Master Degree Project

UNIVERSITY OF SKÖVDE

ANALYSIS OF SOMATIC MUTATIONS IN PAPILLOMAVIRUS POSITIVE TUMOURS FROM YOUNGER AND OLDER OROPHARYNGEAL CANCER PATIENTS

Master Degree Project in Bioinformatics
One year Level 30 ECTS
Spring term 2016

Andreas Ährlund-Richter

Supervisors: Dan Lundh,
Cinzia Bersani and Sebastian DiLorenzo
Examiner: Zelmina Lubovac

ABSTRACT

Background: Human Papilloma Virus positive (HPV^+) Oropharyngeal Squamous Cell Carcinoma (OSCC), dominated by tonsillar cancer (TSCC) and base of tongue cancer (BOTSCC) has low mutation-frequency and better survival for younger than older patients.

Aim: To examine if HPV^+ TSCC and BOTSCC have distinct gene-mutation profiles, for 50 often-mutated genes in cancer, in younger compared to older patients and to test and compare different variant callers to get a deeper understanding of the data.

Materials and methods: DNA had previously been extracted from 299 formalin-fixed-paraffin-embedded (FFPE) tumor biopsies and 13 normal samples, and sequenced on the Ion Proton sequencer, a NGS (Next-Generation Sequencing) platform. Alignment and variant calling had been performed via the Ion Torrent Suite software v5 (ITS), and Torrent Variant Caller (TVC).

UPPMAX, a High-Performance-Computational cluster (HPC) at Uppsala University was used for storing and computing of the sequenced data. Parallel-processing was used to optimize repetitive steps, saving days of computation time. The descriptive analysis, graphical data representations, and more in-depth analysis, were done in R.

Initially, variant calling was performed for 13 tumor/normal paired samples using the novel MuTect2 software from the Genome Analysis Toolkit (GATK) toolset. Variant annotation and statistical analysis was performed on all the 13 paired sequenced samples, using SnpEff. Due to a poor overlap between the above MuTect2 and TVC, after adequate filtering TVC, MuTect, Strelka and VarScan2 were also utilized for comparison.

Result and Conclusion: Having only 13 normal samples, normal-tumor paired variant calling to distinguish germ line and somatic variants could not be performed. To obtain an approximation of the amount of germline variation in our cohort, additional variant callers were used for the tumor normal pairs. The data obtained with the TVC caller formed the basis for further analysis of the tumor samples.

Notably, comparisons of TVC with MuTect2 output revealed major discrepancies and limited overlap. However, when comparing MuTect2 with MuTect, Strelka, or VarScan 2 regarding overlaps with TVC – the overlap were still limited, but higher degrees of overlaps were disclosed between Strelka, MuTect and MuTec2, indicating that MuTect2 was a successful further development and successor of MuTect.

Evidence for presence of distinct gene-mutation profiles correlating to age could not be obtained in the analyzed tumor cohort with the software tool kits applied. Statistical analysis using Wilcoxon-Mann Whitney test did not support a hypothesis of distinct age related mutations for any of the 50 genes analyzed in this tumor cohorts.

The highest p-value for the Wilcoxon-Mann Whitney test was for the gene APC, at $p \sim 0.054$ hints at a possible connection. However, more extensive research with more samples sequenced is necessary to confirm or reject this correlation.

ABBREVIATIONS

AF	Allele Frequency
APC	Adenomatous Polyposis Coli
BAM	Binary Alignment Map format
BOTSCC	Base Of Tongue Squamous Cell Carcinoma
BWA	Burrows-Whealer-Algorithm
CNA	Copy Number Alterations
CHP2	Cancer Hotspot Panel 2
COSMIC	Catalogue of Somatic Mutation in Cancer
CRAN	Comprehensive R Archive Network
DNA	Deoxyribonucleic acid
dNTPs	Deoxy nucleoside triphosphates
FFPE	Formalin-Fixed Paraffin-Embedded
GATK	Genome Analysis Tool Kit
HNSCC	Head neck squamous cell carcinoma
HPC	High Performance Computation
HPV^{+}	Human Papilloma Virus positive
HPV^{-}	Human Papilloma Virus negative
INDEL	Insertion or deletion of bases in the DNA of an organism
IVG	Integrative Genomics Viewer
MNP	Multiple Nucleotide Polymorphism
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
RNA	Ribonucleic acid
SLURM	Simple Linux Utility for Resource Management
SNP	Single Nucleotide Polymorphism
TMAP	Torrent Mapping Alignment Program
TSCC	Tonsillar Squamous Cell Carcinoma
TVC	Torrent Variant Caller
VCF	Variant Call Format

TABLE OF CONTENTS

1.INTRODUCTION

1.1 HPV and cancer

Human papillomavirus (HPV) infections are mainly asymptomatic, but some types cause cancer. HPV is responsible for 70% of oropharyngeal cancers, particularly tonsillar (TSCC) and base of tongue cancer (BOTSCC) in Sweden [1].

HPV is, besides smoking and alcohol, a risk factor for TSCC and BOTSCC and lately the incidence of HPV-positive (HPV^+) TSCC and BOTSCC has increased while smoking induced HPV-negative cancer (HPV^-) has decreased (Fig. 1) [2]. Notably, younger HPV^+ TSCC/BOTSCC patients have a better disease specific survival (DSS) than older patients after therapy (Fig. 2)[2].

Next Generation Sequencing (NGS) is revolutionizing many areas in clinical practice, including the identification and screening of sequencing-based biomarkers, a tool increasingly relevant for diagnosis of multiple diseases and for the selection and monitoring of therapeutic treatments.

A clinical contribution of the proposed project is to improve the prognosis for HPV^+ TSCC/BOTSCC patients. We are applying NGS technique to identify mutations or clusters of mutations in 50 selected genes important for cancer development in an initial cohort of 299 HPV^+ TSCC/BOTSCC patient tumors. The aim is to identify differences in the mutational profile in these patients and to relate that to their age by statistical correlation and clustering methods.

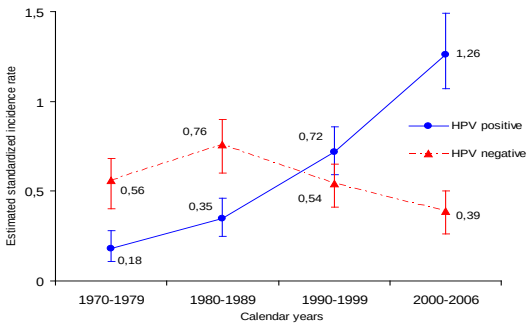


Fig. 1: Estimated age standardized incidence rate with 95% CI of HPV^+ TSCC and HPV^- TSCC cases per 100,000 person years in the County of Stockholm between 1970 and 2006 [1].

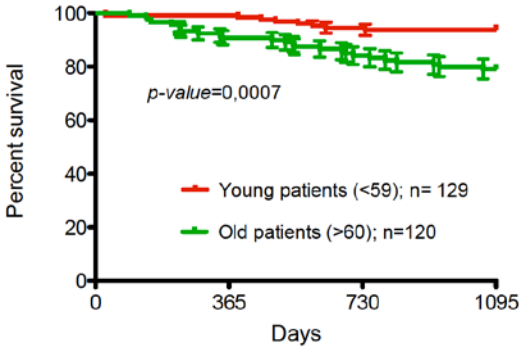


Fig. 2: Age related survival of HPV^+ TSCC and HPV^+ BOTSCC (unpublished data and [2]).

1.2 Previous studies profiling mutational status of HPV+ TSCC/BOTSCC in relation to age of cancer onset

A very limited number of studies have been published where researchers have focused on examining the genetic features of HPV+ TSCC/BOTSCC and relate them to age and response to therapy. Braakhuis and colleagues [3] analyzed mutation in the tumor suppressor p53 in HPV+ oral squamous cell carcinoma in patients younger than 45 years [3]. They showed that p53 mutations are rare in young patients with biologically active HPV. Friedland et al [4] profiled 19 HPV+ TSCC looking particularly at mutations in epidermal growth factor receptor (EGFR), Kirsten RNA Associated Rat Sarcoma 2 Virus (KRAS) and B-Raf oncogene serine/threonine protein kinase (BRAF) genes [4]. They found that gene mutations were rare for these set of genes and that there was a statistically significant improved survival of younger age and non-smokers patients. In another study by Farnebo and co-authors [5], it was shown that the presence of a set of single nucleotide polymorphisms (SNPs) in genes involved in DNA repair mechanism (i.e. XPC, XPD, XRCC1, and XRCC3) represents a risk factor and a survival indicator in head and neck squamous cell carcinomas patients. They showed, for example, that SNP in XRCC3 241M was significantly correlated to age of cancer onset in both men and women and they speculate that this may affect both risk for and survival of HNSCC patients [5].

1.3 Novelty of the proposed research plan

The extent to which specific genetic mutation or clusters of mutations together with age of cancer onset could play a role in survival and response to treatment for HPV+ head and neck squamous cell carcinoma (HNSCC) has not been significantly addressed previously.

A total of 207 amplicons in 50 commonly altered cancer genes were sequenced, allowing the detection of 2,800 COSMIC mutations. For the first time age dependent dynamics to obtain groups of mutated genes for this cancer type were explored.

As for further uniqueness of this study, the cohorts of 299 HPV+ TSCC/BOTSCC patients from Karolinska Institute is by far the largest cohort for this specific cancer subtype considered for targeted sequencing. Moreover, extensive clinical patient data, in terms of survival, age of onset, treatment type, response to treatment, and biomarker expression, are available for all these patients [2]. This will allow deep and detailed correlation studies with the variants found in the analyzed genes.

Another element of novelty in this study considers the use of novel beta version of MuTect2 variant caller software, which is a further development of MuTect, with at the moment no published studies available. MuTect2 was tested here for calling and filtering variants in a set of 13 paired tumor/normal samples included in our dataset and compared to those obtained by the in-house used variant caller, Torrent Variant Caller (TVC). Furthermore, subsequently, also variant callers MuTect, Strelka and VarScan2 were used for comparison to MuTect2.

1.4 Aims

The aims of this project were:

- To analyze and validate DNA sequencing data from 299 HPV^+ oropharyngeal tumors and investigate if there were age related gene mutation patterns, as well as if mutated oncogenes occur in clusters, differing between tumors in young and old patients.

- To study and verify the performance of the novel MuTect2 software, currently available as beta version, for variant calling and variant filtration of 13 paired tumor/normal sequenced samples from the patients above.

- To use MuTect, Strelka and VarScan2 for comparison to TVC and MuTect2 as well as to each other.

Important questions to be addressed

- How is Mutect2 performing in comparison to TVC, MuTect, Strelka and VarScan2?

- Do the results of the mutational analysis confirm data published by others?

- Are there genetic divergences between HPV^+ tumors from young and old patients?

2. MATERIALS AND METHODS

2.1 Tumor material and DNA sequencing

2.1.1 Sample cohort

FFPE tumor biopsies were obtained from 299 patients diagnosed 2000-2011 with HPV+ TSCC and BOTSCC. The DNA extracted from the serum (normal sample) of 13 of the 299 patients was used for germline/somatic mutation discrimination.

2.1.2 DNA extraction

DNA was extracted from the samples using the Roche High Pure RNA Isolation kit (Roche Diagnostics) according to the instruction by the manufacturer, excluding a DNAse treatment, and measured using the Qubit® fluorescence detector (Thermo Fisher) [6].

2.1.3 Library preparation

For most next generation sequencing (NGS) applications, an initial DNA amplification step is required. In this project, the AmpliSeq Cancer Hotspot Panel V2 (CHP2 - Thermo Fisher) was chosen for amplification and library preparation. CHP2 requires only 10 ng of starting DNA material, which is optimal when working with fragmented DNA extracted from FFPE samples. The CHP2 includes a pool of 207 primer pairs that will allow amplifications of specific regions in 50 genes commonly mutated in cancer, covering 2,800 mutations listed in COSMIC (Catalogue Of Somatic Mutations in Cancer), an online database of somatically acquired mutations found in human cancer [7]. See Appendix 7.1 for details.

2.1.4 Template preparation and loading of the chip

During template preparation, one library molecule (amplicon) is bound to one Ion Sphere Particle (ISP) and clonally amplified so that the surface of the bead is coated with many copies of one library molecule only (one amplicon per bead only). This is achieved through emulsion PCR (ePCR), a method that maximizes access of cycle product to the primers on the bead, as compared to free-flowing PCR amplification, thus increasing speed [8]. See Appendix 7.2 for details.

2.1.5 Sequencing

Samples have been sequenced on an Ion Torrent Proton sequencer, which implements basic semiconductor technology on its chip and on its base calling sensor [9, 10]. Instead of alternative sequencing systems that couples luminescent adapters to each nucleotide base (A, T, G, C) with different wavelengths, using lights to detect them, the Ion Torrent Proton sequencer measures pH-changes when each base is added to the DNA polymerase reaction. The average coverage depth per nucleotide in the sequenced files was above 500x for 95% of the samples. See Appendix 7.3 for details.

2.1.6 Sequence alignment and variant calling

The Torrent Ion Suite v5.0 is the primary software used to process the raw base calls that were acquired by the Ion Proton sequencer. The base calls output is in Base Alignment Map (BAM) and FASTQ file formats. The Torrent Mapping Alignment Program (TMAP) [11] is the default read-mapping algorithm for Torrent Ion Suite v5. TMAP utilizes flow-space (pH-signals from the wells) for alignment and includes different alignment algorithms that are based on the Burrows-Wheeler transform for mapping [11]. Downstream of the mapping step, variant discovery is performed by the Torrent Variant Caller (TVC) [12] plug-in built into Torrent Ion Suite v5. TVC produces a VCF file for each of the sequenced samples. The VCF format is tab separated text file having the following columns: 1. Chromosome name; 2. Position; 3. Variant's ID; 4. Reference genome; 5. Alternative (i.e. variant); 6. Quality score; 7. Filter (whether or not the variant passed quality filters); 8. INFO (generic information about this variant).

2.2 Variant validation

For 13 patients in our cohort, we had both tumors and matching normal (serum) samples (germline control). The Torrent Variant Caller plug-in (TVC) did not use these samples as pairs, but rather, variants were called out from each sample independently (i.e. no paired normalization). As an internal validation and examination of the quality of the variants called by TVC, we performed a tumour/serum paired variant calling using an alternative variant caller, namely MuTect2 [13]. This was later followed by the use of an additional three variant callers MuTect, Strelka and VarScan 2 [13-16].

2.2.1 MuTect2

MuTect2 is a method developed at the Broad Institute for the reliable and accurate identification of somatic point mutations in next generation sequencing data of cancer genomes [13]. It has components of its predecessor MuTect (see below), with included assembly-based genotyping, INDEL variant-calling and tolerance for different ploidy.

MuTect2 utilizes Bayesian classification to detect somatic mutations with very low allele fractions, requiring only few supporting reads, followed by carefully tuned filters that ensure high specificity. It operates in four steps. First, low quality sequenced data is removed. Second, variant detection via Bayesian classification using normal samples is performed. Third, a filtering step assures removal of false positives variants caused by sequencing artifacts. Lastly, a classification of variants as somatic or germ line by a second Bayesian classifier is performed. The MuTect2 version used in this thesis, an implementation of the most popular MuTect2 software is still in beta stage, and no publicly available studies showing how it is performing are available yet [13].

2.2.2. MuTect

MuTect, is the predecessor to MuTect2 and a method that applies a Bayesian classifier to detect somatic mutations with very low allele fractions [14]. It requires few supporting reads, with the addition of cautiously tuned filters to obtain high specificity. In comparison to other variant calls, i.e. Somatic Sniper, Joint SNVMix and Strelka, MuTect is regarded to have higher sensitivity with corresponding specificity, with regard to mutations with allelic fractions being as low as 0.1 and under [14]. It can, similar to all the other variant callers in this study, identify small nucleotide polymorphisms (SNP), but in contrast to the others not identify nucleotide insertions or deletions (INDELs) into the DNA.

A major difference between MuTect and MuTect2 is the inclusion of the HaplotypeCaller from GATK in MuTect2 for details see [13]. HaplotypeCaller uses a number of different algorithms to identify SNPs (like MuTect) but also identifies also INDELs. A benefit of adding HaplotypeCaller functionality to a MuTect base, is the addition of being able to handle different allele-frequencies, an attribute in HaplotypeCaller.

2.2.3. Strelka

Strelka, is useful for somatic SNP and small INDEL detection from sequencing data of matched tumor–normal samples [15]. It uses a Bayesian method for continuous allele frequencies for tumor samples and normal samples, leveraging the proposed genotype structure of the normal. The normal sample is represented as a mixture of germline variation with noise, and the tumor sample as a mixture of the normal sample with somatic variation. This allows for that sensitivity can be kept even when the tumor is impure.

2.2.4. VarScan2

VarScan2, is a useful method for uncovering somatic mutations and copy number alterations (CNAs) in exome data from tumor–normal pairs [16]. Its algorithm reads data from tumor and normal samples at the same time. A heuristic and statistical algorithm detects sequence variants classifying them by germ line, somatic, or loss of heterozygosity (LOH). Comparing normalized read depth defines relative copy number changes. It has a high sensitivity and specificity for exome sequence data.

