Norwegian University
of Life Sciences

Master's Thesis 2023 60 stp

The Faculty of Chemistry, Biotechnology and Food science (KBM)

Analysis of mRNA in relation to forensic science

Julie Petrine Stavdal Pedersen

Bioinformatics and Applied Statistics (BIAS) - Bioinformatics

Forord

Med denne oppgaven markerer jeg slutten på mitt to år lange masterstudium innen bioinformatikk hos Norges miljø- og biovitenskaplige universitet. Det har vært krevende, men også svært lærerik og begivenhetsrik prosess.

Jeg vil først og fremst gi en stor takk til mine veiledere, Ane Elida Fonneløp og Thore Egeland, for deres tålmodighet, hjelpsomhet og inspirasjon. Deres kunnskap innen deres fagfelt er helt eksepsjonelt og inspirerende. Tusen takk for at deres tilgjengelighet, råd og veiledning gjennom hele studieløpet.

I tillegg vil jeg takke mine kollegaer på rettsmedisinsk avdeling for deres motivasjon og godhet. Tusen takk til Simon Gustafson Brueberg og Trine Isaksen for god veiledning gjennom ett tett og utfordrerne studieløp.

Vil også utrykke min takknemlighet for støtte, motivasjon og kjærlighet opp igjennom studieløpet fra min familie, kjæreste og venner.

Julie Petrine Stavdal Pedersen

Oslo, 15.07.2023

Sammendrag

Helt siden 1980-tallet, har DNA fra biologisk åstedsmateriale blitt brukt til å identifisere berørte i ulike straffesaker (1). Alle individer har et unikt genmateriale og dermed også en unik DNA-profil. Dette kan utnyttes hvis man ønsker å knytte biologiske spor til en donor slik at en person kan bli rettmessig frifunnet eller for å bringe frem nyttig informasjon til etterforskningen og eventuelt styrke mistanken mot en mistenkt. Kunnskap om ulike celletyper på et åsted, kan også gi vesentlig informasjon om handlingsforløpet til en kriminalsak. Eksempelvis kan kunnskapen om tilstedeværelsen av vaginalt sekret eller sædvæske bidra til å fremme viktige opplysninger innen ulike voldtektssaker (2).

Hos rettsmedisinske laboratorier, er det vanlig å utføre forundersøkelser som gir en indikasjon på om en spesifikk celletype/kroppsvæske er til stede eller ikke. Disse testene er vanligvis basert på kjemisk eller enzymatisk fargeendringer, men er dessverre ikke humanspesifikke og kan være vanskelige å tolke ved små mengder cellemateriale. Derfor må vi forvente at tester som inkluderer fargeendring vil være beheftet med en viss usikkerhet (3).

RNA, ribonukleinsyre, er et molekyl som blant annet er involvert i overføringen av genetisk informasjon fra DNA til proteinsyntese. RNA inneholder derfor viktig informasjon om genuttrykket til ulike celler. Ved å se på gener som er utelukkende utrykt hos spesifikke celler, kan man bruke genene til å identifisere ulike celletyper (4).

I løpet av de siste årene har nye celleidentifikasjonsmetoder blitt introdusert hos rettsmedisinske laboratorier. En av disse metodene inkluderer vevsspesifikke RNA-markører som kan brukes til å lage RNA-profiler. Ved å se på unike kombinasjoner av RNA-markører til en RNA-profil, kan man identifisere flere humanspesifikke celletyper slik som blod, spytt, sæd, vaginalsekret, menstruasjon og hud (5).

Til tross for at en RNA-profil kan bidra til verdifull tilleggsinformasjon om en straffesak, er metoden enda ikke inkorporert i rettsgenetisk saksbehandling. En av årsakene til dette er at RNA-profiler har flere elementer som må tolkes på en annen måte enn hos DNA profiler (5). I tillegg finnes det enkelte markører som er mindre spesifikke som har vist seg å kryssreagere med andre celletyper. RNA-markøren MUC4, en humanspesifikk markør som koder for mucin-4, er for eksempel utelukkende utrykt i vaginalsekret, men har vist seg å bli detektert i både spytt og nesesekret (6).

I denne studien skal vi finne ut om vi kan bruke RNA-profiler til å lage en statistisk modell for celletypebestemmelse. Spesielt da med hensyn på prediksjon av ulike humanspesifikke celletyper i rettsgenetisk sammenheng. Analysen er basert på eksisterende data som er hentet fra avdeling for rettsmedisinske fag hos Oslo universitetssykehus (RESP-OUS). Vi kommer til å gå innom hvilke utfordringer vi kan støte på ved på ved predikering ved hjelp av RNA. For eksempel, kan uspesifikke markører slik som MUC4 bidra til dårlige prediksjon.

