medicine in Cancer: A Review of Current Evidence and Challenges." *Frontiers in Oncology* 10 (October): 532403. https://doi.org/10.3389/fonc.2020.532403.

Zhang, Xiaoyu, Xiaoning Kang, Lijun Jin, Jie Bai, Hui Zhang, Wei Liu, and Zunyi Wang. 2020. "ABCC9, NKAPL, and TMEM132C Are Potential Diagnostic and Prognostic Markers in Triple-negative Breast Cancer." *Cell Biology International* 44 (10): 2002–10. https://doi.org/10.1002/cbin.11406.

Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications* 10 (1): 1523. https://doi.org/10.1038/s41467-019-09234-6.

Appendices

Appendix 1 Denotation for some of the script contents

Appendix 1.1 Denotation of script contents for separation of SNPs from INDELs.

`while read line`	Goes through the input file line by line until the end
`< $1`	Specifies that the input is a file given when the script is run
`do-done pair`	Tells what to do each time we go through file
`$`	Allows us to put the contents of the line as a part of the command
`$line`	Allows us to use the contents of each file
`echo $line`	To print the contents of the line (filename) to the command we want to run for each file
`--gzvcf`	Defines the VCF file to be processed (compressed VCF file)
`--remove-indels`	Excludes sites that contain an INDEL
`--keep-only-indels`	Excludes sites that contain SNPs
`--out`	Defines the output filename prefix for all files generated by vcftools
`--recode`	Used to generate a new file (output file) in VCF from the input VCF file after applying the filtering options (remove INDELs)

Appendix 1.2 Denotation of the script contents of SNP data yielding ANNOVAR compatible format.

`while read line`	Goes through the input file line by line until the end
`< $1`	Specifies that the input is a file given when the script is run
`do-done pair`	Tells what to do each time we go through file
`cat $(echo $line)`	Reads data from the file and gives their content as output
`grep -v`	Matches and displays all the lines except the given pattern
`awk '{print $1,$2,$2,$4,$5}'`	Prints only 5 elements of the line to get just the SNP positions, ref and alt allele

Appendix 2 Gene lists for local and metastatic groups including SNPs and INDELs

Appendix 2.1 Gene list for metastatic group for SNPs and INDELs. Green colored genes are indicated to exhibit SNPs and INDELs both.

Gene	Effect	Gene	Effect	Gene	Effect	Gene	Effect
ABCC9	frameshift substitution	KIF21B	nonsynonymous SNV	RTTN	frameshift substitution	MANBA	frameshift substitution
DMXL2	nonsynonymous SNV	NEFM	frameshift substitution	SSH2	nonsynonymous SNV	MSS51	nonsynonymous SNV
EIF5B	frameshift substitution	NR4A2	nonsynonymous SNV	TAS2R10	frameshift substitution	MTSS1	nonsynonymous SNV
KNTC1	frameshift substitution	PBRM1	nonsynonymous SNV	TMEM5	frameshift substitution	MYO6	nonsynonymous SNV
C2CD3	nonsynonymous SNV	STIM1	nonsynonymous SNV	YLPM1	stopgain	MYO7A	nonsynonymous SNV
CLASP1	frameshift substitution	TTC3	nonsynonymous SNV	MAP4	nonsynonymous SNV	PCDHGA9	nonsynonymous SNV
USP54	nonsynonymous SNV	ZNF646	nonsynonymous SNV	UBR5	nonsynonymous SNV	PDZRN4	nonsynonymous SNV
APOB	nonsynonymous SNV	EPB41L2	nonsynonymous SNV	UNC13B	nonsynonymous SNV	PRRT4	nonsynonymous SNV
ITPR3	nonsynonymous SNV	LYST	frameshift substitution	SACS	nonsynonymous SNV	SLCO5A1	nonsynonymous SNV
GNAS	nonsynonymous SNV	NPC1L1	nonsynonymous SNV	ZNF469	nonsynonymous SNV	SPPL2C	nonsynonymous SNV
PRR12	nonsynonymous SNV	TNRC6A	nonsynonymous SNV	MTOR	nonsynonymous SNV	SPRYD7	frameshift substitution
MGA	frameshift substitution	ADAMTSL3	stopgain	RIF1	frameshift substitution	SRGAP1	nonsynonymous SNV
RYR3	nonsynonymous SNV	BNC1	frameshift substitution	WNK1	nonsynonymous SNV	SVEP1	stopgain
SEZ6	nonsynonymous SNV	CABYR	nonsynonymous SNV	BRD8	nonsynonymous SNV	TMEM168	nonsynonymous SNV
FER1L6	nonsynonymous SNV	CHD7	nonsynonymous SNV	BZW1	frameshift substitution	ZNF114	stopgain
MAP1A	frameshift substitution	DIAPH1	nonsynonymous SNV	CAPN1	nonsynonymous SNV	ZNF184	nonsynonymous SNV
NUP160	nonsynonymous SNV	ESPN	nonsynonymous SNV	CEP97	frameshift substitution	ARHGEF28	nonsynonymous SNV
SBNO2	nonsynonymous SNV	FNBP4	nonsynonymous SNV	COL15A1	nonsynonymous SNV	ATM	nonsynonymous SNV
SWT1	frameshift substitution	FNDC3B	frameshift substitution	CORIN	frameshift substitution	CAMSAP3	nonsynonymous SNV
CREB1	stopgain	GOLGA6C	nonsynonymous SNV	DENND2A	nonsynonymous SNV	DNAH9	frameshift substitution
CENPE	frameshift substitution	IL18RAP	frameshift substitution	EGFLAM	frameshift substitution	KIAA1549	frameshift substitution
CLASP2	nonsynonymous SNV	NR3C1	frameshift substitution	EML6	stopgain	NEXN	frameshift substitution
CSMD2	frameshift substitution	PELP1	nonsynonymous SNV	EXOC8	nonsynonymous SNV	CCDC168	nonsynonymous SNV
DOCK7	nonsynonymous SNV	PRUNE2	nonsynonymous SNV	FYCO1	nonsynonymous SNV		
IL4R	nonsynonymous SNV	PTPRH	nonsynonymous SNV	KIAA0586	frameshift substitution		