2.2.5 SelectVariants

SelectVariants [17] is a variant evaluation and manipulation tool. It can be used for selecting subsets of variants from a larger callset to facilitate the analyses (e.g. comparing and contrasting cases vs. controls; extracting variant or non-variant loci that meet certain requirements, displaying just a few samples in a browser like Integrative Genomics Viewer (IGV) etc.). SelectVariants was used for selecting variants from the TVC file above a specific allele frequency and for all variant callers sorted on a 2% allele frequency of variants.

2.2.6 VCFcompare

VCFcompare is one of the tools provided at the VCFtools suite [18]. It compares positions in two or more VCF files and outputs the numbers of positions contained in one but not the other files; two but not the other files, and so on, which comes in handy when generating Venn diagrams, or overlap tables. The script also computes numbers such as non-reference discordance rates (including multi-allelic sites) and compares actual sequence (useful when comparing INDELs).

2.2.7. Filtering of the variant callers when analyzed together

TVC. Filtration for TVC was performed using the standard filtering for the program as described previously (Section 2.1.6), and where the initial filtering had been performed at 2% minor allele frequency.

MuTect2. For MuTect2 most variants (>95%) were rejected by with the label "alt_allele_in_normal". The "alt allele" filter sorts out any variant present in any frequency in both normal and tumor. However, this does not account for loss of homozygosity, and the fact that deep sequencing tends to increase of artifacts and errors, due to chance, very low amount of a variation can be found at low levels in the normal sample. To compensate, this particular filter was ignored, and a custom filter was applied. The custom filter was written in R, and removed any variants with less than 3 times the prevalence in tumor compared to normal. This effectively accounted for differences in coverage by ensuring that for any variant allele the amount of that variant found in normal had to be significant to be removed. Normal variant alleles found below this frequency were probable artifacts.

MuTect. MuTect had a similar issue with the deep-sequence data as MuTect2, and was filtered in the same manner as MuTect2.

Strelka. Using the default settings, no variants were obtained. Strelka uses an aligner-specific filter corresponding to different alignment algorithms. However, the Smith-Waterman algorithm used by Torrent Mapper was not among the available algorithms, so the filter was excluded. No allele-frequency filters needed to be applied, as Strelka already includes a filter allowing for some normal-contamination and allele-frequency noise.

VarScan2. Default settings were used, adequate amounts of variants detected. VarScan2. samtools mpileup was used to generate a mpileup of tumor and normal data for each sample which was piped into VarScan2, using default parameters, for variant calling. The variant callset produced did not need to be filtered to remove non-plausible variants using the defaults. See the *somatic* tool [19] for parameters.

2.3 Variant annotation

Variant annotation is the process of assigning functional information to DNA variants [20]. There are many different types of information that could be associated with variants, from measures of sequence conservation to predictions about the effect of a variant on protein structure and function [20]. Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, a variant annotation software can identify whether variants can cause protein coding change and what amino acids are affected, as well as identify whether a variant is reported as a known SNP and at what allele frequency (referring to e.g. 1000 Genome Project [21] or Exome Aggregation Consortium [22]). For this project we chose the SnpEff variant annotation software [23].

2.3.1 SnpEff and SnpSift

SnpEff [23] is a variant annotation and effect prediction tool, and it can annotate up to 1,000,000 variants per minute. SnpEff accepts input files in the Variant Call Format (VCF) and can annotate SNPs, multiple nucleotide polymorphisms (MNPs), insertions and deletions. Among the variant effect that can be calculated by SnpEff, there are “Synonymous Coding”, “Non Synonymous Coding”, “Frame Shift” and “Stop Gained”. Additionally, SnpEff provides a simple assessment of the putative impact of the variant (e.g. high, moderate or low impact). VCF is SnpEff’s default input and output format. The output VCF file annotated using SnpEff contains all the additional information in the “Info” column of the original input VCF file. Moreover, the program performs some statistics. However, the statistics of this program was not used. All variants were later analyzed with the same methods in R.

SnpSift [24] is a toolbox that allows you to filter and manipulate SnpEff annotated files. Once the genomic variants in the sample have been annotated, an additional filtering step is performed in order to find the interesting and relevant variants. Given the large data files, this is not a trivial task (e.g. you cannot load all the variants into XLS spreadsheet). SnpSift helps to perform this VCF file manipulation and filtering required at this stage in data processing pipelines. For further details of the filtering see section 3.1.1.

Here, we run SnpEff on our 299 tumor samples using GRCh37 as human reference genome [25]. The SnpEff TVC output files were then used by SnpSift to filter variants according to impact category “moderate” and “high”. The impact category “low” with mutations not leading to change of protein sequence were thus not included here. This way, only variants with putative deleterious effects (deletion, insertion/deletion frame shift, donor splice site disruptions, acceptor splice site disruptions, stop codon gains, stop codon losses, start losses) affecting protein sequence were left in the final TVC file.

2.4 Hardware and software

2.4.1 UPPMAX

UPPMAX is a HPC cluster and a large scale storage intended as the hardware for this project [26]. UPPMAX allows typical bioinformatics pipeline steps after sequencing, like mapping and annotation for large datasets, with the advantage of reduced calculation time. Several databases are stored on UPPMAX, e.g. Uniprot, Swissprot, and reference genomes. The UPPMAX hardware can support demanding algorithms and programming languages that are usually laborious on processing power (e.g. verbose R packages and cross validations).

2.4.2 Bash

Many bioinformatics pipelines allow for parallel processing and for data files piping into the modules. Different modules require specific formatting of the data and the UPPMAX bash terminal provides built-in functionality for this purpose.

Bash, the Bourne-Again Shell [27], refers both to a particular Unix shell program and its associated scripting language. It is the default shell of the GNU Operating System (Linux) and Apple's OS X. It is a powerful shell, and features command line editing, command history, command substitution, functions and arrays, to mention some.

Bash is the terminal language used for commands to UPPMAX, via SLURM [28] (an open source workload manager designed for Linux clusters). Bash automatizes running of programs allowing parallel computing on UPPMAX nodes with efficiency and in a short amount of time. This is particularly advantageous in our case, where we demand automated processing and handling of large libraries of files and data.

2.4.3 R language

R is a programming language and environment constructed mainly for statistical computing and plots [29]. R's strength includes the possibility and simplicity for the user to create graphical plots, call statistical functions, and write new functions and packages. Through the comprehensive R archive network, a multitude of scientific packages (more than 8000) are readily accessible for download and use.

2.4.4 Venny

Various Venn diagram programs are freely available to assist in the facile visual interpretation of biological datasets. Venny [30] is a simple Venn diagram browser program that compares a list of strings, and outputs shared items in a Venn diagram and has been used in this project for comparison of list of variants from different tumor samples.

2.5 Statistical methods for data analysis

2.5.1 Mann-Whitney-Wilcoxon tests

The Mann-Whitney-Wilcoxon test is similar to the t-test in formula, but like many other non-parametric tests it uses rank-sum instead of mean to calculate statistical significance, making it more robust when used on skewed variable distributions. Here, we use Wilcoxon rank-sum tests to determine significant differences in mutation rate for the different genes across five age groups [31].

2.5.2 Hierarchical clustering and heat map

For a simple visualization of the dataset and identification of possible mutational clusters, we used dendrograms and heat maps based on hierarchical clustering [32, 33].

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree represents a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The cluster heat map is a rectangular tiling of a data matrix with dendrograms appended to its margins, which compacts large amounts of information into a small space to bring out coherent patterns in the data.

A multitude of different hierarchical clustering algorithms are available in R, where one is described by Ward, J. H. [33]. In the same study, clusters of genes derived from single- and average-linkage hierarchical clustering tend to produce worse-than-random results. With this in mind, complete-linkage will be used as hierarchical clustering algorithm [27]. Particularly, we will use the hclust function in R, which uses the complete linkage method for hierarchical clustering by default. This clustering method defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. The process is repeated until the whole data set is agglomerated into one single cluster.

3. IMPLEMENTATION AND RESULTS

3.1 Variant validation – *part one*

In this project, 50 genes important in cancer [34] in a total of 299 HPV^+ TSCC/BOTSCC patient tumors biopsies described in more detail previously [1, 2] have been sequenced. These tumor biopsies were not microdissected before DNA extraction and library preparation, meaning that a variable percentage of the extracted DNA might be originated from other cell types than tumor cells (e.g.: stroma cells, pericytes and tumor infiltrating immune cells). The DNA contained in non-tumor cells represents the genetic background of the patients and variants found in this DNA are considered as germline mutations (germline variants) or individual single nucleotide polymorphisms (SNPs). These genetic variations are generally silent and differ from population to population. However, when these variants occur in a coding region of a gene and cause critical changes in the coded amino acid and in the translated protein, it can be associated with a disease.

Only DNA isolated from normal samples (in this case serum containing non-tumor cells) of 13 of the 299 patients was available for sequencing. The TVC program called variants from the 13 normal samples and the corresponding 13 tumor samples separately, without performing comparison and germline correction/extraction from the tumor samples. For this reason, aligned BAM files were used for these 13 patients (both 13 normal and 13 tumor samples BAM files) to perform variant calling and filtration using the novel MuTect2 software. The output file corresponds to one TVC file for each patient, containing exclusively mutations present in the DNA of patient tumor cells.

3.1.1 Variant calling using MuTect2

The default MuTect2 parameters were applied to identify potentially discordant variants between tumor and normal DNA (Fig. 3). The following files were provided as input files:

- A “dbSNP.vcf” file containing a list of mutations frequently found i.e. based on their population frequency was used. This list selected mutations most likely to be germline mutations and was used to reject germline mutations.
- A “COSMIC.vcf” file containing a list of mutations found only in cancer tissue was used for identification of somatic mutations (whitelist).
- A reference FASTA file (Homo_sapiens_assembly19.fasta) required as reference genome of the human genome version 19 (hg19).
- 26 BAM files corresponding to each of the 13 tumors and 13 normal sample paired patients’ sequenced samples.

Implementations (by custom R script) were necessary in order to generate "dbSNP.vcf" and "COSMIC.vcf" files that were compatible with the FASTA and BAM files originated from the TVC software (e.g. consistency in the reference ordering and usage of the correct human genome assembly (i.e. GRCh37)).

Also, MuTect2 has the capability to process each sample file in a separate processor core. A script in Bash was generated for running MuTect2 on UPPMAX for each tumor/normal pair separately, thus allowing to shorten down the time required for variant calling to a mean of 35 min instead of around 9 hours (Fig. 4).

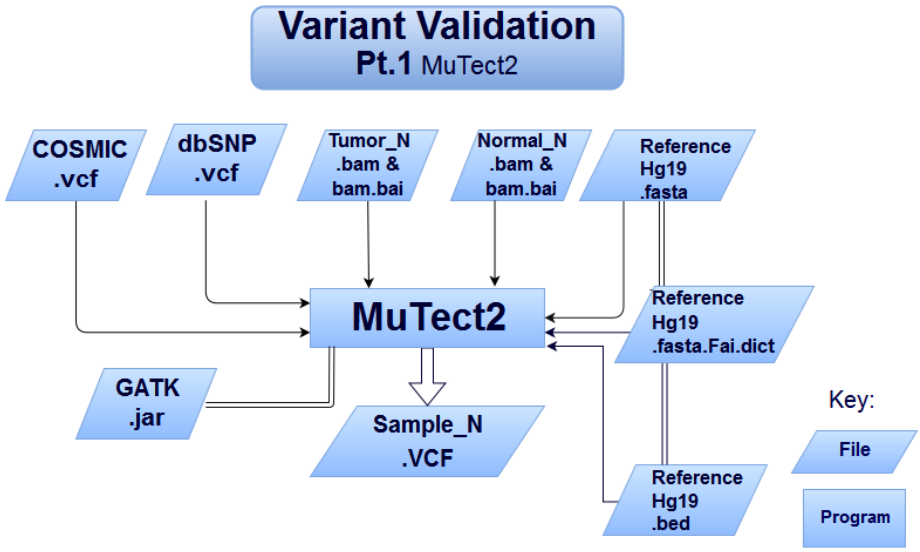


Fig. 3: MuTect2 variant calling. Flowchart describing a MuTect2 run. "Tumor N", and "Normal_N"; ".bam & .bai", are the base reads and corresponding indexes for each tumor pair. "N" can stand for any of the 13 tumor-serum pairs (see 3.1.1 for details). "dbSNP.vcf" is a list of common single nucleotide polymorphisms, and it is used as a blacklist to exclude germline mutations. "COSMIC.vcf" is a list of somatic mutation in cancer and it is used as a whitelist for identification of somatic mutations. "Sample_N.vcf" file contains the list of cancer related somatic mutation found in uniquely in the tumor sample and not in the normal sample.

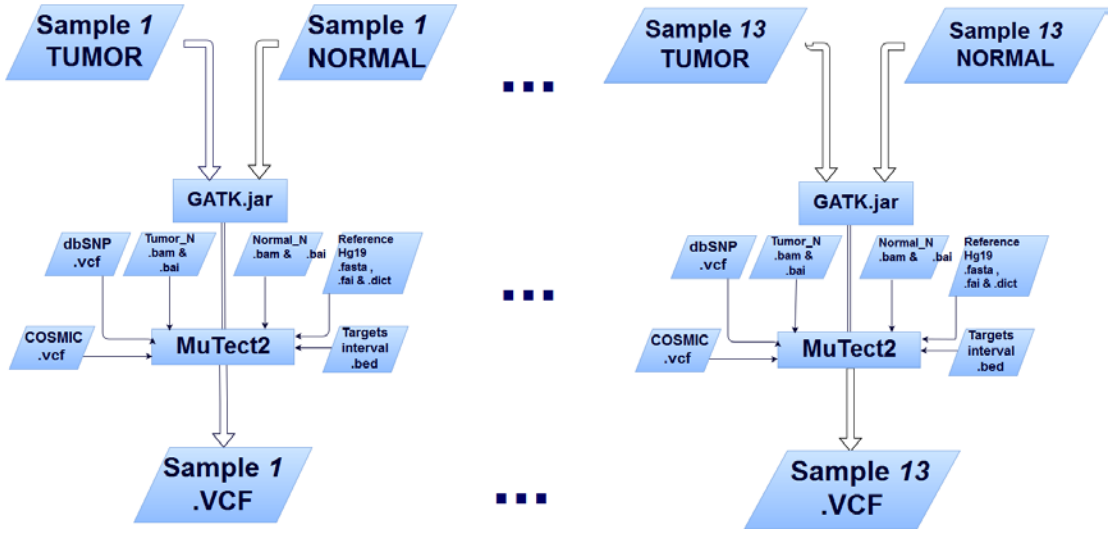


Fig. 4: Parallel computing with Sbatch scripts: Example of parallel processing with multiple Sbatch commands to the SLURM cluster. Each comparison was run on a separate processor core, minimizing any bottle necking or queuing of jobs. The "**...**" represents sample-pair 2-12.

3.1.2 Variant Filtering

Once the 13 Mutect2 output VCF files containing uniquely somatic variants present in the DNA of the tumor of the patients were obtained, the issue was to compare them with the corresponding 13 VCF files from tumor samples generated by the TVC plug-in via the Ion Torrent Suite software v5. The aim of this step was to verify the performance of the MuTect2 software for variant calling of tumor/normal sequenced samples and examine if all the variants left were also present in the tumor TVC VCF file from the corresponding patient.

While comparing MuTect2 and TVC output VCF files, it was disclosed that the allele frequency (AF) threshold for variant calling was set at 2% for the TVC software (meaning the all variants found at a frequency lower that 2% were not called). This frequency of 2% was decided when this project was initiated. However, the threshold was set at 3% for the MuTect2 software. In order to standardize the files, the GATK SelectVariants tool [17] was applied for cleaning out from the TVC VCF files all variants with AF lower than 3%. This way the two VCF files could be compared.

Another discrepancy found while comparing the variant from MuTect2 and TVC output VCF files was that TVC output VCF files contain information for each aligned position. This means that a record underneath the headers indicating reference allele ("REF") and alternative allele(s) observed ("ALT") is always given, no matter if the alternative allele corresponds to the reference allele. Differently, the MuTect2 output VCF files show only nucleotides that have been called and that correspond to single or multiple variants, insertions and deletions. An R script was therefore created for removal of records corresponding to chromosomal position where the alternative allele was identical to the reference allele in the TVC output VCF files. After this implementation, it was possible to perform the variant comparison

3.1.3 Variant comparison

In order to compare MuTect2 output VCF files containing uniquely somatic variants present in the DNA of the tumor of the patients to the corresponding VCF files from tumor samples generated by the TVC plug-in via the Ion Torrent Suite software v5, the function "VCFcompare" from the VCFtools suite was used [18].

VCFtools is a suite of functions for use on genetic variation data in the form of VCF and BCF (binary variant call format) files. The VCFtools provide functions used mainly to summarize data, run calculations on data, filter out data, and convert data into alternative file formats. The function VCFcompare compares positions in two or more VCF files and outputs the numbers of positions contained in one but not the other files, which comes handy when generating Venn diagrams.