Abstract

Ever since the 1980s, DNA from biological crime scene material has been used to connect suspects to crime scenes and victims that are affected in various criminal cases (1). All individuals have a unique genetic material that can be used to create a unique DNA-profile. A DNA-profile is an efficient way to link biological traces to a donor so that a person can be rightfully acquitted or to assess the investigation further and possibly strengthen the suspicion against the suspect. Knowledge about the different cell types at a crime scene, can also provide important information about the course of action from the event that occurred. For example, can the presence of vaginal secretion or seminal fluid give valuable information in various rape cases (2).

In forensic laboratories, it is common to perform presumptive tests that confirm whether a specific cell type/body fluid is present or not. These tests are usually based on chemical or enzymatic color changes but are unfortunately not human-specific and can be hard to interpret if there are small amounts target material. Therefore, we must expect that some tests must be confirmed with a certain uncertainty (3).

RNA, ribonucleic acid, is a molecule involved in transferring genetic information from DNA to protein synthesis. This means that RNA contains valuable information about the gene expression in various cells. By looking at genes that are exclusively expressed in specific cells, we can use the knowledge to identify different cell types (4).

Today, several new cell identification methods have been introduced in forensic laboratories. One of these cell typing methods include tissue specific RNA markers that can be used to create an RNA profile. By looking at the unique combinations of RNA markers for an RNA profile, one can identify several human-specific cell types such as blood, saliva, semen, vaginal secretions, menstruation and skin (5).

Even though an RNA profile can contribute to valuable additional information about a criminal case, the method has not yet been widely incorporated into forensic genetics case management. One of the reasons for this is because an RNA profile has several factors that must be interpreted in a different way than DNA profiles (5). In addition, there are certain markers that have been shown to be less specific and can cross-react with other cell types. The RNA marker MUC4, a human-specific marker that codes for mucin-4, is for example,

exclusively expressed in vaginal secretions, but has been shown to be detected in both saliva and nasal secretions (6).

In this study, we will find out if we can use RNA profiles to create a statistical model for cell type prediction. We will specifically focus on prediction of human specific cell types and body fluids related to forensic casework. The data used in this analysis is based on existing data acquired from the Department of Forensic Medicine at the University Hospital in Oslo (RESP-OUS). The data were used in a statistical model for cell type determination. In the study we will mention some of the challenges that we may encounter while predicting different cell types/body fluids with RNA profiles. For example, non-specific markers such as MUC4 may contribute to poor prediction.

Innholdsliste

1. Introduction

Forensic science is the application of scientific techniques and principles that is used to provide important evidence or context to legal investigations. The first applications in forensic science, can be traced all the way back to ancient Chinese and Roman societies. Antistius, from the ancient Rome, did for example perform the first autopsy on Julius Caesar around 44 B.C.E. While a jurist from China, Sung Tzhu, wrote one of the greatest works within forensic medicine: the “Hsi Yuan Chi Lu” (The Washing Away of all Wrongs) in AD 1247 (7) (8).

During the industrial revolution and the early nineteenth century, new scientific methods within the forensic field were introduced. Methods such as: fingerprint analysis, UV spectrophotometer for detection of organic material and the comparison microscope for compering microscopic patterns (7).

However, it was not until 1985 that DNA profiling technique was initially introduced in a laboratory by the British geneticist, Sir Alec Jeffery (9). DNA profiling stands as one of the most important advancements in forensic science, allowing the comparison of DNA samples from different sources to determine a person’s genetic makeup. In 1987, the DNA profiling technique was first used in a police forensic test, where it played a crucial role in confirming the veracity of a suspect's confession and securing the conviction of a murderer involved in the rape and murder of two teenagers (10). The first criminal case that used the technique in Norway was in 1988 and was also regarding a rape-murder case (11).

Today many severe criminal cases can be solved or aided with the help of DNA profiling. The technique is a critical tool in forensic investigation since it can be used to identify individuals related to evidence in a criminal investigation or exonerate innocent individuals. A DNA profile is based on a unique genetic signature that comes from a small variation in the genome among individuals. This means that every individual, except identical twins, has their own specific profile that can be used to link biological samples to a donor. (12)

The aim of forensic studies is to aid and assist the court in reaching a valid conclusion or judgement of a criminal investigation. Knowledge about the identity of the donor is unfortunately not always enough evidence to provide a proper conclusion in court. In some cases, the question is more centered around how the evidentiary traces got deposited, rather than the individual that deposited it.

As an example, can the prosecution scenario (H1) differ from the alternative scenario (H2) based on how the cellular material, got deposited. The cellular material from the suspects can, for example, be found on the murder weapon due to direct contact between the suspect and the weapon after the suspect used the weapon to murder the victim (H1). Or as alternative scenarios (H2): the cellular material has been deposited after the murder was committed, the suspect could have been in contact with the weapon before the murder took place or the findings could have been transferred from the suspect to the actual killer due to secondary transfer. (13)

In some cases, is it important to know the localization and the type of body fluid/tissue that were deposited. Some body fluids are more likely to be localized at certain areas than others. For example, we do not expect to find vaginal discharge on hands or certain objects. The identification and localization of the cellular material can therefore be a critical aspect to be considered in rape cases or other cases where the action of the activity is uncertain.