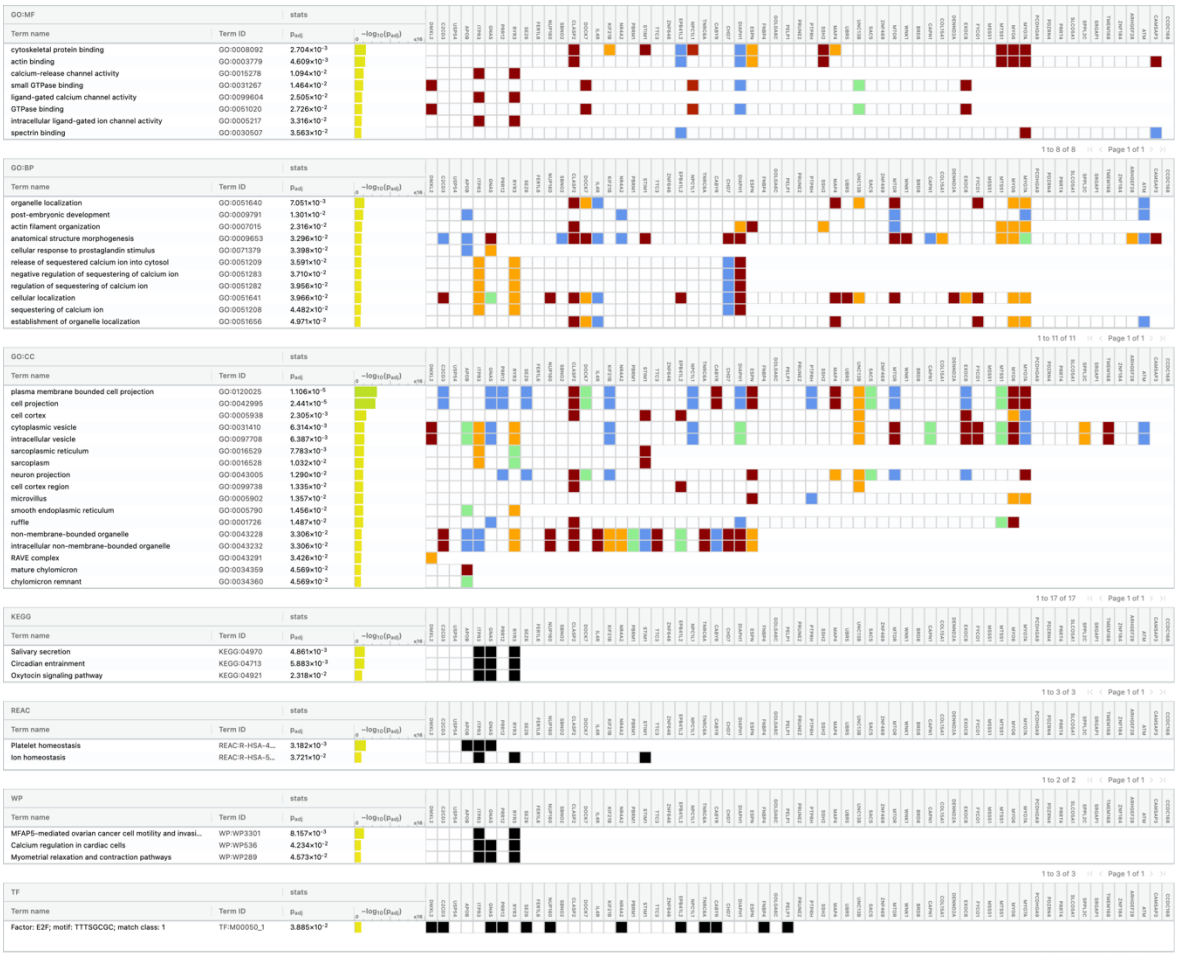
Appendix 2.2 Gene list for local group for SNPs and INDELs. Green colored genes are indicated to exhibit SNPs and INDELs both.