In our analysis, the output selected was Venn diagram numbers, genotype info, as well as debug output files (Fig. 5). The same combination of Bash scripts and SLURM scripts were used for parallelization while running VCFcompare. VCFcompare works on compressed VCF files, increasing computational speed. However, VCF files needed to be first compressed and then indexed, and this was resolved by using tabix and bgzip programs respectively. Also, tabix and bgzip programs could not pipe their output files (TBI file and GZIP file) to VCFcompare directly, but it had to be done in separate steps. Finally, the Venny 2.1, a Venn's diagrams drawing tool was used for visualization of intersections and complements between compared VCF files.

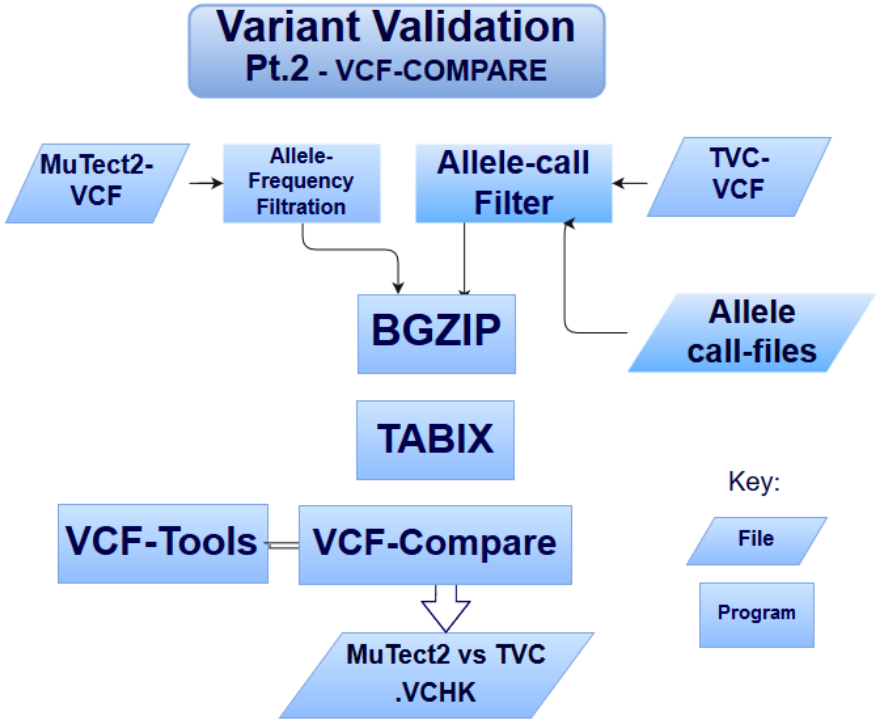


Fig. 5: Variant Comparison. Modifications in relation to the original planning: A quality filter removing poor quality calls from the TVC VCF files was added. Another change was the addition of Filtration the MuTect2 VCF files from variants with allele calls lower than the ones used in TVC filtration. Following these steps, the procedure was the same for both VCF files. VCF files of interest were compressed into binary format by BGZIP into GZIP ".gz" files. The GZIP files were then used to generate the index files in the TBI format via TABIX. The variants inside the TBI files were compared by VCF Compare, and outputted into .VCHK files displaying similar/dissimilar/missing variants.

The output from VCFcompare showed poor overlap between the TVC and MuTect2 VCF files (Fig. 6). By using Venny, the overlap between VCF files could be further explored (Fig. 6, left panel). The results showed greater overlap between different samples from the same variant caller, than with their sample counterpart processed in another program. This persisted after variant filtration as well (Fig. 6, right panel).

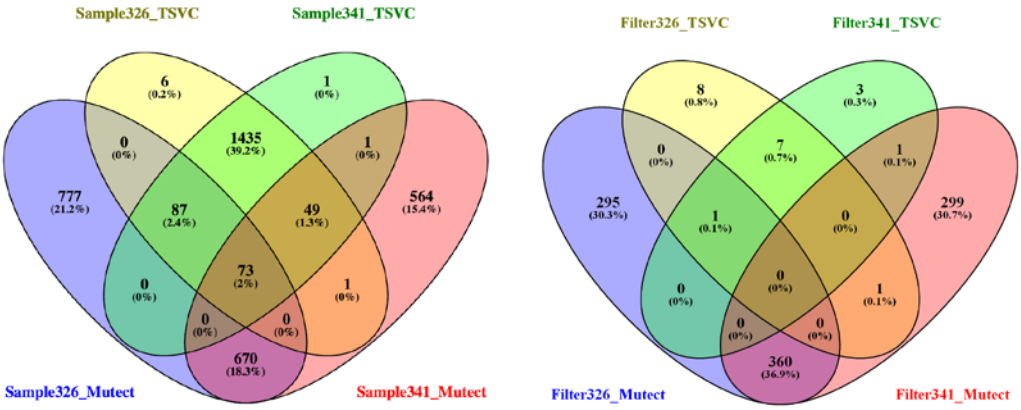


Fig. 6: Venny output. Left Venn diagram displays overlap for two TVC (denoted TSVC) and MuTect2 (denoted Mutect) output VCF files before filtering. Right Venn diagram displays overlap for the same two pairs after filtering.

3.2. Comparison between several variant callers

Due to that MuTect2, is recently developed and not been extensively used in previous work, it is necessary to also use other variant callers for comparison to TVC. This was performed in a similar way as described above in the section 3.1 for TVC and MuTect2, but in this second section comparing MuTec2, its predecessor MuTect, as well as Strelka and VarScan2 in parallel to TVC, and more extensively.

3.2.1. Variant validation of TVC, MuTEC2, MuTec, Strelka and VarScan2, and comparison

In Table 1, the data from all 13-paired tumor and normal samples are depicted for the different variant callers.

The absolute numbers of variants varied between the different variant callers especially for some samples e.g. PROM 137 and PROM 341, where TVC and VarScan2 show much fewer numbers of variants than the other variant callers.

In contrast for PROM 100 and PROM 141 the absolute numbers of variants obtained with the different variant callers were relatively similar.

However, Table 1 does not show whether the aforementioned variants overlap. For this purpose a number of Venn diagrams were made of which Fig. 7 and 8 depicting INDEL and SNP variants for two patients PROM 36 and 137 respectively, with the complete Venn diagrams for all samples found in the Appendix 7.4.

Table 1: Numbers of variants (SNPs & INDELs) for the respective variant callers for the 13 paired tumor/ normal samples.

	PROM 36	PROM 92	PROM 100	PROM 107	PROM 108	PROM 110	PROM 122
MuTect	2	0	16	0	16	121	0
MuTect2	9	12	13	17	41	89	3
Strelka	44	38	37	36	45	64	29
TVC	0	17	14	12	12	15	1
VarScan2	15	5	7	7	16	3	2
	PROM 137	PROM 141	PROM 145	PROM 322	PROM 326	PROM 341	
MuTect	234	15	0	0	0	292	
MuTect2	520	18	17	529	588	599	
Strelka	327	50	52	368	430	381	
TVC	9	13	12	11	17	12	
VarScan2	24	10	3	28	46	37	

Incomplete overlap was obtained, but some variant callers were more similar to each other than others.

MuTect2 and Strelka often showed a similar absolute number of variants as depicted in Table 1, and there were also repeated overlaps between the two variant callers in the 13 paired samples that were analyzed here as also shown in Table 5 (See below section Analysis of the Results). For details see Appendix 7.4.

For INDELs no variant in the same sample was called for all variant callers, while 16 variations in specific samples were covered by three variant callers. For SNPs, one variant in one specific sample was covered by all five variant callers, and five variants were called by four variant callers, while >50 variants were covered by three variant callers.

That all variant callers overlapped was seen only for PROM 137 (for the gene PTEN) and is depicted in Fig 8, and data not shown. Overlaps between four of the variant callers could e.g. be observed for PROM 107, 145 (both for the gene PIK3CA) and for PROM 341 (for the genes JAK2 and KDR) (Addendum 7.4, and data not shown). TVC and MuTect showed in general the least overlap, see Table 5 (Results and Analysis) and Appendix 7.4.

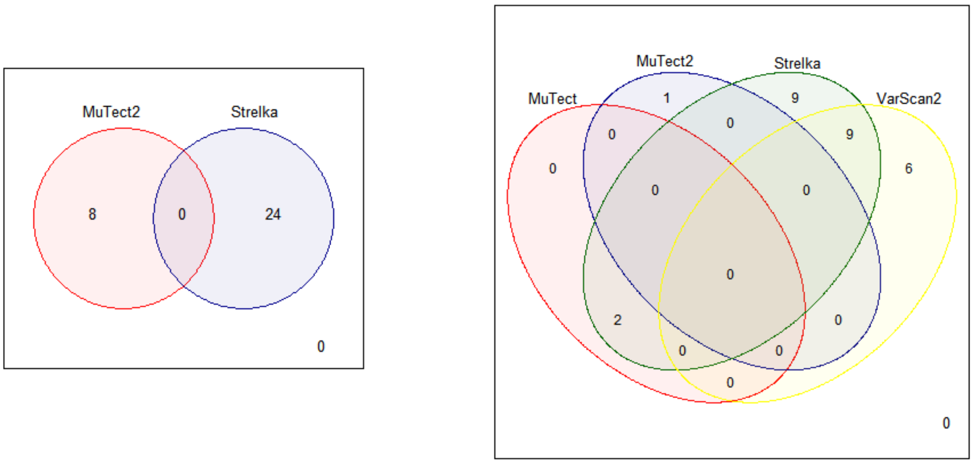


Fig. 7. Venn diagram for PROM 36, with no overlap for INDEL (left) and limited overlap for SNP (right).

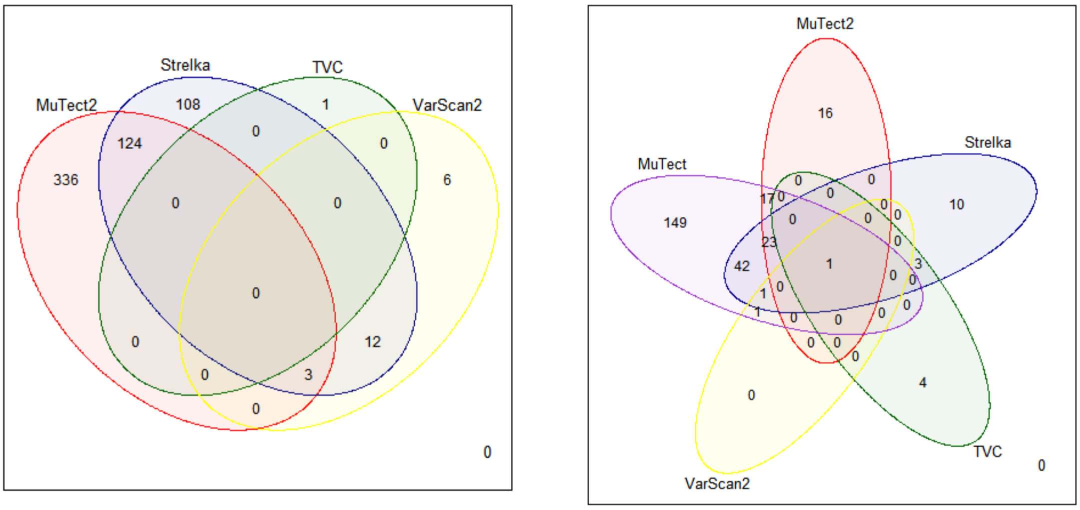


Fig. 8. Venn diagram for PROM 137, showing several overlaps for INDEL (left) as well as for SNP (right).

3.3 Variant annotation

The main aim of this thesis project is to investigate if there are age-related gene mutation patterns, as well as if mutated oncogenes occur in clusters, differing between tumors in young and old patients. To pursue this, DNA sequencing data from 299 HPV^+ TSCC/BOTSCC tumors was analyzed. Here, SnpEff was run on the 299 tumor samples using GRCh37 as human reference genome for annotation, followed by a first filtration to exclude low impact variants using SnpSift and a second filtration to exclude all variants not present in the COSMIC database using an R script (Fig. 9).

3.3.1 Annotation using SnpEff and filtration using SnpSift and R

SnpEff was used to annotate all 299 HPV+ TSCC/BOTSCC tumors samples with default settings. The first step was to remove the "No call" (low quality variants) and "Absent" (equal to reference genome) variations from TVC VCF files as described for the validation section (see paragraph 3.1.1). Then, the SnpEff output files were further filtered with SnpSift toolbox to keep only variants with impact category "moderate" and "high", that is, only variants with putative deleterious effects (deletion, insertion/deletion frame shift, donor splice site disruptions, acceptor splice site disruptions, stop codon gains, stop codon losses, start losses).

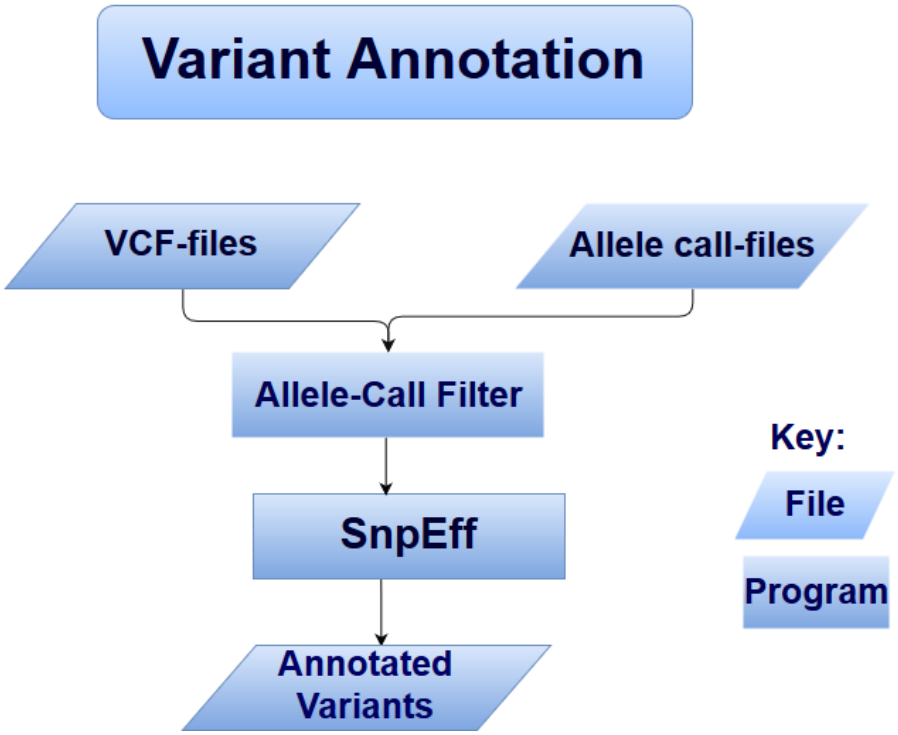


Fig. 9. Variant Annotation outline. A new introduced method was allele call files. These were from the Torrent Variant caller and were used to filter out calls of bad quality, and absent calls from the VCF files before SnpEff.

Additionally, in order to include only variants previously reported as somatic, and therefore to filter out possible germ line mutations, variants not annotated in COSMIC database were excluded. This filtration was managed via an R script. Likewise, any sample not carrying any variations where filtered out (Fig. 9).

To simplify statistical analysis and interpretation, all changes in a gene for a patient was counted as one change for that gene.

3.3.2 Variant annotation results

After filtering out low quality variants, non-COSMIC variants and non-phenotype causing variants from the TVC VCF files, the average number of variants per sample was 5.2 (±2.24 SD) variants per sample, ranging from a minimum of 1 to a maximum of 27. Moreover, 16 samples out of 299 were removed, as they contained no variants after filtering (Table 2). Only 29 genes out of 50 were found mutated after the filtering. The top 10 mutated genes and their corresponding type of mutation detected are listed in Table 3. For more specifics of each step, see Table 4.

Table 2: Clinical attributes of data before and after filtering. Number of samples in cohort, range of ages in cohort and genes included and affected in at least one sample in the dataset.

		Before		After
Attribute	**Range**	**Number of**	**Range**	**Number of**
Samples	-	299	-	283
Genes		50		29
Age	30-91	-	30-84	-

Table 3: List of the top 10 mutated genes after annotation and filtering. The types of mutation detected are also listed.

Gene	Number of mutations	Type of mutation
APC	249	c.4479_4480delGG; c.3949G>T; c.4328delC
CSF1R	242	c.*35_*36delCAinsTC
DERL3	79	c.*742G>A; c.*727C>T
TP53	9	c.455delC; c.513delG; c.560-2A>T; c.736delA; c.511G>T; c.460_461delGGinsAT; c.733G>T; c.328C>T; c.652G>A;
PTEN	8	c.640C>T; c.733C>T; c.801+1G>T; c.391delA; c.1003C>T; c.389delG
ATM	6	c.*29C>G; c.*5C>T; c.*44A>G;
PIK3CA	5	c.1209C>A; c.331A>G; c.*29T>C; c.3139C>T;
RB1	4	c.940-1G>A; c.*7431T>C; c.951_954delTTCT
SMAD4	3	c.1573A>G
STK11	3	c.475C>T; c.772delG

Table 4: Total variants and samples removed after each filtering step[1].