Unfortunately, a DNA profile gives little information about the timing, activity nor cellular origin of the dispositioned biological fluid/tissue that might be crucial information for further investigation.

There are fortunately several methods that are used for identification of body fluids and different tissues. Some of the most common body fluid/tissue identifying methods in forensic science are based on chemical and enzymatic assays. These tests are often called "presumptive" and are used to verify the presence of a specific compound, usually by enzymatic color changing reactions. An acid phosphatase (AP) test, is an example of a standard chemical reaction used in forensic labs to detect the presence of an enzyme that is present in sperm cells (13).

However, none of these conventional tests are error free. Occurrences of false positives or false negatives are not uncommon (14). As these rapid stain identification methods (RSID) are not performed simultaneously and often require a significant amount of cellular material one can lose both time and important evidence while performing these tests. Another methodological problem one might encounter is cross reactivity among species and the lack of sensitivity and specificity towards body fluid and tissue identification (15).

An RNA profile can give information about the cellular origin based on their gene expression pattern. Different cells in the body produce a specific combination of mRNA molecules that are correlated to the cell's unique protein production. We can therefore create a tissue specific assay based on the expression of mRNA of the cell to identify different body fluids and tissues. While RNA profiling is not yet widely used in forensic investigations, it holds promise as a complementary tool to DNA profiling. This potential results from its ability to provide additional information that DNA profiling alone cannot offer, making it a valuable tool among forensic methods. (5)

1.1 RNA

1.1.1 Structure and Function

Ribonucleic acid (RNA) works as the intermediate product of protein biosynthesis and share a lot of similarities with deoxyribonucleic acid (DNA). Both nucleic acids can be found in all living organisms and are built up of the intracellular organic molecule: nucleotides (4). A nucleotide consists of a pentose (deoxyribose in DNA and ribose in RNA) and a nitrogenous base that share a covalent bond to one or more phosphates. In DNA the nitrogenous bases are Cytosine & Guanin and Thymine & Adenine. The nitrogenous bases are joined together in pairs by a hydrogen bond on each side of a double stranded helix. In RNA, the Thymine is substituted with the pyrimidine Uracil (4).

Transcription is a cellular process where a segment of DNA is copied into a messenger RNA (mRNA). This mRNA contains genetic information essential for protein synthesis within the cell. The mRNA acts as a messenger, transporting the DNA information from the cell's nucleus to the cytoplasm, where proteins are produced.

Even though RNA shares a lot of structural similarities with DNA, its function is limited due to its relatively short half-life. RNA is usually observed single stranded and is therefore more exposed to degrading cellular mechanisms such as RNA- degrading enzymes (RNases). The RNA can get degraded internally, by the endonuclease, and from the 5' and 3' end by the 5'exonuclease and the 3'exonuclease (18).

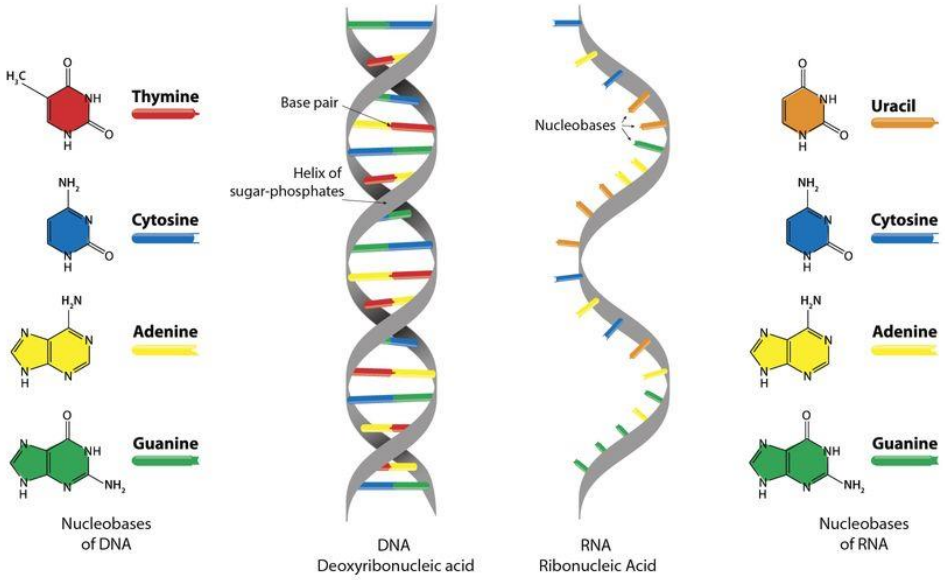


Figure 1.1: *Taken from www.technologynetworks.com/ /genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719 (retrieved 02.03.23)*