Gene	Element	Gene	Element	Gene	Element	Gene	Element
MPDZ	nonsynonymous SNV	PAQR7	nonsynonymous SNV	ZNF257	stopgain	RAB27A	frameshift substitution
WASF3	nonsynonymous SNV	TCHH	nonsynonymous SNV	CACNA1G	nonsynonymous SNV	RNASE10	frameshift substitution
CCDC181	nonsynonymous SNV	TBX6	nonsynonymous SNV	SPTBN2	nonsynonymous SNV	SEC23A	stopgain
ARSH	nonsynonymous SNV	WNK2	nonsynonymous SNV	APOBR	nonsynonymous SNV	SLC37A3	frameshift substitution
DCLRE1C	nonsynonymous SNV	DNAH5	nonsynonymous SNV	INTS5	nonsynonymous SNV	SLC9A8	nonsynonymous SNV
FGFR1	nonsynonymous SNV	ARID2	nonsynonymous SNV	UBTF	nonsynonymous SNV	STAU1	nonsynonymous SNV
MECOM	frameshift substitution	ABL1	frameshift substitution	WASHC2C	nonsynonymous SNV	STX1A	nonsynonymous SNV
SELP	frameshift substitution	AMDHD2	nonsynonymous SNV	AMZ1	nonsynonymous SNV	TECPR2	stopgain
TRIM16	nonsynonymous SNV	ANO8	nonsynonymous SNV	ARMC12	nonsynonymous SNV	TRIM2	frameshift substitution
BDH1	nonsynonymous SNV	ARHGEF25	nonsynonymous SNV	CCDC113	nonsynonymous SNV	UBXN11	nonsynonymous SNV
C8orf34	nonsynonymous SNV	ATP10A	nonsynonymous SNV	CCM2	nonsynonymous SNV	UGT1A3	nonsynonymous SNV
ERMAP	nonsynonymous SNV	C19orf25	nonsynonymous SNV	CLOCK	frameshift substitution	USP49	nonsynonymous SNV
FAHD1	nonsynonymous SNV	CACNB2	frameshift substitution	EHD1	nonsynonymous SNV	CSMD1	nonsynonymous SNV
FUT3	nonsynonymous SNV	GDF6	nonsynonymous SNV	FAM102A	nonsynonymous SNV	CCDC116	nonsynonymous SNV
GAB3	frameshift substitution	LRRC4B	nonsynonymous SNV	FOXD4L6	nonsynonymous SNV	KIFAP3	nonsynonymous SNV
IRAK4	nonsynonymous SNV	N4BP2L1	nonsynonymous SNV	HNRNPA3	nonsynonymous SNV	OSBPL5	nonsynonymous SNV
JPH2	nonsynonymous SNV	PIK3CG	nonsynonymous SNV	HNRNPU	stopgain	EPB41L1	nonsynonymous SNV
OR4X1	frameshift substitution	STT3B	nonsynonymous SNV	HSD17B4	nonsynonymous SNV	ERN2	nonsynonymous SNV
PALM2	nonsynonymous SNV	WSB1	nonsynonymous SNV	IL17REL	nonsynonymous SNV	MYH13	nonsynonymous SNV
THRA	nonsynonymous SNV	C11orf95	nonsynonymous SNV	LRP5L	nonsynonymous SNV	RASA1	frameshift substitution
UCHL1	nonsynonymous SNV	TRMT1	nonsynonymous SNV	LRRC36	nonsynonymous SNV	WIPF3	nonsynonymous SNV
ZFAND3	stopgain	ATP2A1	nonsynonymous SNV	NME7	frameshift substitution	ZC3H13	frameshift substitution
ZSCAN32	nonsynonymous SNV	CAPN12	nonsynonymous SNV	NTNG1	nonsynonymous SNV	ZNF517	nonsynonymous SNV
ZYG11A	nonsynonymous SNV	RIMS1	stopgain	OR9I1	nonsynonymous SNV	ABCA4	nonsynonymous SNV
MTMR10	frameshift substitution	RUNDC3A	nonsynonymous SNV	POTEE	nonsynonymous SNV	ARMC2	frameshift substitution