Step	Samples	Variants
Original VCF	285	825788
Removing "No Call"	288	794433
Removing "Absent"	288	4085
After Annotation	285	3755
After removing low impact	284	1187
Removing non-Cosmic	283	943

[1]The "No Call" filtration was done via the XLS-files, where an extra column denotes call-quality and type, as well as absence of variant, as in the "Absent" filtration step. "After Annotation" where data extracted from the SnpEff output annotated VCF-files, both output filtrated from SnpSift and without was saved in the SnpEff step.

4. ANALYSIS OF THE RESULTS

4.1. Variant Validation – analysis

As mentioned previously, due to the fact that only 13 normal samples were available, normal-tumor paired variant calling to distinguish germ line and somatic variants could not be performed for the whole cohort. To obtain an approximation of the amount of germline variation in our cohort, additional variant callers were used for the tumor-normal pairs, in comparison with the same 13 tumor samples used by TVC unpaired. These data showed limited overlap in the section above.

The result from the comparison between the different variant callers in terms of total number of overlaps for INDEL and SNPS for the paired 13 tumor/normal samples is shown in Fig. 10. In Table 5 the absolute numbers of unique and overlapping variants (SNPs) are shown for the five variant callers. As stated above, there were most overlaps for Strelka and MuTect2, but as expected there was overlapping between MuTect2 and MuTect.

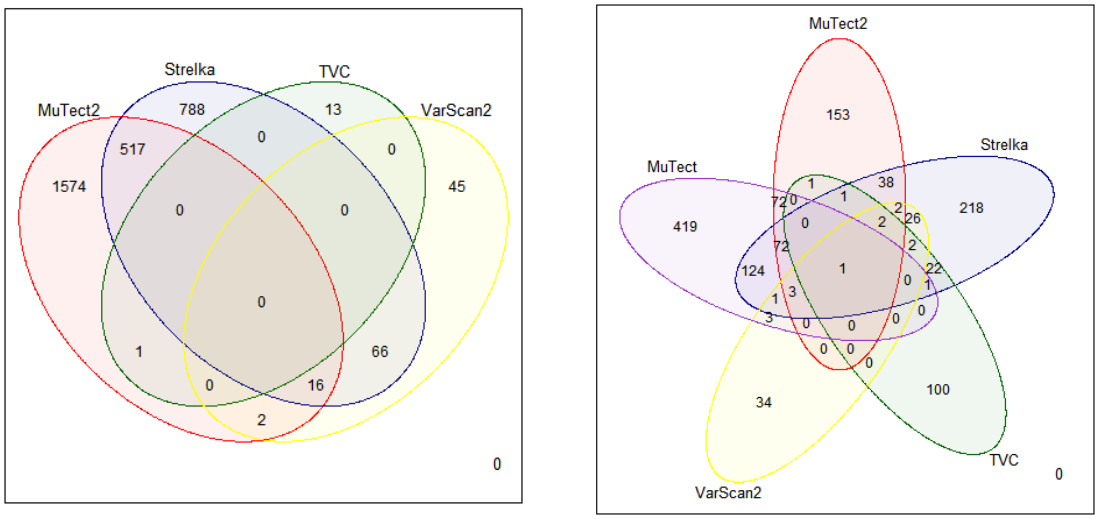


Fig. 10. All denoted overlaps of all the investigated samples for the different variant callers (left) INDELs and (right) SNPs.

Table 5: Numbers of SNP variants overlapping for each different variant caller.

	TVC	MuTect	MuTect2	Strelka	VarSan2
TVC	**130***				
MuTect	2	**693***			
MuTect2	4	148	**342***		
Strelka	28	202	119	**512***	
VarScan2	4	8	6	37	**74***

*Total number of SNPs per variant caller.

The data for SNP and INDEL variants respectively in Fig. 10 are also presented in Tables 6 and 7 respectively, where also the percentages are disclosed.

Table 6. Uniqueness and/or overlap sharing for each denoted variant (SNPs) in sample detected with the variant callers MuTect, Mutect2, Strelka, TVC, and VarScan2.

Variant detected per variant caller[1]	Number of variants	Percentage of variants
1	924	71.35
2	286	22.08
3	79	6.10
4	5	0.39
5	1	0.08

[1]Number denotes in how many variant callers the specific variant is detected

Table 7. Uniqueness and/or overlap sharing for each denoted variant (INDELs) in sample detected with the variant callers Mutect2, Strelka, TVC, and VarScan2.

Variant detected per variant caller[1]	Number of variants	Percentage of variants
1	2420	80.08
2	586	19.39
3	16	0.53

Tables 8 and 9 respectively, depict the percentage of SNP and INDEL variants respectively grouped per variant caller and are shown as unique to that variant caller/or shared with other variant callers. TVC has many unique SNP variants, while MuTect2, Strelka, VarScan2 show more overlap. Likewise, TVC shows many unique INDEL variants, as compared to the remaining variant callers that show more sharing.

Table. 8. Percentage SNP variants unique or shared for MuTect, Mutect2, Strelka, TVC, and VarScan2. (Variants of each "Variant Consensus" level, divided by total variants in caller).

Variant detected per specific variant caller unique or shared with other variant callers	MuTect	MuTect2	Strelka	TVC	VarScan2
1[1]	60.20	44.35	42.50	76.92	45.95
2	28.59	32.17	40.94	17.69	39.19
3	10.63	21.74	15.40	3.08	6.76
4	0.43	1.45	0.97	1.54	6.76
5	0.14	0.29	0.19	0.77	1.35

[1]Number denotes in how many variant callers the specific variant is detected

Table. 9. Percentage INDEL variants unique or shared for Mutect2, Strelka, TVC, and VarScan2. (Variants of each "Variant Consensus" level, divided by total variants in caller).

Variant detected per specific variant caller unique or shared with other variant callers	MuTect2	Strelka	TVC	VarScan2
1	74.60	56.81	92.86	34.88
2	24.64	42.03	7.14	52.71
3	0.76	1.15	0.00	12.40

The data can also be presented in a different way in order to show the sum of sharing between specific variant callers. This is done in Tables 10 and 11 for SNP and INDEL variants respectively. As already mentioned above, while TVC shows a generally low overlap with other variant callers, MuTect2 and Strelka share to a higher degree. However, as expected MuTect2 and MuTect also exhibit sharing.

Table.10. Percentage overlap for *SNPs* in each sample, between variant callers compared to total variants between the two variant callers compared.

	MuTect	MuTect2	Strelka	TVC	VarScan2
MuTect	100				
Mutect2	28.4	100			
Strelka	33.4	27.7	100		
TVC	0.5	1.7	8.7	100	
VarScan2	2.1	3.8	12.6	3.9	100

Table.11. Percentage overlap for *INDELs* in each sample, between variant callers compared to the total variants between the two variant callers compared.

	MuTect2	Strelka	TVC	VarScan2
Mutect2	100			
Strelka	30.5	100		
TVC	0.1	0	100	
VarScan2	1.6	10.8	0	100

To conclude, the initial data on Mutect2 with TVC showing limited overlap was not unique for MuTect2, but rather more specific for TVC. The data obtained later using the additional variant callers, MuTec, Strelka and VarScan2 showed a generally higher overlap. However, the combined data together did not impact on that TVC was used for the continued statistical analysis.

4.2. Analysis of mutations in relation to patient age.

The TVC variant caller was used to perform the statistical analysis since it is commonly used when performing sequencing with the Ion Torrent system. Furthermore, it includes both INDEL and SNPs and gives a moderate estimation of the number of mutations. In addition, the study by Hwang et al., [35] demonstrated that this TVC was the least error prone when performing whole genome sequencing analysis. In addition, the validation of the different variant callers (Section 3.2) did not confer a convincing cause to choose another variant caller.

An outline of the statistical analysis is depicted in Fig. 11.

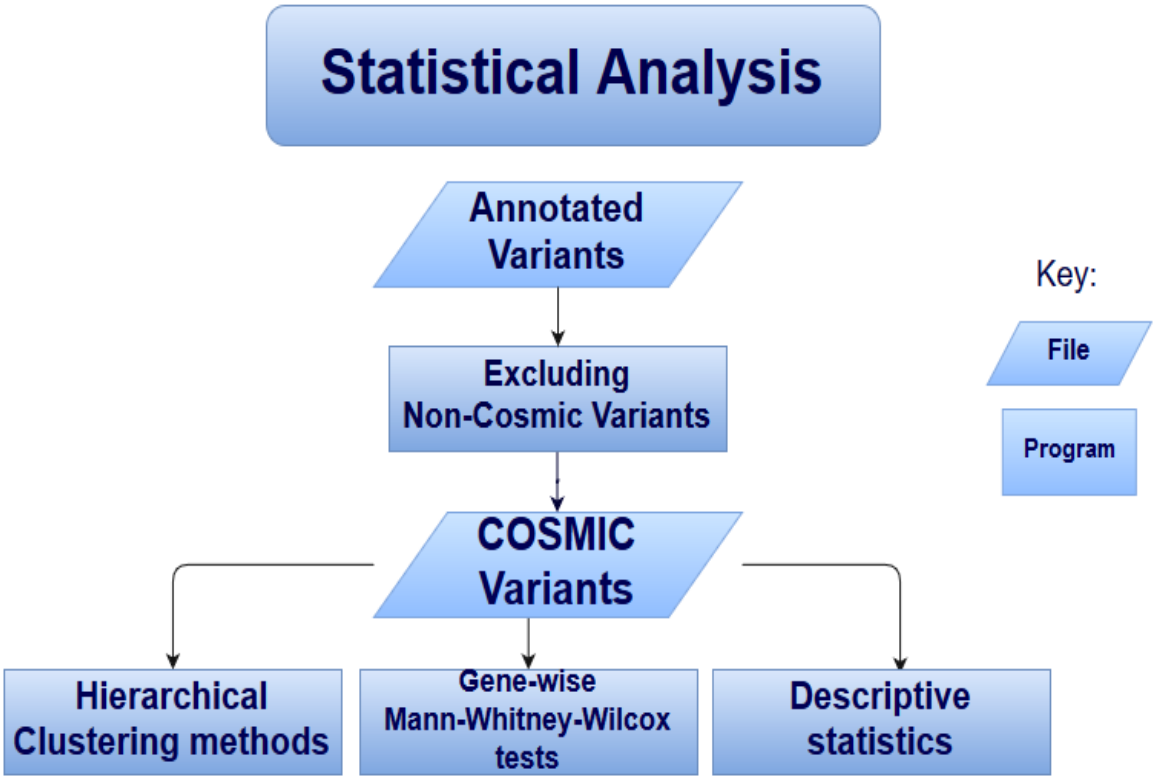


Fig. 11. Outline of the statistical analysis.

Performing a hierarchical clustering of filtered genes and samples did not reveal any specific clustering, with regard to variant distribution between patients (Fig. 2) or between age and mutation frequencies see Appendix 7.5.

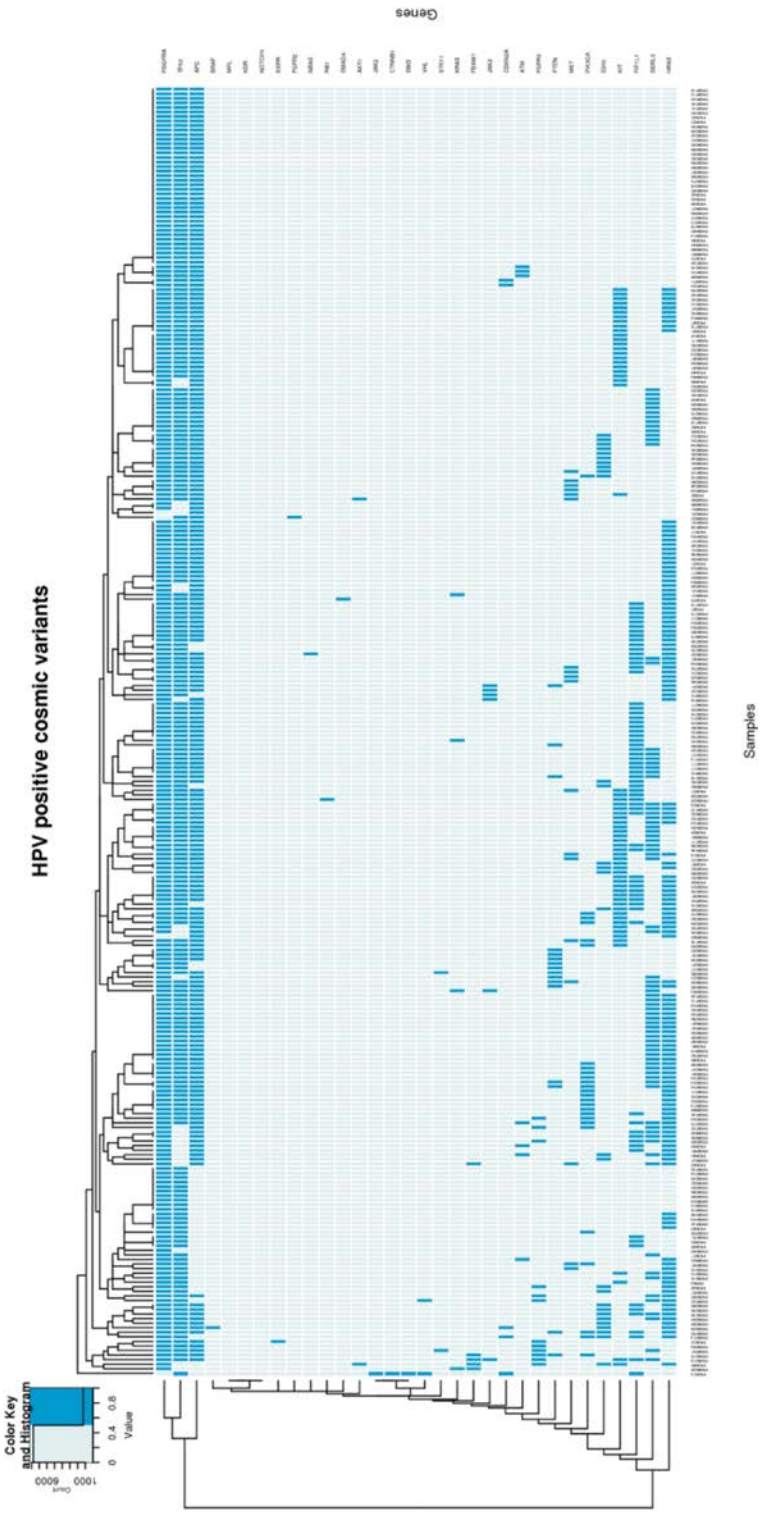


Fig. 12. Hierarchical Cluster heat map: Complete-linkage Hierarchical Clustering of filtered genes and samples.

Performing Mann-Whitney-Wilcoxon tests revealed no patterns of significant correlation between oncogene/suppressor mutations and age, for any of the examined genes (Table 12).

Table 12: Mann-Whitney-Wilcoxon test comparing mutation for each gene with the rank sum of age. The list shows ascending p-value for genes with p<=1.0. For specifics on filtering of the listed genes please see section 3.3.2.

Gene name	P-value
APC	0.054
BRAF	0.086
CDKN2A	0.127
AKT1	0.168
FBXW7	0.221
SMAD4	0.250
TP53	0.291
EGFR	0.304
PDGFRA	0.322
RB1	0.327
PTEN	0.335
CTNNB1	0.437
FGFR2	0.437
JAK2	0.437
SMO	0.437
JAK3	0.457
VHL	0.471
KIT	0.492
PIK3CA	0.496
IDH1	0.650
NRAS	0.654
HRAS	0.658
MET	0.679
STK11	0.709
FGFR3	0.714
KRAS	0.737
FIP1L1	0.752
DERL3	0.920
ATM	0.972
KDR	1
MPL	1
NOTCH1	1

Calculating mutations per person (frequency) for unique age groups in the cohort, revealed ~4-7 mutations in the tested oncogenes and suppressor genes per sample as shown in a histogram in Fig. 13. However, as depicted no clear difference with respect to age was found.