There are many different types of RNAs that is either directly involved in the biosynthesis of proteins or play a part in the regulatory network of gene expression. Among the total amount of all RNA transcripts, the transcriptome, most of the RNA are non-coding RNA (ncRNA). The ncRNA does not encode for proteins but regulates cellular functions and physiology (19). On one hand, we have: MicroRNAs, long-noncoding RNAs and circular RNAs that are examples of ncRNAs that are involved in transcription regulation and expression. On the other hand, we have the messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA), that plays an important part in the protein biosynthesis. (4)

1.1.2 Protein Production

mRNA is a product of the first steps of the protein biosynthesis: the transcription. This is where a sequence of a DNA is transcribed into a copy of the genome sequence. The copy of the genome is called the mRNA which is further transformed into proteins due to the translation. Translation is performed in ribosomes that is built up of ribosomal RNA: rRNA. Transfer RNA (tRNA) is the link between a specific codon in the mRNA sequence that corresponds to an amino acid. The tRNA is a central key to the translation proses where mRNA is transformed into an unprocessed polypeptide. (4)

mRNA is produced during the first step of protein biosynthesis, known as transcription. During transcription, a DNA sequence is copied to a shorter mRNA sequence. The mRNA is then used for protein synthesis through a process called translation. Translation takes place in ribosomes, which are made up of ribosomal RNA (rRNA). Transfer RNA (tRNA) plays a crucial role in translation by connecting specific codons in the mRNA sequence to corresponding amino acids. This makes tRNA a central component of the translation process, converting mRNA into an unprocessed polypeptide. The flow of genetic information from DNA to RNA to protein is called the central dogma. (4)

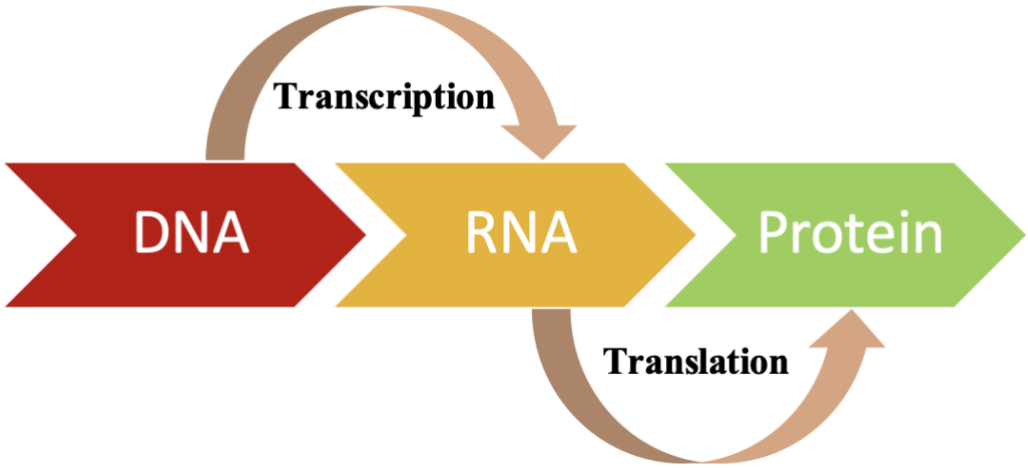


Figure 1.2: *Figure showing the flow of genetic information from DNA to RNA to proteins. The transformation from DNA to RNA is done through transcription. The transformation from RNA to proteins is done through translation. Figure 1.2 is created in word.*

The number of proteins produced in a cell is directly related to the levels of mRNA present in the cell. These mRNA levels can vary based on different phenotypic traits that influence protein production and cell function. Different cells have unique protein patterns due to variations in gene activity, which can also differ between species and genomes. (20)

1.2 Biomarkers

Biomarkers, also known as signature molecules, are measurable indicators of a biological process that can be found in a variety of sources including blood, urine, tissue, and other bodily fluids. A biomarker is a broad term that is primarily used as a medical expression for a measurement that is used to detect a potential hazard in the body (21). These measurements are linked to special characteristic of body functions that are either biological, pathological, or pharmacologic and may give valuable information about cell abnormalities, diseases, or cell

identity. Examples of biomarkers can be anything from: blood pressure, pulse to the level of gene expression within cells (22).

1.2.1 Different types of Biomarkers

In current research, a wide range of biomarkers has been identified and used for diverse applications. Among these biomarkers, genomic, epigenetic, and metabolic markers are notable examples used to detect and analyze various physiological abnormalities within the body.

Genetic biomarkers are based on a distinctive individual's genetic information that can be associated with mutations or enhanced/suppressed genes that may increase or decrease the production of certain proteins. Examples of genetic biomarkers include BRCA1 and BRCA2 mutations that are associated with increased risk for breast cancer (23).