Gene	Element	Gene	Element	Gene	Element	Gene	Element
BARHL1	nonsynonymous SNV	ASPA	nonsynonymous SNV	DDR1	stopgain	KLHL4	stopgain
DENND4A	stopgain	ATP5F1	frameshift substitution	DECR1	frameshift substitution	KNG1	nonsynonymous SNV
DSTYK	nonsynonymous SNV	ATP6AP1L	frameshift substitution	DENND1C	frameshift substitution	KPRP	frameshift substitution
GORASP1	nonsynonymous SNV	AVIL	stopgain	DENND2A	frameshift substitution	KRT77	stopgain
MTERF3	nonsynonymous SNV	AXDND1	nonsynonymous SNV	DHRS12	nonsynonymous SNV	KRTAP21-3	nonsynonymous SNV
POU4F2	nonsynonymous SNV	B3GALT4	stopgain	DNAJC12	stopgain	KRTDAP	nonsynonymous SNV
RASAL2	stopgain	BCAP29	stopgain	DPY30	nonsynonymous SNV	LAP3	frameshift substitution
RUNX2	nonsynonymous SNV	C17orf97	nonsynonymous SNV	EPC2	nonsynonymous SNV	LEF1	nonsynonymous SNV
SASH1	nonsynonymous SNV	CBWD7	nonsynonymous SNV	EXOC5	nonsynonymous SNV	LEKR1	nonsynonymous SNV
SLC22A1	nonsynonymous SNV	CCDC7	stopgain	FANCD2OS	frameshift substitution	LRRC23	frameshift substitution
SLC38A7	nonsynonymous SNV	CCND3	nonsynonymous SNV	FCGR1A	frameshift substitution	MCL1	frameshift substitution
SRL	nonsynonymous SNV	CD151	nonsynonymous SNV	FLG2	stopgain	MLEC	nonsynonymous SNV
TBX5	nonsynonymous SNV	CDH24	frameshift substitution	FSD1L	nonsynonymous SNV	MLST8	nonsynonymous SNV
TESMIN	nonsynonymous SNV	CDKL4	frameshift substitution	GALNTL6	nonsynonymous SNV	MORN1	nonsynonymous SNV
VSTM2A	nonsynonymous SNV	CFHR2	stopgain	GGACT	nonsynonymous SNV	MPC1L	nonsynonymous SNV
PPP1R26	nonsynonymous SNV	CHN2	stopgain	GPC3	nonsynonymous SNV	MRGPRX4	nonsynonymous SNV
ZNF827	stopgain	CLEC4G	nonsynonymous SNV	GPR179	stopgain	MSS51	stopgain
KIAA1671	nonsynonymous SNV	CLN3	frameshift substitution	GPR34	frameshift substitution	MTIF3	frameshift substitution
FAM65A	nonsynonymous SNV	COG4	stopgain	HBS1L	stopgain	MYO19	frameshift substitution
ACSL1	stopgain	COMMD7	frameshift substitution	HCLS1	nonsynonymous SNV	NAT16	stopgain
ACTR2	nonsynonymous SNV	CRAMP1	nonsynonymous SNV	HERC5	frameshift substitution	NBPF26	nonsynonymous SNV
AMPD1	frameshift substitution	CRYGA	nonsynonymous SNV	HSPD1	stopgain	NECAB1	nonsynonymous SNV
ARHGEF38	frameshift substitution	CST3	nonsynonymous SNV	IRF6	stopgain	NEK2	nonsynonymous SNV
ARMT1	frameshift substitution	CYP2B6	nonsynonymous SNV	ITGA2B	stopgain	NKX2-1	frameshift substitution
ASH2L	nonsynonymous SNV	DCUN1D4	nonsynonymous SNV	KCNAB1	stopgain	NPVF	stopgain