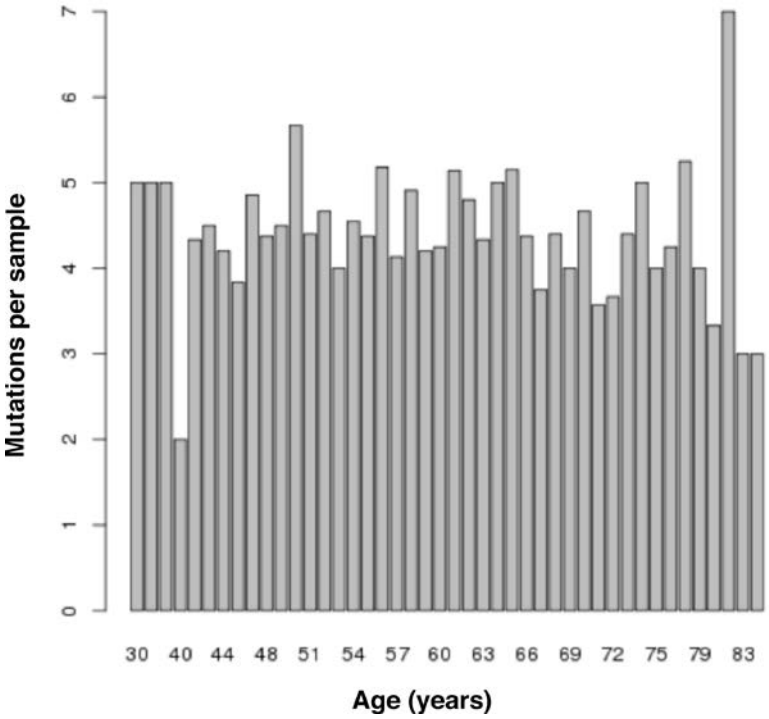


Fig. 13. Histogram of Mutation frequency per age group: Graph represents data divided by each unique age in the sample. Y axis corresponds to all mutations divided by all samples in each age group. All samples are divided into groups based on all values of age found in the dataset, 58, 59 etc. The height of the staple represents all mutations among patients with the age the staple represents, divided by the amount of samples in the staple.

However, when examined for individual genes, it was shown that mutations of the gene APC tended to be slightly more prevalent in younger patients, in that more APC mutations were detected in the youngest (≤ 40 years), with a slightly lower mean-age (Table 13).

Table 13: The descriptive statistics for carriers of APC, and the rest of the cohort.

	Mutated APC	Non-Mutated APC
Attribute	Value	Value
Standard Deviation	9.951	9.782
Mean	58.89	59.33
Number of samples	243	282
IQR	13	12
Range of ages	30-84	30-84

We also were interested to examine the age distribution and if it could have affected our results. For this purpose a box plot was made and shown in Fig. 14. In general there was a concentration of the patients towards the middle as could have been expected.

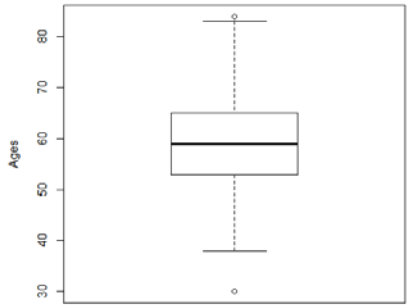


Fig. 14. Boxplot of cohort age distribution: The distribution is normally divided but quite wide, with a few outliers, the most extreme around 30 years of age.

Due to the close-to-significant p-values for the three genes APC, BRAF, and CDKN2A, a generalized linear regression and linear regression for correlation with age was performed. P-values for APC and BRAF were significant. (Table 14) All correlations with age were negative. The coefficient for BRAF was quite large (-23.00). However, the standard error for BRAF was considerable (10.98).

Table 14: General linear regression and linear regression for the three most common genes. For all these genes, we can see correlation with a decrease in age of carriers of the mutation.

	APC	BRAF	CDKN2A
Linear Regression			
P-value	0.0388	0.037	0.2238
Coefficient	-3.494	-23.00	-6.056
Std. Error	1.683	10.98	4.967
General linear regression			
P-value	0.0388	0.037	0.2238
Coefficient	-3.494	-23.00	-6.056
Std. Error	1.683	10.98	4.967

Further, a PCA was performed on the dataset for age and all 29 mutations. For all the principal components proportional variance was 3% (data not shown). Here, no predictive profile of variables could be deduced.

5. DISCUSSION AND CONCLUSION

5.1 Discussion of results in relation to the aim

The aim of this project was to investigate if there was a difference in mutation distribution or mutation patterns observed among any of the examined 50 oncogenes or tumor suppressor genes that could be correlated to the different age of the HPV^+ TSCC and BOTSCC patients. The reason for this was that it has been shown that younger patients have a better response to treatment than older patients, and this area had not been investigated for patients with HPV+ TSCC and BOTSCC.

This was done using sequence data from exons of 50 oncogenes and suppressor genes from 299 tumors. In addition, to get a deeper understanding of the data information, and having a an innovative input, the novel MuTect2 variant caller was compared to the TVC variant caller by analyzing 13 paired tumor and normal samples, and then these variant callers were both compared to MuTect, Strelka and VarScan2.

The initial lack of overlap in the variant calls between MuTect2 and TVC was an unexpected finding. It could be explained by the fact that some of the new implementations by adapting MuTect2 to the HaplotypeCaller framework as well as newly implementing INDEL calling, could have created issues with the variant calling.

However, since novel MuTect2 had not been described in the literature, an additional analysis was performed utilizing other variant callers MuTect, Strelka and VarScan2. Previous to this analysis a filtering step to better adapt the variant callers to each other was performed. This did still not result in high overlap with TVC. Nevertheless, MuTect2, showed best sharing with Strelka, but also as expected with its predecessor MuTect.

There can be several reasons for the different overlap. Different variant callers have different strengths and weaknesses. Some can be due to e.g. parameters that the user cannot influence or that the variant callers use different ways to filter the data. Furthermore, it has previously been published that it is not infrequent that different variant callers do not always show complete or excellent overlap, especially when using the Ion Proton sequencing platform, and many variant callers may perform better on the Ilumina platform [36]. More comparative investigations regarding specific differences could be of interest.

Based on the data obtained with the different variant callers, it was not obvious that TVC should be exchanged for the continued analysis and TVC was therefore utilized for the continued investigation of numbers mutations (thus including both germline and somatic mutations) within the tumor samples per age group. According to the analysis, there were no statistically significant differences in mutations of the analyzed oncogenes or tumor suppressor genes between older and younger individuals.

Still, mutations in the APC gene were more frequently observed in the young patients (See Tables 12-14), but the low p-value makes this observation inconclusive. This is something that could readily be followed up in future studies.

On average, we found 4-6 mutated genes per person for most ages. A widely accepted paradigm is that the increase of cancer with age is linked to an accumulation of mutations necessary to present the cancer phenotype. Specifically, a need for 6-7 oncogenic mutations for cancer development was suggested [37, 38]. Considering that we are screening only 50 genes using our cancer gene panel, our result is in concordance with earlier reports in the scientific community.

5.2 Discussion of the used method(s)

DNA from tumors of 299 patients had been sequenced on the Ion Proton sequencer, while alignment and variant calling has been performed via the Ion Torrent Suite software and Torrent Variant Caller plug-in, respectively [27]. The average coverage depth per nucleotide in the sequenced files was above 500x for 95% of the samples, (Section 2.2.5). Since, for detecting human genome mutations, SNPs, and rearrangements, a depth of coverage from 10× to 30× is often recommended, with a good degree of certainty our dataset should be of appropriate quality.

However, despite a total sample size of 299 HPV+ tumors, having only 13 non-tumor samples posed a challenge in the analysis with regard to the possible influence of endogenous germline mutations that might have passed the filtering steps applied.

To get an approximation of germline to somatic mutations MuTect2 was compared to TVC for the 13 paired normal samples and tumor samples, this resulted in a very poor overlap and the analysis was extended to include additional variant callers and is discussed in more detail below.

Hierarchical clustering was performed in R using several different approaches, clustering on samples only, genes only, as well as on both genes and samples, and still, no clear clusters could be found. Complete linkage-clustering non-significant results (Fig. 2 and Appendix 7.5).

The Mann-Whitney-Wilcoxon tests were performed for all genes to calculate whether age was a relevant factor for their prevalence (Table 12). Since age was normally distributed, then a t-test could have been performed. However this was not done, since there were some outliers below the lower whisker (in the lower age group, below 1.5 IQR (Fig. 14). These outliers would be especially important considering that they are the ones least prevalent in the most significant gene in the study, APC. As Mann-Whitney-Wilcoxon tests do not rely on absolute values as t-tests, but the ranking of values, this slight asymmetrical distribution is somewhat mitigated.

For binary dependent variables (mutated or not mutated), and continuous independent variables (age), the Mann-Whitney-Wilcoxon tests suitability to ordinal data is more fitting, compared to the t-tests usage of mean values for variables. Multiple sample correction was not included; in retrospect this could have given more accurate results.

The validation of TVC variant calling using MuTect2 gave a very poor overlap (Fig. 6). This was not expected as previous versions of MuTect performed very similarly to many other variant callers[39]. However, MuTec2 is still in development and no comparisons have been published. The INDEL calling was added since the last version.

The filtering performed on MuTect2 before comparison of MuTect2 and TVC variants was only allele frequency adjustment, as most other thresholds were in concordance. Overlap was measured before and after filtration, filtration reduced uniquely called genes, but also removed some of the few overlapping genes. We can only speculate that the cause of poor overlap could be found in issues with the variant caller, or specific weaknesses of one of the variant callers found in this particular dataset.

In order to compare the data obtained above, the analysis was repeated after filtering output from MuTect2, MuTect, as well as Strelka and VarScan2. Notably, there was not any major overlap between TVC and any of the other variant callers and as mentioned above this could be due to several reasons.

When the initial comparison was performed with only MuTect2 and TVC, this caused some concern, since MuTecT2 was novel and there were no published comparisons to other variant callers. However, when comparing to additional more established and tried variant callers, the data were consistent with that TVC in general had a low overlap with other callers. Notably, there was a higher degree of overlap between MuTect2 and other variant callers especially Strelka as well as MuTect2s predecessor MuTect, and this suggested that the initial analysis that was performed in this project was indeed adequate. In fact, our data suggest that MuTect2 is a successful improvement to MuTect, at least in this context.

Similar to the present findings, it has been reported before that when different variant callers are compared a complete overlap is not generally found and it has been reported that different variant callers have different strengths and weaknesses and find different variants [35, 36, 39, 40]. In the future the reasons for this could be of interest to pursue further.

Nonetheless, the data that were obtained with the different variant callers did not argue for not pursuing the mutation to age analysis using TVC and this was eventually done.

For most steps, the High Performance Computation cluster did not present clear advantages. However, for the more computationally demanding tasks of annotation and variant calling, designing the scripts in a way that partitioned the work on samples, gave significant time saving, that in turn was very useful for debugging the code when issues had arisen, even when the error was encountered well into submitted jobs.

The use of PANTHER and/or DAVID pathway analysis software was initially taken in consideration, as they could have been an informative way of exploring possible connections between mutation profiles and age related pathways. However, due to the results obtained from the Mann-Whitney-Wilcoxon tests, this was impossible to be performed. At least a few correlated genes would have been needed to give any clear pathway connections.

Weaker individual correlations that manifested as correlation with age in specific mutation profiles, could have been discovered trough either Multifactor analysis, Rough Sets algorithms or supervised PCA, but had to be abandoned due to time constraints. However, all of the above mentioned methods do require some measured clinical correlation with mutations to be effective, and for age, this presence is debatable due to the high p-values for the Mann-Whitney-Wilcoxon tests.

The APC gene was briefly investigated in the GenAge database and was not found as an entry. The other genes with larger p-values, where considered not relevant and not included in the search.

5.3 Discussion by relating to other relevant work in the field

This project aimed to identify whether HPV$^+$ TSCC and BOTSCC have distinct gene mutation profiles, for 50 commonly mutated genes in cancer, in young patients as compared to older patients.

To achieve bioinformatics novelty, to understand the data more thoroughly, since there were only 13 paired tumor normal samples from the 299 patients, and to get an approximation of germline to somatic mutations, the novel MuTect2 was compared to TVC for the paired normal samples. However, a poor overlap was obtained between MuTect2 and TVC and the analysis was therefore extended to include the variantcaller MuTect a predecessor to MuTect2, as well as the callers Strelka and VarScan2. TVC showed in general a limited overlap with all variant callers, but there was overlap between MuTec2, Mutect and especially Strelka.

The obtained data indicated that MuTect2 in fact has potential improvements to its predecessor, and could be a useful successor to MuTect, since beside new functionality as identifying both SNPs and INDELS, it had better overlap with Strelka. The fact that Strelka had high sensitivity in sequence data with a low allele fraction had been published before [15], making overlap for other variant callers a positive. However, as mentioned previously, it has been reported before that some variant callers perform better in sequences obtained using an Ilumina platform (luminescensce based data) rather than in sequences obtained using an Ion Proton platform (PH-based data) [26]. Nevertheless, a high overlap is not always obtained or necessarily correlated with a correct variant call [40, 41]. Obviously, there are many unanswered questions that would need further investigation.

No statistically significant differences in number of mutations according to age were found. However, there was possibly a trend for more mutations to be found in APC, since more APC mutations were found in the younger patients.

The low p-values of the Mann-Whitney-Wilcoxon test did not support the alternate hypothesis of distinct age related mutations. However, since information on human genes in this context is still incomplete, it is possible that other HPV$^+$ BOTSCC/TOSCC related genes are more readily mutated with age, also the fact that only 50 genes are covered here is an important consideration. Further studies are necessary to unravel such possibilities. For the material used in this study (FFPE tumor tissue), targeted sequencing using a cancer panel was a necessary limitation, but other studies with different material might be able to extend the analysis to more genes. Earlier mentioned databases, such as GenAge, house hundreds of discovered senescence correlated genes, conserved across multiple phylum of species, and would benefit from studies with a larger amount of studied genes.

To the knowledge accumulated in this project, no similar study has been done earlier. Previous studies on age related genetic alterations in human cancer have focused on

specific genes, or limited gene combinations and relatively limited numbers of patients with HPV$^+$ TSCC and BOTSCC are described. Moreover, there are no studies showing increasing numbers of APC mutations with age. Additional studies are however necessary before any conclusion could be drawn.

In a study by Batchelor et al [42], the authors were able to obtain sequencing data from 60 glioblastoma cases and analyzed the frequency of age dependency and prognostic effects of the p53 mutation, deletions of CDKN2A/p16 and EGFR amplification, genes also of interest for TSCC and BOTSCC. The authors found that p53 mutations were significantly associated with patient age, and that the prognostic effects of p53 mutations, CDKN2A/p16 deletions and EGFR amplification were dependent on age.

To study, especially p53 and p16 here, was of importance, since these genes are known to be of interest in head and neck cancer. However, there were no differences in numbers of mutations in these genes, but this could be that there were in general few mutations among these genes in the HPV$^+$ TSCC and BOTSCC patients.

A review of more general interest hypothesizes whether loss of heterozygosity (LOH), i.e. loss of e.g. a chromosome, increases with age and whether this can lead to the inactivation of tumor suppressor genes and thus an increase of cancer incidence with age [43]. The authors discuss that understanding the effect of aging on this type of mutation is important. The authors point out that there are recent studies in model organisms showing increased rates of LOH with age, and that repair of DNA damage occurs via a different pathway in old cells versus young cells.

Another study focused on whether specific translocations between chromosomes were more frequently observed with age (<1 year to 91 years of age) [44]. They studied normal peripheral mononuclear blood cells (PMNBC) and could show that with age a translocation of 14:18 was more common with age, but absent in children <10 years of age.

5.4 Highlight novelty

- This is the first study attempting to correlate mutations in 50 oncogenes and suppressor genes in HPV^+ TSCC and BOTSCC according to patient age in a sample size of HPV^+ tumors (299) that is larger than most other studies on the subject.

- MuTect2, a novel variant caller and a development of MuTect was compared to TVC, as well as MuTect, Strelka and VarScan 2. MuTect2 showed overlap with MuTect as expected and especially with Strelka. The data indicate that MuTect2 could be a successful successor to MuTect, since it shows SNPs variant calling similar to, and possibly slightly improved to MuTect, as well as comparable overlap for the newly included variant calling of INDELs.

- The poor overlap for TVC and the other variant callers not constructed for ion proton data, and also limited overlap for the other variant-callers among themselves, suggest that ion proton data should possibly be investigated for variants using TVC. However, a directed effort using synthetic and controlled data for comparison is warranted.

- After more thorough investigation of TVC has been performed, a possible project for other bioinformaticians is the production of new variant callers more suited for ion proton data, overcoming possible issues highlighted here.

- There were no major differences in the frequency of mutations among 50 specific oncogenes or tumor specific suppressor genes according to the age of patients with HPV^+ TSCC and BOTSCC.

- There were however more mutations of APC among younger patients which may render this gene of interest to follow up.

5.5 Description of ethical aspects and impact on society

Ethical permission: The project was performed according to ethical permissions 99-237; 01/-296; 2005/1330-32; 2009/1278-31/4 from the Regional Ethical Committee in Stockholm.

Tumor biopsy samples and blood samples were collected between 2000-2011 at the Clinic of Pathology, and the Dept of Oto-Rhino-Laryngeology, Head and Neck Surgery Karolinska University Hospital and handed in coded form to the laboratory. Key to the samples are not available to the researchers and the presented data cannot be traced back to individual patients.

Motivation for the research question addressed and the expected results:

This study is a continuation of a project ongoing since the 1990s, demonstrating that HPV is a causative agent of TSCC and BOTSCC, and that HPV^+ tumours have a better prognosis than the corresponding HPV^- tumors. Furthermore, the incidence of HPV^+ TSCC and BOTSCC has increased epidemically the past decades. Today, patients with HPV^+ and HPV^- tumors are treated in the same way with heavy irradiation and cytostatics leading to life long adverse side effects affecting quality of life and resulting in costs for society. The currently used aggressive treatment regimen is likely unnecessary for patients with HPV^+ tumors.