Epigenetic biomarkers are proteins that can influence gene expression without changing the underlying DNA sequence. DNA methylation is an example of this, where methyl molecules get added to a specific region of the DNA and in the prosses alter the gene expression. This can affect how genes are turned on or off, affecting various cellular processes. (24)

In contrast to genetic and the epigenetic biomarkers, metabolic and microbial biomarkers are not based on the genetic information but are rather based on the presence of small molecules and microbiota in body fluids or tissues. For example, high glucose levels can be a molecular biomarker for the detection of diabetes in blood. (25)(26)

1.2.2 Biomarkers in forensic science

In forensic science, biomarkers can be used as valuable tools for victim/suspect identification and provide crucial insights to the circumstances surrounding the crime and the crime scene. These biomarkers are often based on genetic markers.

One of the most used biomarkers in forensic science is DNA. DNA can be extracted from a variety of biological samples, including blood, skin cells, and other bodily fluids. A DNA analysis can give additional information about a person's identity and other additional information such as their sex and ancestry.

In forensic genetics, messenger RNAs (mRNAs) have gotten growing interest due to their ability to distinguish body fluids from other tissues of forensic significance (3). Different cell types and body fluids express a various amount of mRNA due to the need of specific proteins associated with the cell type and its functions. Blood cells, for example, produce the characteristic protein hemoglobin, that contributes to the transport of oxygen from the respiratory system to other peripheral tissues in the body.

The gene *HBB* encodes for the subunit beta-globin protein and can therefore be used as a biomarker for identification of venous blood cells. Another biomarker for blood identification is the gene CD93. CD93 encodes for a protein component that is connected to a larger receptor complex called C1q. This receptor is a cell surface glycoprotein that may play a significant function in cell–cell adhesion and removal of apoptotic cells (27).

The ALAS2 gene is also specifically expressed in blood cells. This gene contributes to the synthesis of the enzyme 5'-aminolevulinate synthase 2, that plays an essential role in the early stages of the enzymatic pathway of the heme-synthesis. This gene can therefore be used as a biomarker for identification of blood (28).

HTN3 or Histatin 3 is a protein in the histatin - family and is encoded by the HTN3 gene. Histatins are salivary proteins found on tooth surfaces and in saliva and can therefore be used to distinguish saliva from other body fluid types (29)(30). Another gene that is specific to saliva, is the STATH gene which encodes for a satherin protein produced in the saliva glands. This protein prevents precipitation of calcium phosphate in saliva and contributes to the ossification in the cavity (31).

Nasal secretion is regarded as the most difficult body fluid to distinguish because of its shared similarities with saliva or virginal secretion. The satherin protein has been reported to be strongly expressed in nasal secretion and in some vaginal secretion samples despite only being modestly expressed in saliva. The histatin mRNA, seems to not be expressed in neither nasal nor vaginal secretion and is therefore more specific than the STATH marker (30).

The protein-coding gene, BPIFA1, encodes for a lipid binding protein that has antibacterial activity against Gram-negative bacteria. The antimicrobial protein Plays an important role in the activation of an airway immune response in the nasopharyngeal regions and the upper airways (32). This gene is therefore one of the biomarkers associated with nasal secretion identification.

The presence of semen can be determined by examining the gene expression of: SEMG1, PRM1, and KLK3. SEMG1 encodes for Semenogelin 1, a protein predominantly found in semen. This protein is expressed in basal prostatic cells and prostatic glandular cells in the seminal vesicle, making it a vital marker for tissue/cell identification within forensic science. Additionally, Semenogelin 1, SEMG1, plays a crucial role in the formation of a gel matrix that encapsulates ejaculated spermatozoa (33).

The PRM1 gene is responsible for encoding the protein protamine 1, which is synthesized in the testis and serves as a replacement for histones during the haploid phase of spermatogenesis (34). KLK3, on the other hand, encodes for Kallikrein Related Peptidase 3, a single chain glycoprotein synthesized in the prostate gland. This protein fulfills various physiological functions, including the liquefaction of the seminal coagulum, and is present in the seminal plasma (35).

Differentiation between blood and menstruation blood can be crucial in criminal investigation, especially in rape cases where the presence of menstruation blood can indicate sexual activity. The matrix metalloproteinases: MMP7, MMP10 and MMP11 have proven to be reliable markers for differentiation between blood and menstruation blood with MMP11 as the most specific and reliable marker (36). Matrix metalloproteinases are proteins that can degrade extracellular matrix (ECM) which is especially important during the angiogenesis and embryonic development (37).

Similarly to nasal secretion, vaginal secretion is one of the more difficult body fluids to identify. This is due to cross reactions among some of the marker used for vaginal fluid identification, like the MUC4 gene, that shows incomplete specificity and detectability (38). Mucin 4, MUC4, is a mucin protein that is encoded by the MUC4 gene. Mucins are components of the mucus layer that overlie the intestinal epithelial cells and are important for the maintenance of the physiologic hemostasis (39). Since there are mucous membranes covering the vaginal region, this gene can be used as an RNA marker for the identification of vaginal mucosa or secretion. Other RNA markers that can be used for vaginal identification are the MYOZ1 and CYP2B7P1 genes.