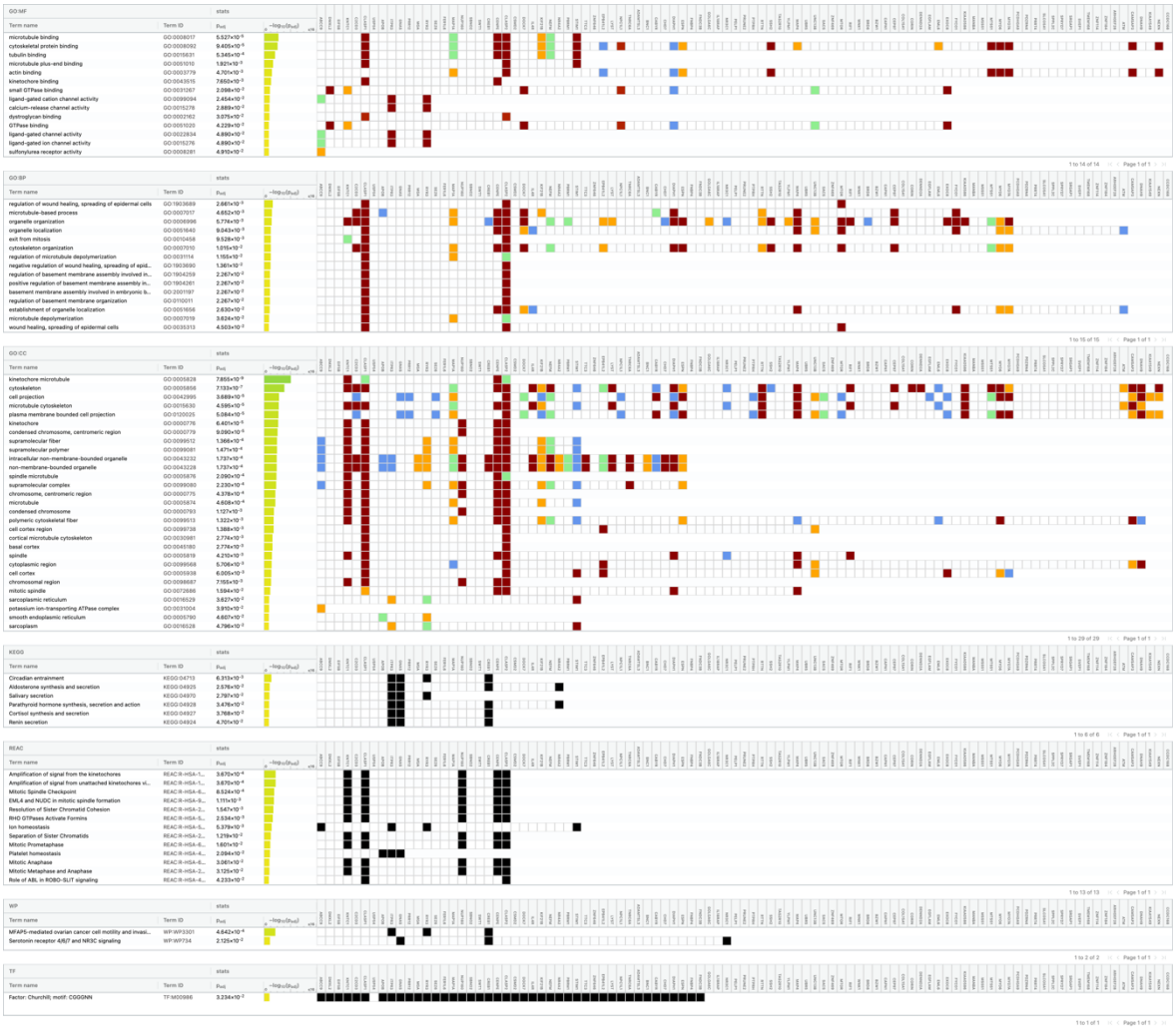
Gene	Element	Gene	Element	Gene	Element
NUP35	frameshift substitution	RFPL1	nonsynonymous SNV	YWHAZ	nonsynonymous SNV
OAZ3	nonsynonymous SNV	RIMBP3	frameshift substitution	ZBED9	stopgain
OLFM3	stopgain	SERPINB3	frameshift substitution	ZBTB34	stopgain
OR1E2	nonsynonymous SNV	SERTAD2	frameshift substitution	ZDHHC23	frameshift substitution
OR2D3	nonsynonymous SNV	SFTPC	nonsynonymous SNV	ZFP92	nonsynonymous SNV
OR2M7	nonsynonymous SNV	SH3KBP1	stopgain	ZIM3	nonsynonymous SNV
OR6C70	nonsynonymous SNV	SKAP1	nonsynonymous SNV	ZMAT3	stopgain
OXCT2	nonsynonymous SNV	SLC6A6	frameshift substitution	ZNF23	stopgain
PAX2	stopgain	SMAGP	stopgain	ZNF311	frameshift substitution
PDPR	nonsynonymous SNV	SNU13	stopgain	ZNF418	frameshift substitution
PKMYT1	frameshift substitution	SP3	stopgain	ZNF521	stopgain
PLAT	nonsynonymous SNV	ST3GAL3	nonsynonymous SNV	ZNF585A	stopgain
PNLDC1	nonsynonymous SNV	STARD10	nonsynonymous SNV	ZNF770	stopgain
PPA2	nonsynonymous SNV	TERB1	stopgain	SCAF4	nonsynonymous SNV
PPIL2	nonsynonymous SNV	TEX29	frameshift substitution		
PPP2R3B	nonsynonymous SNV	TFAP2D	stopgain		
PRIMPOL	nonsynonymous SNV	TIMP3	nonsynonymous SNV		
PROKR2	startloss	TK1	nonsynonymous SNV		
PROP1	nonsynonymous SNV	TKFC	nonsynonymous SNV		
PTPRQ	stopgain	TM4SF18	stopgain		
PYGO2	frameshift substitution	TMEM209	stopgain		
RAB43	nonsynonymous SNV	TMEM45A	nonsynonymous SNV		
RAD51AP1	frameshift substitution	TNFSF8	frameshift substitution		
RBM38	stopgain	TRIM62	nonsynonymous SNV		
RBM46	stopgain	TRMT12	stopgain		
RCBTB1	frameshift substitution	USP16	frameshift substitution		