It also addresses the possibility to investigate the novel MuTect2 still in beta development and a successor to MuTect and to compare it with TCV and other variant callers.

Possible gains:

The novel MuTect2, is likely a useful successor to MuTect, in that it shows overlap with other variant callers and identifies both SNPs and INDELs.

Furthermore, this project adds to the information available on different variant callers in FFPE tumor material.

The ultimate aim of this project was to find possibilities to better individualize treatment, i.e. identify factors that can identify patients that need less therapy, or that could receive more targeted therapy. The data suggest that APC is a gene that should be looked into further among patients with HPV^+TSCCC and BOTSCC.

5.6 Description of future directions

As previously mentioned, the “APC” gene differed mainly from other genes, in its greater prevalence in younger subjects. However, the number of younger patients may have been slightly limited (20 < 40 years) to obtain significant statistics. The importance of this finding needs further investigation, with a larger proportion of younger (< 40 years) patients.

The novel MuTect2 showed a promising profile as compared to its predecessor MuTect as well as the other variant callers Strelka and Varscan 2 despite none of these variant callers had a high overlap with TVC

Had there been more time it could have been of interest to further investigate where the discrepancies in variant calling actually originate.

Using several variant callers, an analysis within the same population, this time including a normal sample from each patient, and with a broader array of genes and whole genome sequencing to detect more mutations and their possible connection with age could be intriguing to perform.

It would elucidate the strengths and limitation of different variant callers as well as elucidate the whole genomes mutational profile.

6. REFERENCES

1. Nasman, A., et al., *Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: an epidemic of viral-induced carcinoma?* Int J Cancer, 2009. **125**(2): p. 362-6.
2. Tertipis, N., et al., *A model for predicting clinical outcome in patients with human papillomavirus-positive tonsillar and base of tongue cancer.* Eur J Cancer, 2015. **51**(12): p. 1580-7.
3. Braakhuis, B.J., et al., *TP53 mutation and human papilloma virus status of oral squamous cell carcinomas in young adult patients.* Oral Dis, 2014. **20**(6): p. 602-8.
4. Friedland, P., et al., *Human papillomavirus and gene mutations in head and neck squamous carcinomas.* ANZ J Surg, 2012. **82**(5): p. 362-6.
5. Farnebo, L., et al., *DNA repair genes XPC, XPD, XRCC1, and XRCC3 are associated with risk and survival of squamous cell carcinoma of the head and neck.* DNA Repair (Amst), 2015. **31**: p. 64-72.
6. Simbolo, M., et al., *DNA qualification workflow for next generation sequencing of histopathological samples.* PLoS One, 2013. **8**(6): p. e62692.
7. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic acids research, 2015. **43**(Database issue): p. D805-11.
8. Zhu, Z., et al., *Single-molecule emulsion PCR in microfluidic droplets.* Anal Bioanal Chem, 2012. **403**(8): p. 2127-43.
9. Merrian, B., et al. *Progress in ion torrent semiconductor chip based sequencing*. 2012, Electrophoresis 33(23):3397-417.
10. Gabriel, C., et al. *HLA typing by next-generation sequencing- getting closer to reality.* Tissue Antigens, 2014 83(2)65-75.
11. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
12. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Briefings in bioinformatics, 2010. **11**(5): p. 473-83.
13. *MuTect2 - GATK | GATK | Tool Documentation Index* https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php
14. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
15. Saunders C.T., et al. S*trelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs.* Science & Mathematics Bioinformatics. 2012. **28**(14)1811-17.
16. Koboldt, DC. *VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing* Genome Res. *2012. **22**: 568-576*
17. GATK | Tool Documentation Index. <https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_SelectVariants.php%3E.
18. Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. 27(15): p. 2156-8.
19. http://varscan.sourceforge.net/using-varscan.html
20. Pabinger, S. et al. *A survey of tools for variant analysis of next-generation genome sequencing data.* Brief Bioinform. 2014. **15**(2): 256–78.
21. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
22. *Exome Aggregation Consortium (ExAC), Cambridge, MA.* (URL: http://exac.broadinstitute.org).
23. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
24. Cingolani, P., et al., *Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift.* Frontiers in genetics, 2012. **3**: p. 35.
25. *GRCh37 | 1000 Genomeshttp*://www.1000genomes.org/category/grch37/

26. Lampa S, H.J., Spjuth O, *UPPNEX - A solution for Next Generation Sequencing data management and analysis*. EMBnet. journal, 2012.
27. Gibbons, F.D. and F.P. Roth, *Judging the quality of gene expression-based clustering methods using gene annotation.* Genome research, 2002. **12**(10): p. 1574-81.
28. Andy B. Yoo, M.A.J., Mark Grondona, *SLURM: Simple Linux Utility for Resource Management.* 2003: p. pp 44-60.
29. Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 1996. **5**(3):299--314.
30. Oliveros, J.C., *VENNY. An interactive tool for comparing lists with Venn Diagrams.* http://bioinfogp.cnb.csic.es/tools/venny/index.html., 2007.
31. Kasuya E. Mann–Whitney *U* test when variances are unequal, Animal Behaviour 61(6)1247-9.
32. Ward, Joe H. *"Hierarchical Grouping to Optimize an Objective Function".* Journal of the American Statistical Association. 1963, **58** (301): 236–44.
33. D. Defays. *"An efficient algorithm for a complete link method" (PDF).* The Computer Journal. British Computer Society. 1977, **20** (4): 364–366.
34. Kim S. et al. *High-Throughput Sequencing and Copy Number Variation Detection Using Formalin Fixed Embedded Tissue in Metastatic Gastric Cancer,* PLoS One, 2014, 9(11): e111693.
35. Hwang S et al. *Systematic comparison of variant calling pipelines using gold standard personal exome variants.* Sci Rep. 2015, **7**(5):17875.
36. Merriman, B. and J.M. Rothberg, *Progress in ion torrent semiconductor chip based sequencing.* Electrophoresis, 2012. **33**(23): p. 3397-417.
37. Armitage, P. and R. Doll, *A two-stage theory of carcinogenesis in relation to the age distribution of human cancer.* British journal of cancer, 1957. **11**(2): p. 161-9.
38. Armitage, P. and R. Doll, *The age distribution of cancer and a multi-stage theory of carcinogenesis.* British journal of cancer, 1954. **8**(1): p. 1-12.
39. Xu, H., et al., *Comparison of somatic mutation calling methods in amplicon and whole exome sequence data.* BMC genomics, 2014. **15**: p. 244.
40. Rashid M, Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in the cancer genomes. Bioinformatics aplications note 2013. **29**(17)2208-10.
41. Cornish et al. A comparison of variant calling pipelines using genome in a bottle as a reference. BioMed Research Int. 2015:456479
42. Batchelor, T.T., et al., *Age-dependent prognostic effects of genetic alterations in glioblastoma.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2004. **10**(1 Pt 1): p. 228-33.
43. Carr, L.L. and D.E. Gottschling, *Does age influence loss of heterozygosity?* Experimental gerontology, 2008. **43**(3): p. 123-9.
44. Dolken, L., et al., *High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay.* RNA, 2008. **14**(9): p. 1959-72.

7. APPENDIX

7.1 Library preparation.

The first step in library preparation is targeted amplification achieved through 20 cycles of a multiplex Polymerase Chain Reaction (mPCR). Primer pools are incubated together with 10 ng of genomic DNA, deoxynucleotides (dNTPs) and a DNA-polymerase. Multiplex PCR, allows concurrent use of several DNA-primers to amplify/replicate multiple genes or DNA sequences.

The second step in the library preparation is the partial digestion of the primer sequences. Primer sequences on previously obtained amplicons are partially digested and amplicons phosphorylated.

The third step implies ligation of adapters and barcodes to the amplicons. Here, the use of 96 different barcodes allows combination and accommodation of 96 different libraries (i.e. 96 different tumor samples) in a single sequencing run. Moreover, the adapters are important for binding to beads/ISPs later used in sequencing of the amplified DNA library. A Bioanalyzer instrument (Agilent) is then used to check quality and quantity of the obtained barcoded libraries. Finally, each of the 96 libraries is pooled in a single tube at a final concentration of 100pM.

7.2 Emulsion PCR

A clonal amplification is performed in an oil-aqueous emulsion that contains large enough oil droplets to accommodate one bead, an amplicon, DNA polymerase and dNTPs[8]. In each droplet regular PCR cycles are made. This process produces millions of beads covered with millions of different DNA fragments. Beads are then flowed across an Ion PI Chip (Thermo Fisher) for sequencing using a semiconductor based sequencer (Ion Proton sequencer).

7.3 Ion proton sequencing technology

The semiconductor based system of the Ion Proton sequencer is utilized for NGS. The technology is based on the fact that during the polymerase reaction, after the addition of a new nucleotide to the DNA polymer chain, a hydrogen ion is released, a.k.a a proton (H^+). This hydrogen ion has a charge that the Proton System's ion sensor can perceive using a "tiny pH-meter". Any nucleotide added to a DNA template will be spotted as a voltage change, and the sequencer will call the base. A nucleotide, which is not a match for a specific template, will result in no voltage change, and a base will not be called for that template.

7. 4 Venn diagrams

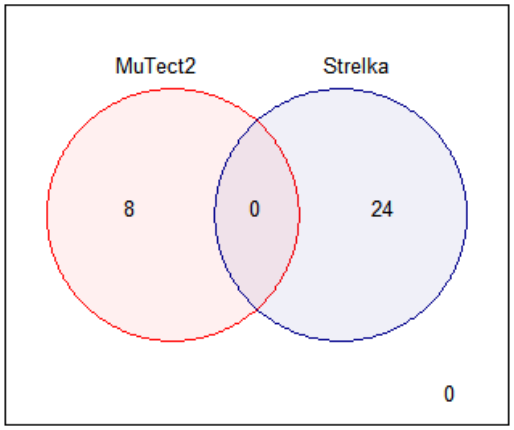
Below Venn diagrams have been made for the five variant callers for all 13 tumors with paired normal samples (PROM 36, 92, 100, 107,108, 110, 122, 137, 145, 322, 326 and 341). Each PROM is depicted on a separate page.

Figure information

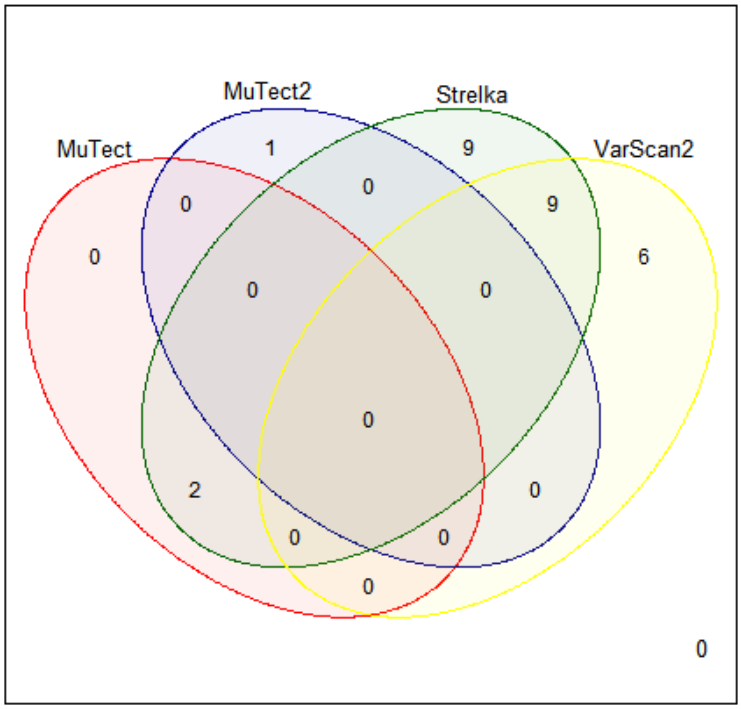
The top figures denote INDEL overlaps and the bottom figures on each page denote SNP overlaps for each PROM.