MYOZ encodes for a protein in the myozenin family and is primary found in skeletal muscle. Myozinins work as intracellular binding proteins that link Z-disk proteins together (40). CYP2B7P1is a pseudogene which is a nonfunctional DNA that does not encode for a protein.

As opposed to MUC4, CYP2B7P1 is exceptionally specific with no noticeable cross reactivity with other body fluids/tissues (38).

1.2.2.1 Housekeeping genes

Housekeeping genes are genes that are expressed in all cells under normal and physiological conditions. These genes are usually essential for the survival of the cell and contribute to the maintenance of basic cell functions. In forensic science, one can use these genes as a positive control to make sure the sample contains the targeted mRNA markers. The most common housekeeping genes that are currently used in RNA and DNA profiling are the 18s rRNA gene, and the gene ACTB (41).

The 18s rRNA is one of the basic components of the eukaryotic cytoplasmic ribosomes and is therefore also one of the basic components for all eukaryotic cells. Even though, the gene marker is a widely used control for several gene expression assays, the expression of the gene can sometimes be too strong and can cause disturbance in other gene expression assays in the same reaction (42) (43).

ACTB is considered a housekeeping gene because of its major role as beta-actin in the non-muscle cytoskeletal actins. Beta actin especially important for cell mobility, structure, and integrity in the cell (44).

1.2.2.2 Sex specific genes

Other genes such as the X inactive specific transcript (XIST) and Ribosomal protein S4 (RPS4Y1), are commonly used to identify the gender of the cellular host.

XIST is a non-coding-RNA molecule that is essential for the inactivation of the X-chromosome. During the early stages of female mammal development, the XIST RNA-gene will exclusively be expressed and produce a coat that will inactivate one of the X-chromosomes. Consequently, this non-coding RNA can be used as a sex marker for female identification in forensic RNA assays (45).

RPS4Y1 is a Y-chromosome-specific marker and therefore only expressed in males. This gene has the highest average expression difference between females and males and encodes for a Y-specific ribosomal protein (46).

1.3 Multiplex PCR and profiling of RNA

All types of cells in the body have a unique combination and expression of mRNA that differ between each cell type. This makes it possible to create a tissue-specific assays based on the cell type specific genetic expression. New techniques within forensic science have made it possible to detect multiple RNA transcripts from DNA stains found in various crime scenes (47). The RNA is first co-extracted from the evidentiary sample and then amplified with reverse transcription polymerase chain reaction (RT-PCR). The detection of the expressed RNA transcripts is most frequently done by capillary electrophoresis, but some MPS panels have also been developed (48)(49).

In the next section the different techniques that were used to acquire the RNA and the values describing the expression rate of the different transcripts will be more thoroughly introduced.

1.3.1 Multiplex PCR of the RNA samples

Polymerase chain reaction (PCR) is a groundbreaking laboratory technique that is used to multiply a specific sequence of DNA. PCR allows for the amplification of billions of targeted segments using primer sequences that hybridize to the selected regions that will be amplified. The enzymatic process involves repetition of heat-cool cycles that enables enzymes to separate and replicate the two complimentary strands of the sample DNA (50).

By adding more primers to the mixture in the PCR reaction, one can copy multiple regions simultaneously in one simple PCR reaction. This is called multiplex PCR and is used to amplify multiple targets from multiple samples simultaneously. This can save both time, money and makes it easier to compare the expression of multiple genetic markers from one DNA sample (50).

Unfortunately, RNA is not suitable for the PCR technique because of its single stranded structure and its relative short half-life due to degradation. Also, RNA cannot be amplified directly because PCR requires DNA as a template for amplification and not RNA. Consequently, one must transform the RNA into its complementary DNA (cDNA) by using a technique called reverse transcription (RT). Once the cDNA has been synthesized by the enzyme reverse transcriptase, one can continue to use the standard PCR techniques to amplify the cDNA (51), (52).

RNA transcripts can be detected with RT-PCR followed by gel or capillary electrophoresis. By combining these techniques, one can detect multiple RNA transcripts at the same time and create an RNA -profile that can be used for body fluid identification.

By adding an additional fluorescence dye to the multiplex PCR primers, different amplicons can be detected and separated by emitting different fluorescence dyes during capillary electrophoresis. The targeted amplicon will get labeled by covalently binding to the dye which is attached to the primer. By having an additional fluorescence dye for each primer labeling the different targeted fractions, we can create even larger multiplexes since we then can differentiate the fragments not only by size, but also color of the dye. (50)

1.3.2 Capillary Electrophoresis

Capillary electrophoresis (CE) is an analytic separation method that detects and separates DNA fragments based on their electrophoretic mobility and size (53). The migration speed of the fragment correlates with its size, making it possible to estimate the length of the fragment using CE. Furthermore, CE can utilize fluorescence emission to quantify the fragments, in addition to its detect and estimate the size of the fragment.