Appendix 3 Gene enrichment analysis results generated by g:Profiler



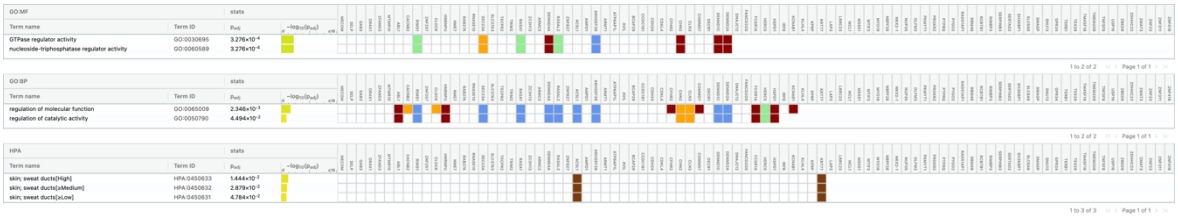
Appendix 3.1. Gene enrichment analysis results generated by g:Profiler for the aggressive group including only SNPs.

Appendix 3.2. Gene enrichment analysis results generated by g:Profiler for the aggressive group including only INDELs.



Appendix 3.3. Gene enrichment analysis results generated by g:Profiler for the metastatic group including SNPs and INDELs.

Appendix 3.4. Gene enrichment analysis results generated by g:Profiler for the local group including only SNPs.



Appendix 3.5. Gene enrichment analysis results generated by g:Profiler for the local group including only INDELs.

Appendix 3.6. Gene enrichment analysis results generated by g:Profiler for the local group including SNPs and INDELs.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

1.1.1

Mina Laura Luhari (sünnikuupäev: 04.08.1993)

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose Bioinformatic analysis of recurrent lung cancer whole-exome sequencing data for detection of clinically relevant mutations, mille juhendajad on Olli-Pekka Smolander ja Kersti Oselin

1.1 reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2 üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.

3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

_____ (allkiri)

_____ (kuupäev)