PROM 36

INDEL

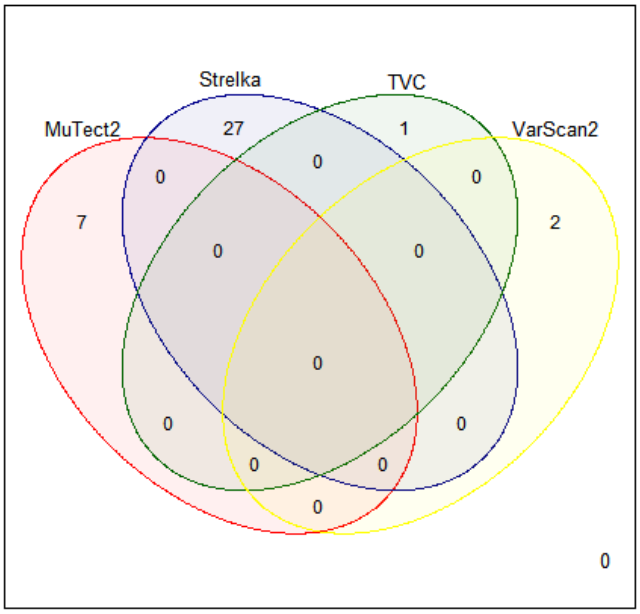


SNP

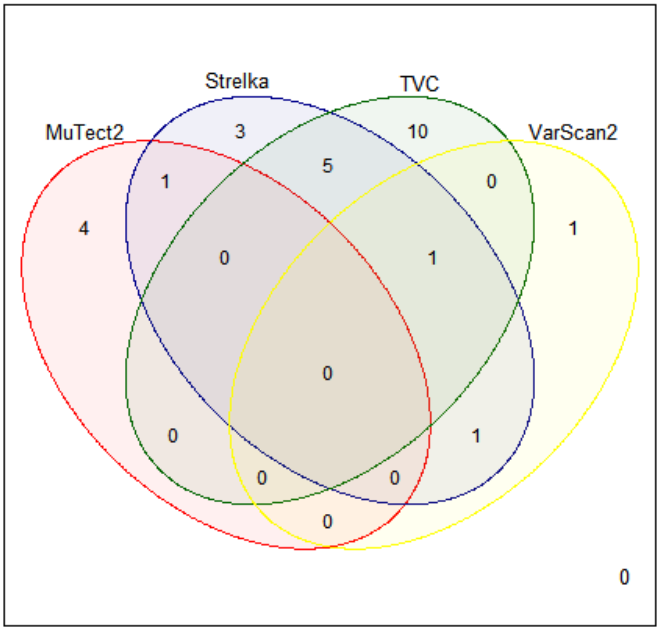


PROM 92

INDEL

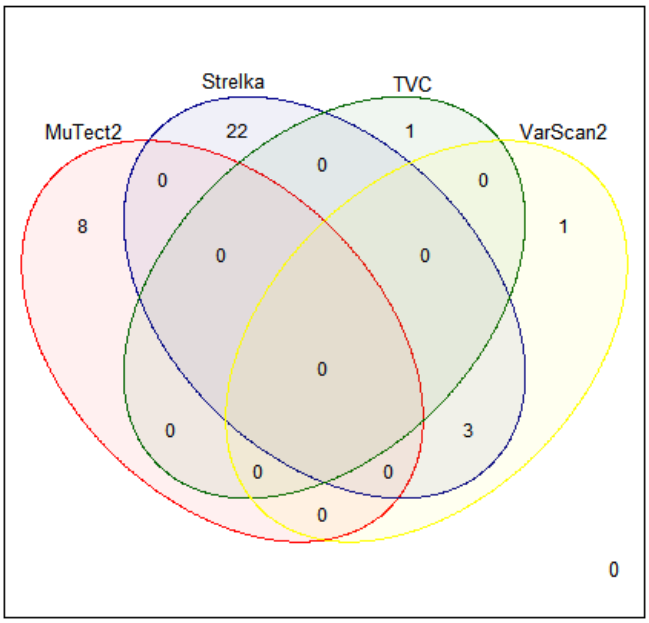


SNP

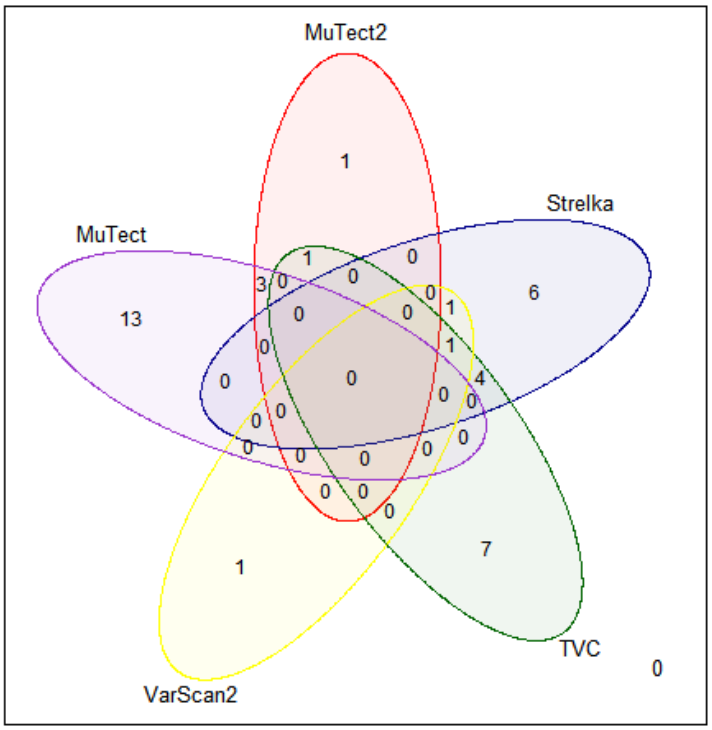


PROM 100

INDEL

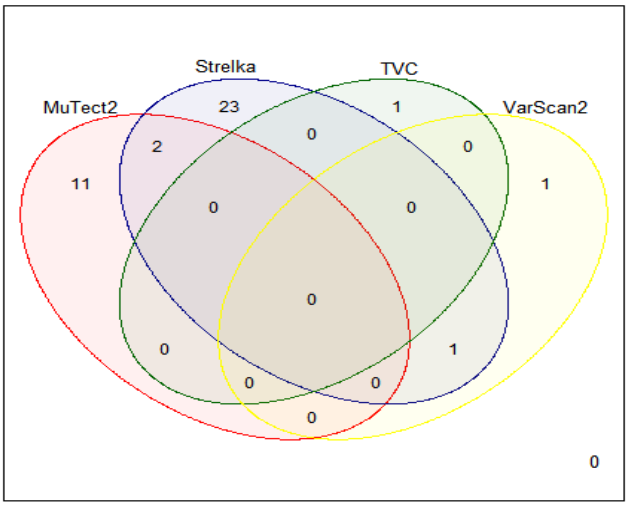


SNP

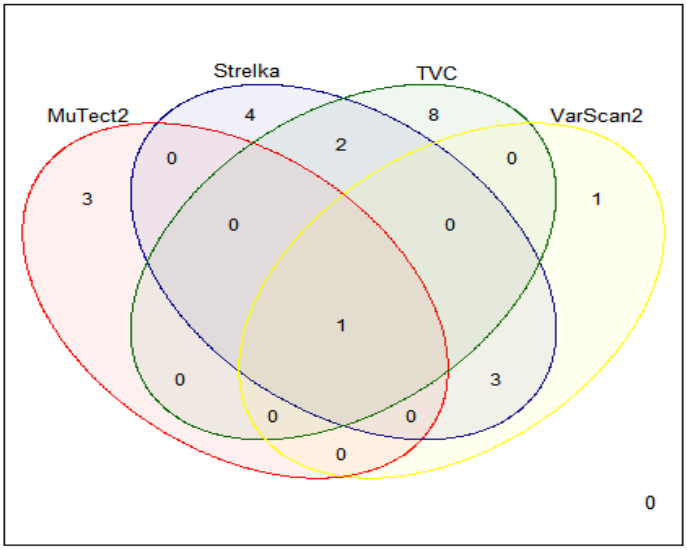


PROM 107

INDEL

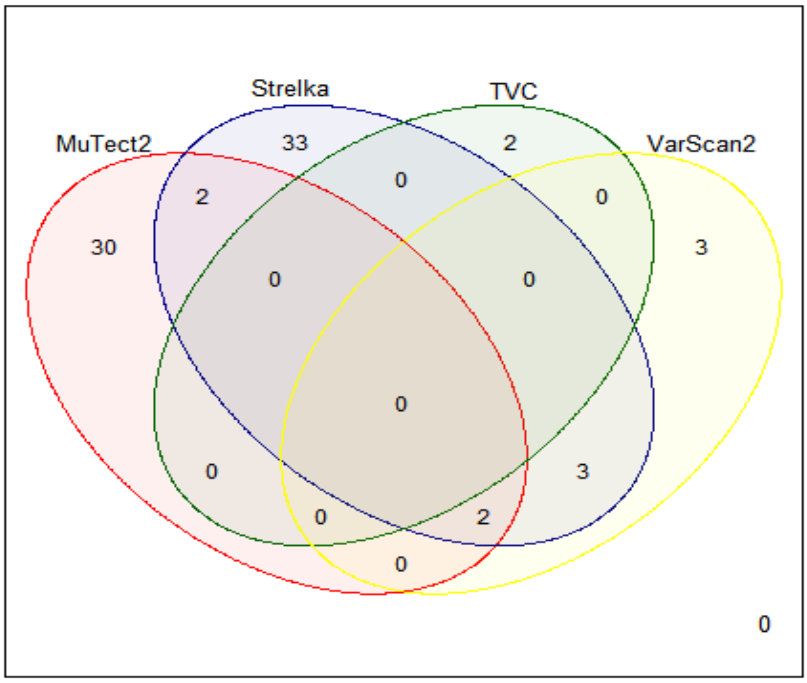


SNP

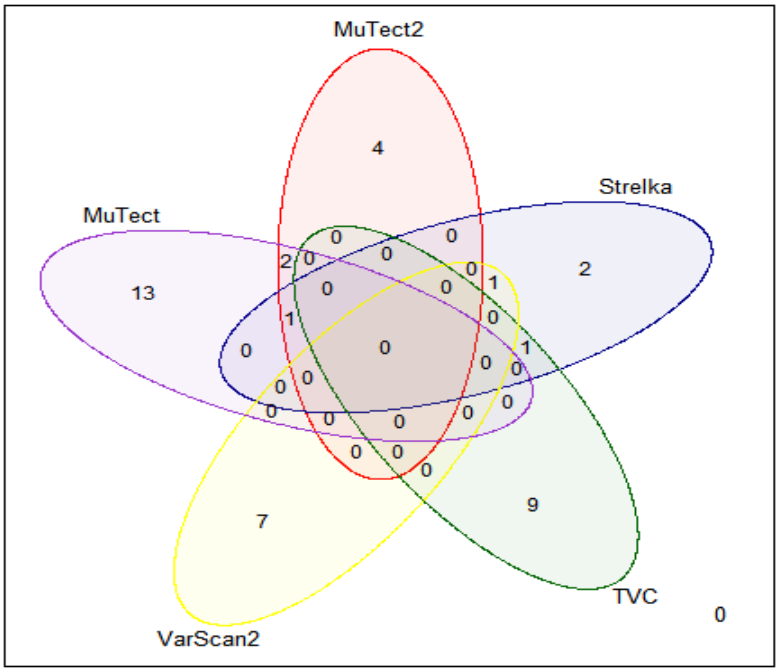


PROM 108

INDEL

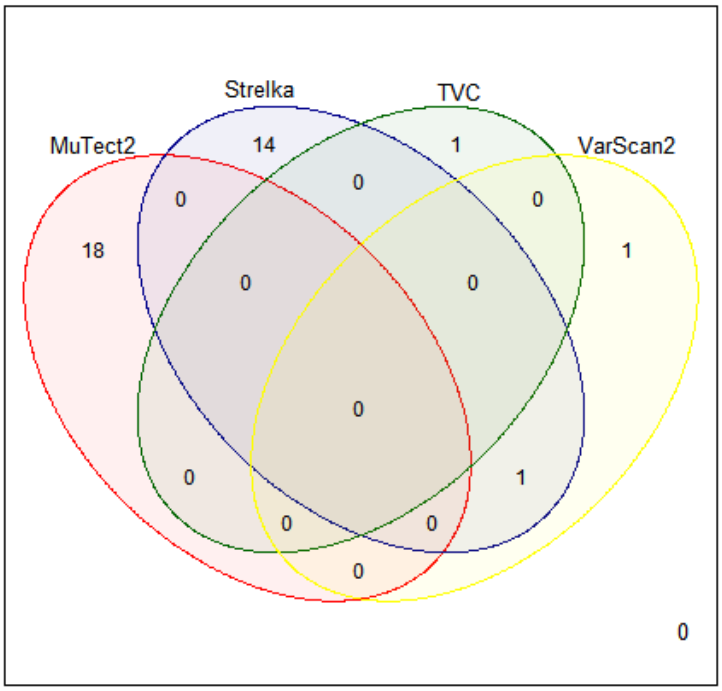


SNP

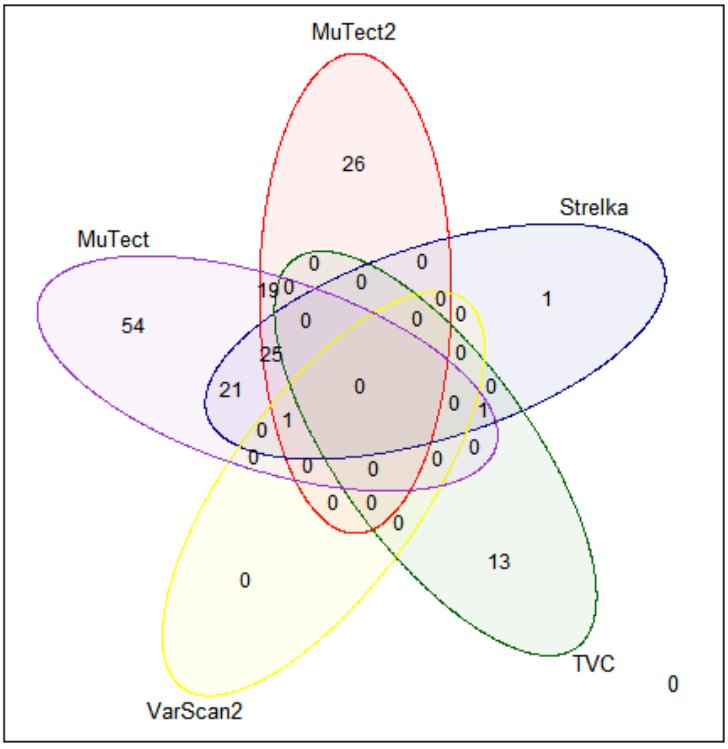


PROM 110

INDEL

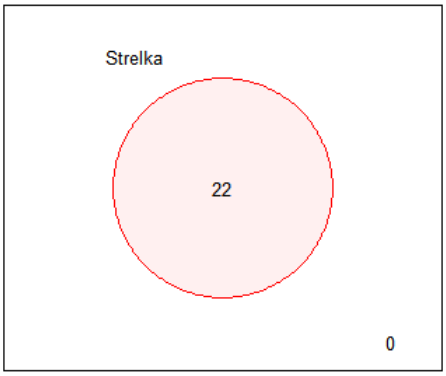


SNP

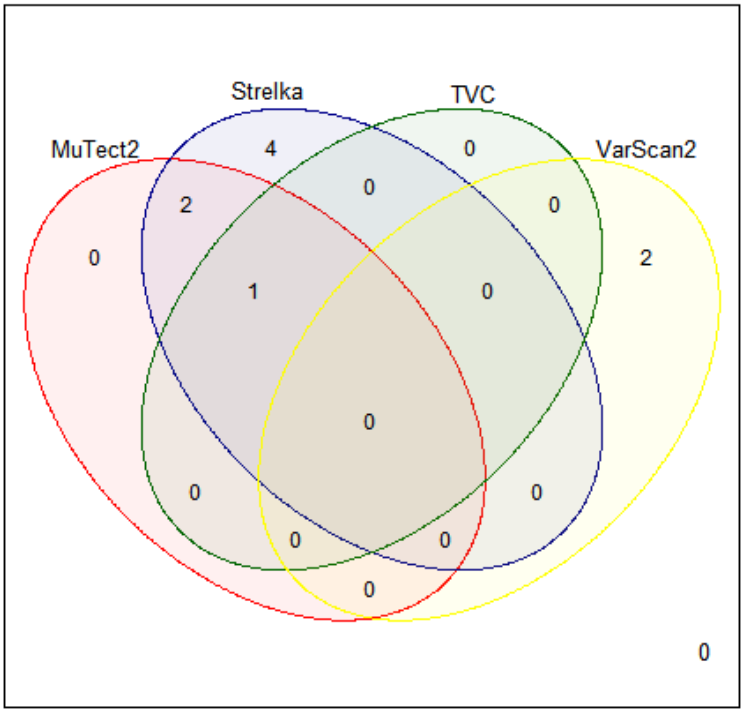


PROM 122

INDEL

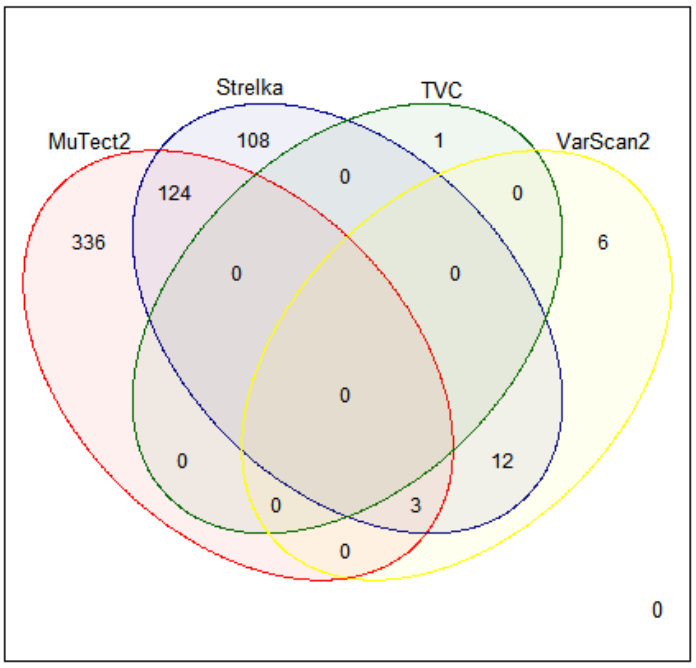


SNP

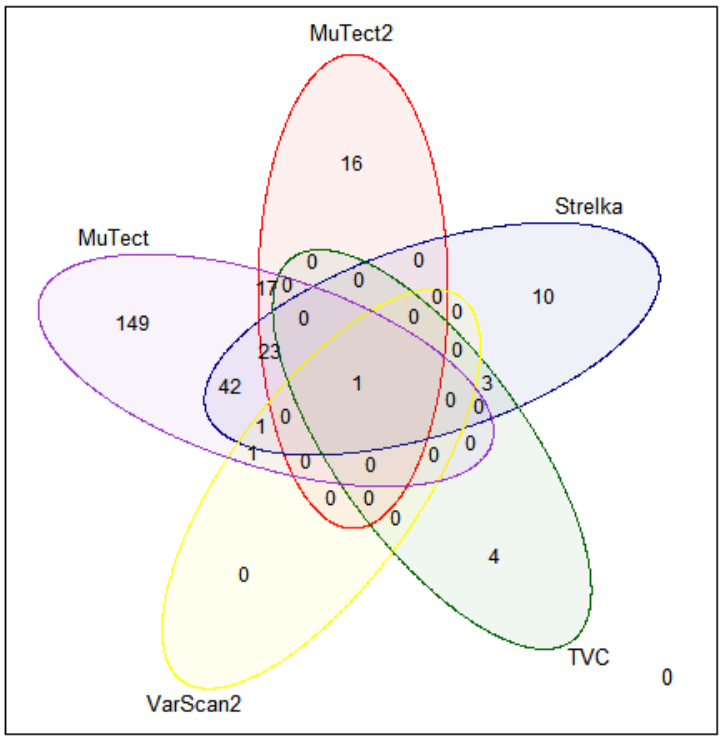


PROM 137

INDEL

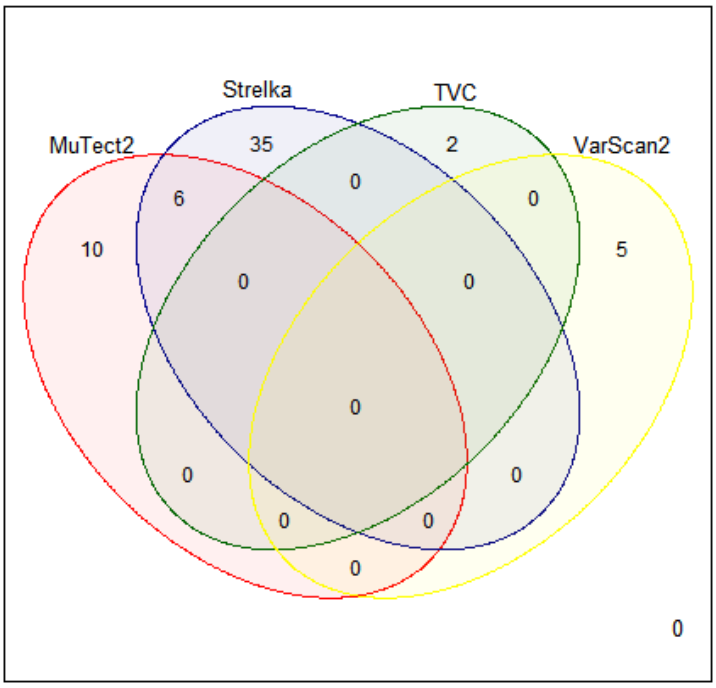


SNP

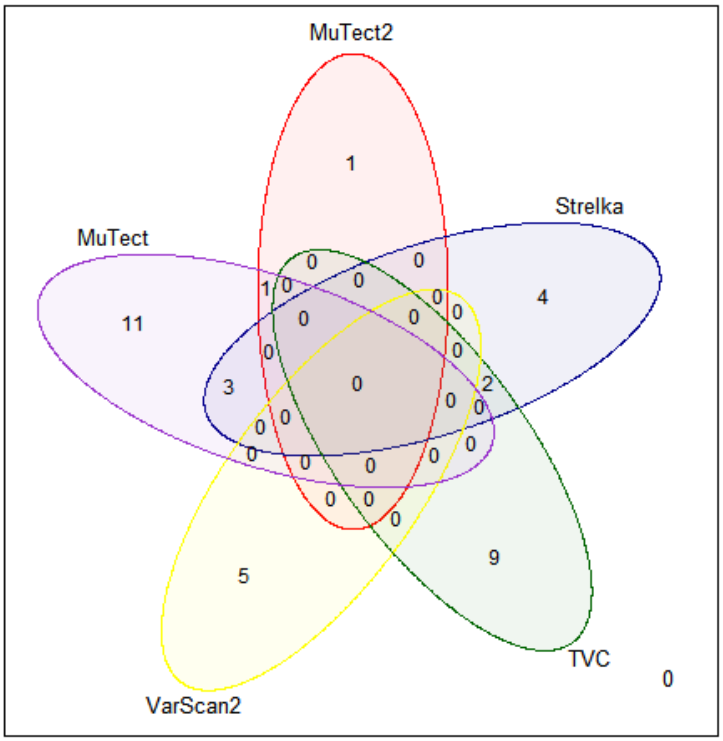


PROM 141

INDEL

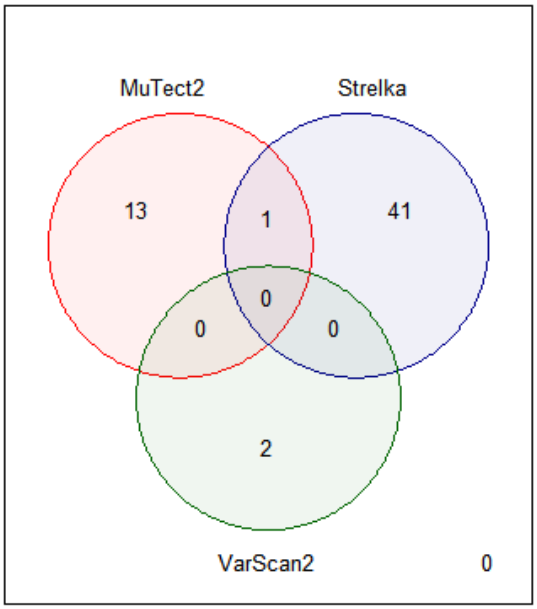


SNP

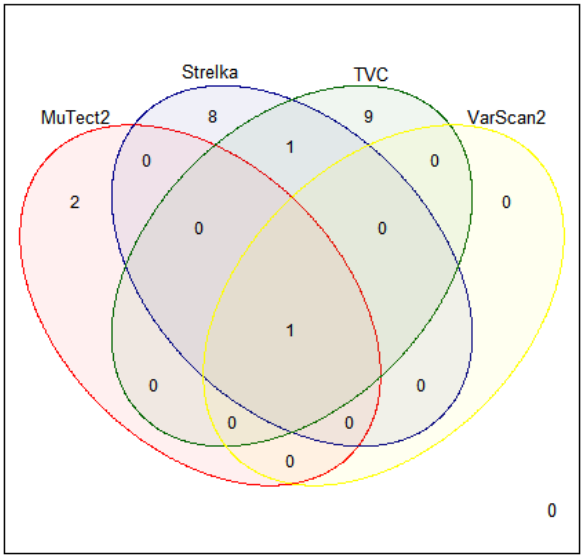


PROM 145

INDEL