Following the amplification process, the DNA samples undergo dilution and fluorescence labeling before being injected into the capillary. Typically, the samples are diluted with high-quality deionized formamide at varying ratios (1 to 10), which ensures low conductivity. During the DNA amplification, primers with distinct fluorescent dyes are used to label the DNA fragments. This approach makes it possible to detect and separate of different amplicons based on their emitted fluorescence from the respective dyes (53).

Migration of the fragments happens when a high voltage is applied to a cathode and an anode on each side of the thin capillary. The fragments from the DNA sample will then start to migrate from the negative cathode to the positive cathode. During the migration, the different fragments will be separated by size from smallest. This size-based separation is facilitated by a polymer solution acting as a sieving medium inside the capillary.

The capillary of the CE instrument is approximately 50cm long and 50μm in diameter and has a negative charged interior wall. The longer the capillary, the better separation of fragments and better resolution. However, a long capillary can also hold a disadvantage due to the increased running time for each sample (53).

Capillary electrophoresis has a high efficiency and automation due to the thin capillaries and a high voltage supply. Also, it only requires 0.1- 10 μL sample which is really low compared to other separation methods.

A specific wavelength/light occurs as a laser excites the dye of the fluorescence labeled amplicon at the end of the capillary. An optical filter detects and separates the emitted light that is later separated by a multi wavelength analyzer that captures the wavelength of the light through a detection window that is often located at the end of a capillary in the CE. The intensity of the light is correlated with the quantity of the dye that is absorbed. A detector then records the fluorescence signature of each fragment as well as the time the fragment use to migrate through the polymer. The raw data are then sent to a computer program and later analyzed in a computer software. (53)

2. Methods

This study is based on an already available dataset that was obtained during the validation of mRNA analysis at the Department of Forensic biology at the University Hospital in Oslo (RESP). The dataset consists of 38 different samples, taken from six different body fluids and tissue types: blood, vaginal secretion, nose secretion, menstrual blood, semen, and salvia. All samples have been collected and processed through a set of different methods like extraction, separation and detection methods which resulted in 90 unique RNA-profiles, one for each of the collected samples.

The RNA- profiles are based on 19 different markers that are either related to a specific body fluid/tissue, a housekeeping gene, or a gender specific mRNA marker (sex-marker). For each marker that was detected, there will be a corresponding fluorescence intensity (RFU) – value, that represents the presence and quantity of the gene that is expressed from the biological sample.

One of these 90 RNA profiles had to be removed due to lack of detection among the body fluid specific mRNA markers including the sex specific markers. In the main parts of this study, we will exclude the housekeeping genes from the analysis, as they were found to be present in all the samples, and their inclusion would not provide additional discriminatory power.

The participants donated one of the six different body fluids, providing maximum one sample per body fluid/tissue type. The samples were then placed in an extraction tube where the genetic material was co-extracted to provide one RNA and one DNA fraction. DNAase treatment was then performed on the RNA fraction before the reverse transcription of the RNA.

The reverse transcription is a necessary step in the process because PCR, which is used to amplify the fragments, cannot use the single stranded RNA as a template. Reverse transcriptase transforms the single stranded RNA to cDNA making PCR and amplification possible. The targeted and amplified cDNA fragments were then separated by size with capillary electrophoresis (CE). The data collected from the CE was processed in the GeneMapper® software version 1.6 (Applied Biosystems™) and further analyzed in the R-software software version 4.1.3 for data and statistical analysis. The workflow of the study is presented in *figure 2.1*.

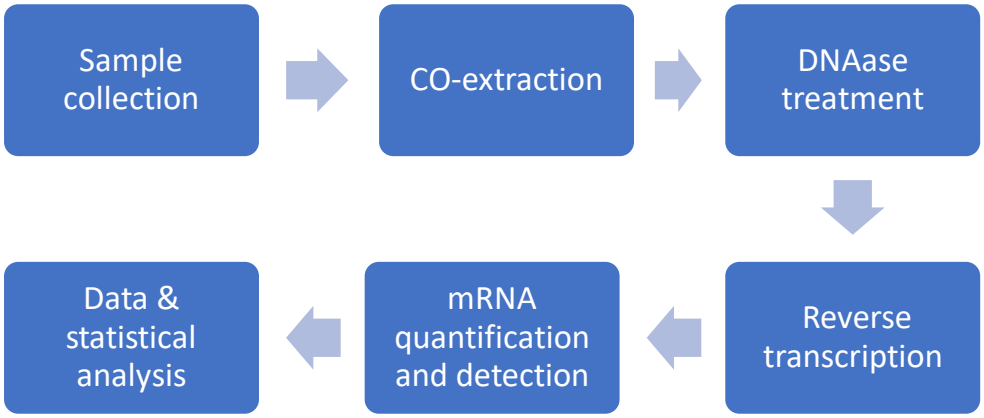


Figure 2.3: *An overview of the workflow of the study.*

2.1 Sample Collection

The samples used in this study were collected from voluntary participants from the Department of Forensic biology. All samples were taken after informed consent and approved for further data analysis. All samples in this study are and will be kept anonymous.