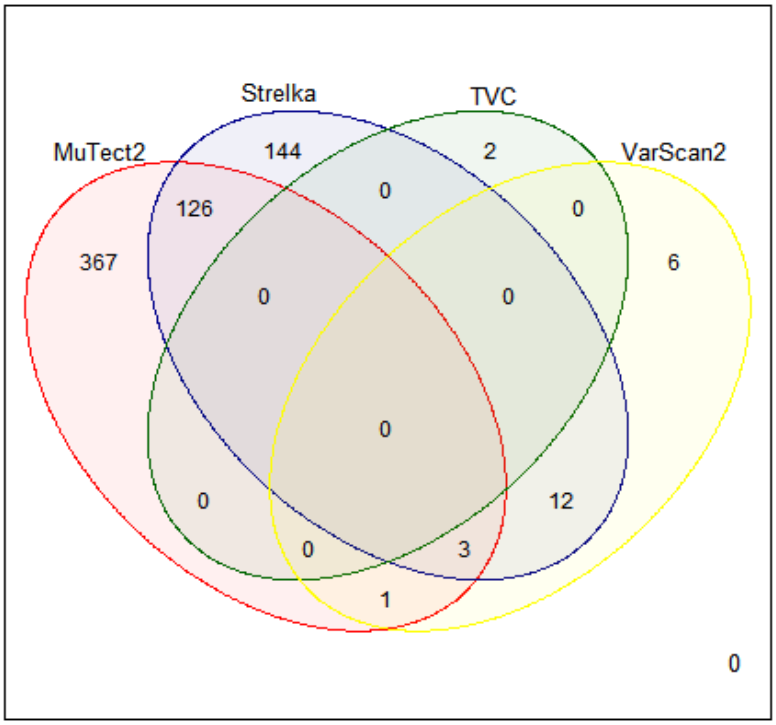


SNP

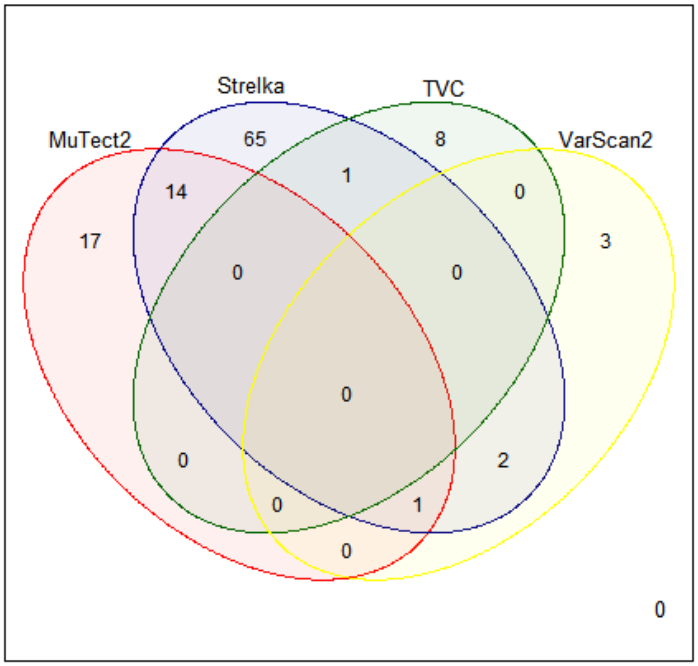


PROM 322

INDEL

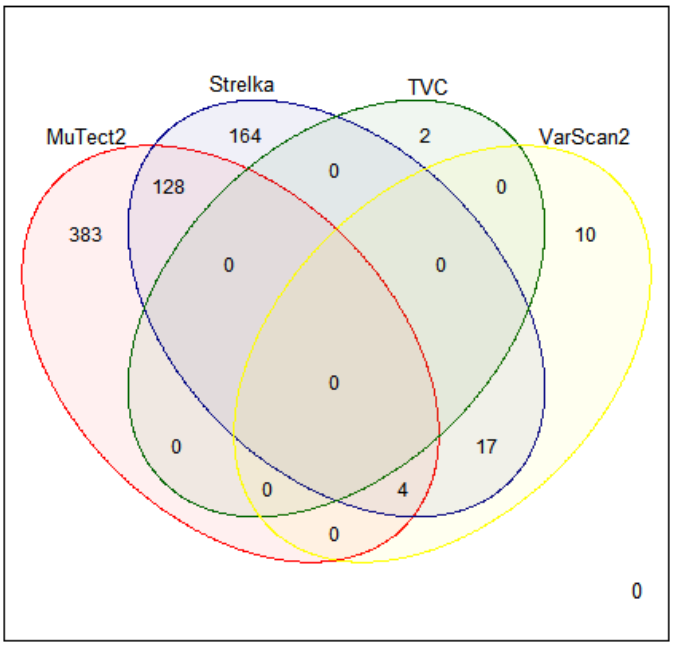


SNP

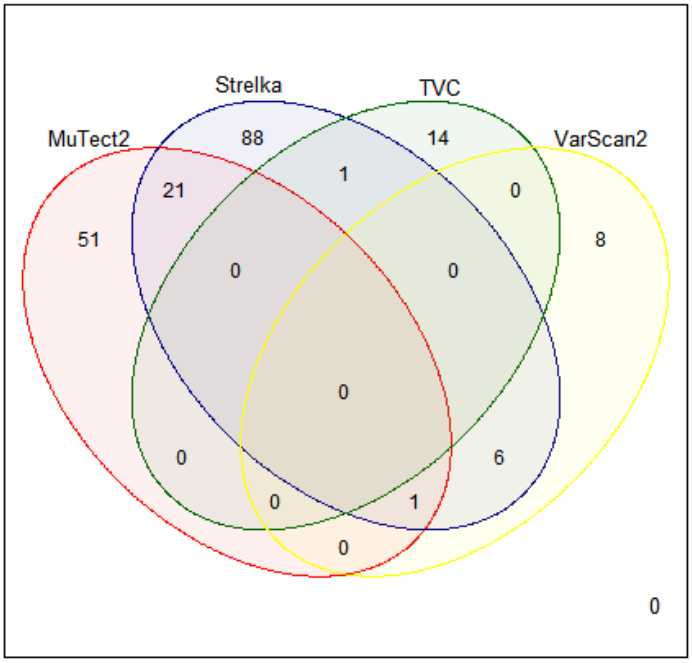


PROM 326

INDEL

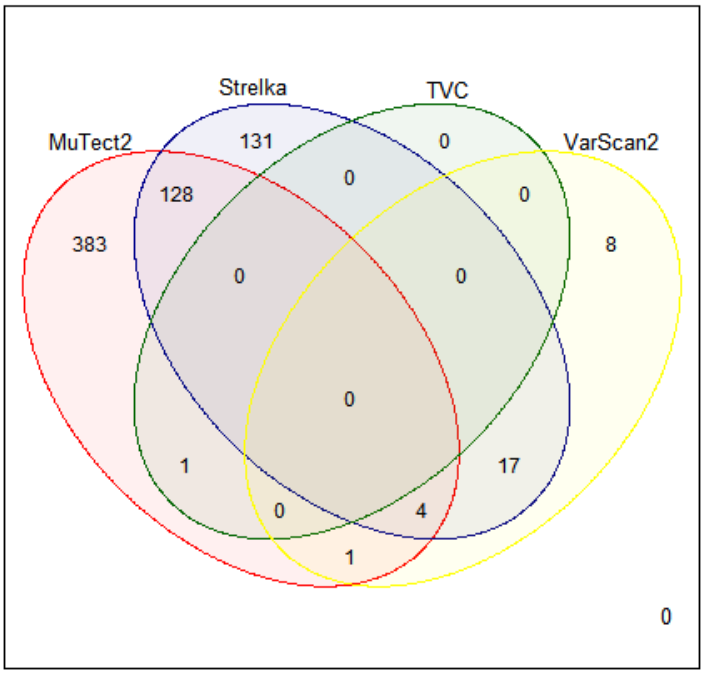


SNP

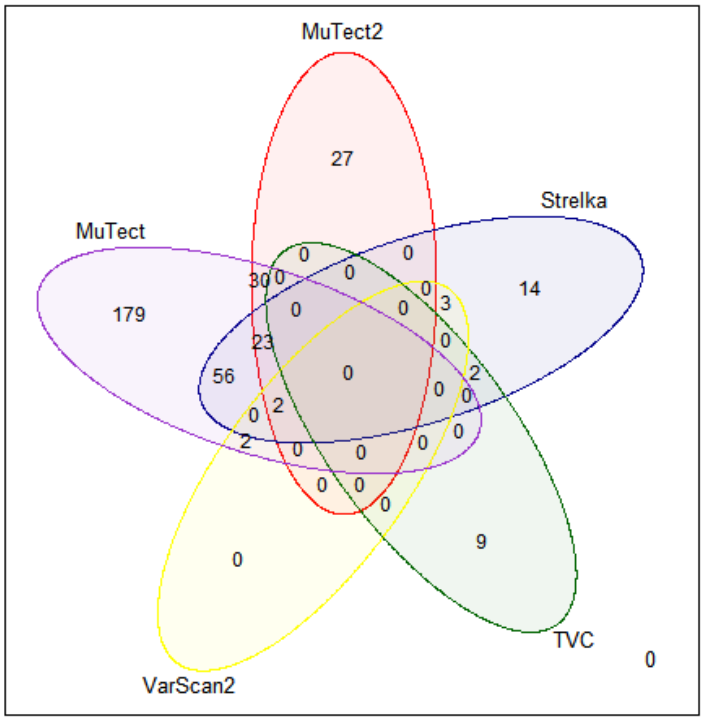


PROM 341

INDEL



SNP

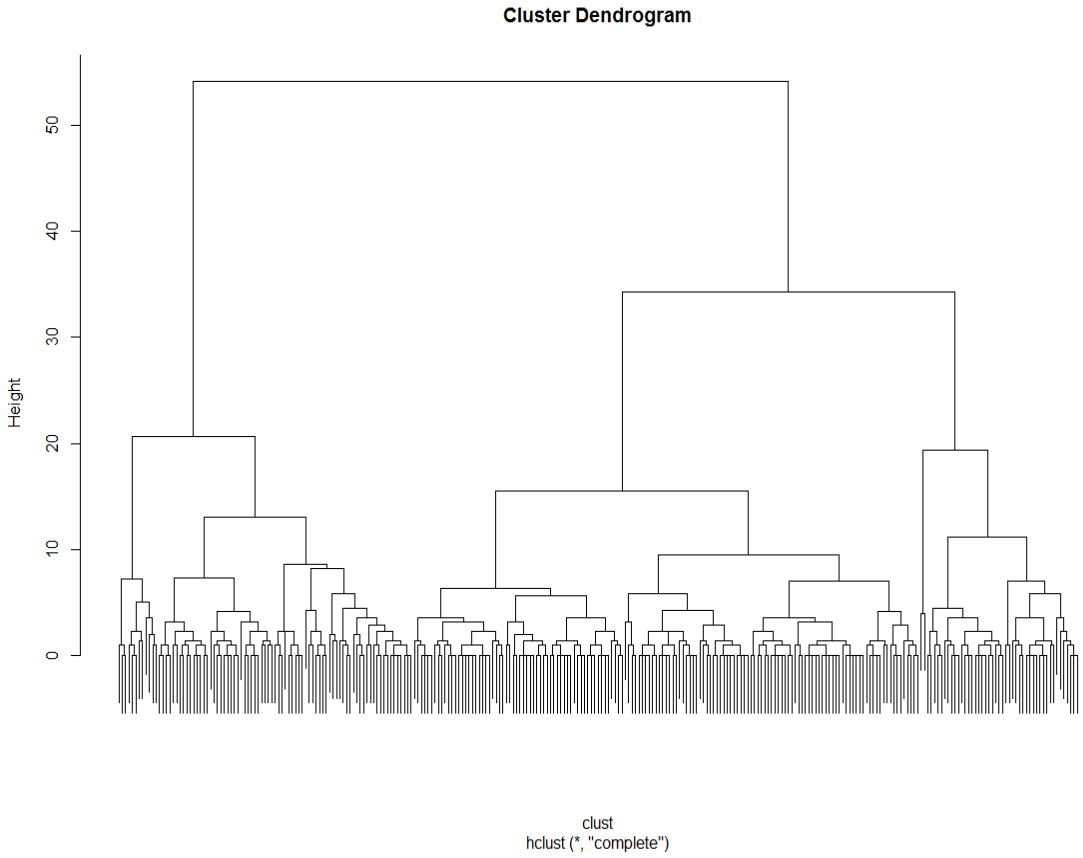


7. 4 Hierarchical clustering

The hierarchical clustering below (see page 61) includes ages and amount of variants for 283 patients with HPV^+ TSCC and HPV^+ BOTSCC.

Figure information

The y-axis denotes the Euclidean distance between cluster members. The two foremost clusters both had a mean of variants called of ~4.5, while the mean age for the leftmost cluster was 54 years of age, while the cluster furthest to the right was at ~71 years of age. These differences were not deemed significant enough to warrant further investigation in this statistical model.


Cluster Dendrogram
Height
0
10
20
30
40
50
clust
hclust (*, "complete")