There are in total 38 samples from the experimental design from the RESP department. The distribution of samples among the various sample types are described in *Table 2.1*.

Table 2.1*: Summary of the collected samples.*

Sample Type	**Number of samples taken from the RESP department at OUS**
Blood	5
Menstrual Blood	3
Nasal Mucosa	4
Semen	6
Saliva	16
Vaginal	4
sum	**38**

The participants, both female and male, were given clear instructions on how to collect the specific samples. All samples, except blood, were collected by cotton swabs and carefully placed directly in labeled sample bags. Twenty μL of blood were taken from the participant's finger and absorbed into a cotton swab. The cotton swabs were then cut with sterile scissors and collected in an extraction tube under clean conditions to avoid contamination from external genetic materials. All samples were then co-extracted and amplified using a multiplex PCR before an RNA-profile was created.

Twenty-eight samples were amplified using 0.5, 1.0 and 3.0 μL cDNA input. The remaining 10 samples were amplified using only 1.0 μL input. This gave a total of 94 RNA -profiles, where four of the profiles, two of the semen samples and two of the saliva samples, were removed due to undetected values of the ACTB and/or 18S-rRNA marker (housekeeping genes). One saliva sample with volume 0.5 was also removed from the study due to lack of detection among the body fluid specific mRNA markers including the sex specific markers. This gave us a total of 89 RNA-profiles from the six different body tissue/fluids. (See table 2.2 and 2.3 for a summary of the distribution of RNA-profiles).

***Table 2.2**: The total amount of RNA profiles from each body fluid type.*

Sample Type	**Total Number of samples**
Blood	11
Menstrual Blood	9
Nose secretion	12
Sperm	12
Saliva	33
Vaginal Secretion	12
Total	89

***Table 2.3**: A summary of the total 89 RNA profiles divided into their representative PCR volume.*

Sample Type	**Volume of PCR**	**Number of Samples**
Blood	0.5	3
Blood	1.0	5
Blood	3.0	3
Menstrual Blood	0.5	3
Menstrual Blood	1.0	3
Menstrual Blood	3.0	3
Nose secretion	0.5	4
Nose secretion	1.0	4
Nose secretion	3.0	4
Semen	0.5	3
Semen	1.0	5
Semen	3.0	4
Salvia	0.5	9
Salvia	1.0	14
Salvia	3.0	10
Vaginal Secretion	0.5	4
Vaginal Secretion	1.0	4
Vaginal Secretion	3.0	4
Total		**89**

2.2 Co-extraction

The co-extraction of the collected samples was done using a phenol-chloroform method with the QIAamp DNA mini kit (QIAGEN) and mirVANA™ miRNA isolation kit. The applied method was retrieved from the paper by Lindenbergh et al. and by Johannessen et al. (3)(4).

2.3 DNAase Treatment

DNAase treatment was performed on the extracted RNA fractions with the TURBO DNA-free™ Kit. The procedure followed the manufacturer's protocol by Invitrogen from Thermo Fisher Scientific: (Invitrogen by Thermo Fisher Scientific) (54).

2.4 Reverse transcription

The transformation of the RNA to the complementary DNA (cDNA) was performed using the SuperScript® IV Reverse Transcriptase (invitrogen by Thermo Fisher Scientific). The transformation was performed in the same manner as described by Johannessen et al (6).

2.5 RNA Quantification and Detection

The amplification of the targeted RNA sequences was done using the QIAGEN multiplex PCR Kit (QIAGEN) with an RNA 19-plex that is optimized by the NFI. There was a total of 19 mRNA markers that were used in the 19-plex primer mix. All 19 markers and its primer sequence are described in table 2.4. Each mRNA marker was labeled with a dye and the fragments were later separated by size through the capillary electrophoresis.

***Table 2.4**: This table is collected from the scientific study by Helen Johannessen and described by Van den Berge et al. (2017). Two of the markers from the paper are replaced by two gender specific markers (XIST and RPS4Y1) as described in (4). The name of each mRNA marker and its corresponding body fluid are listed in the two left columns whereas the primer sequence and the length of the mRNA marker sequence are listed in the two right columns.*

Body fluid	mRNA marker	Primer sequence (5' → 3')	Size (bp)
Blood	HBB	fw: GCACGTGGATCCTGAGAACTTCAG rv: ATGGGCCAGCACACAGACCAG	61
	ALAS2	fw: TTCTGCACCAGAAGGACTCAGCC rv: TAAATCTCGCACCCTGGCAGGATC	103
	CD93	fw: GCTCTGGGGCTACTGGTCTATC rv: TCCCAGGTGTCGGACTGTACTG	151
Saliva	HTN3	fw: CTTCACTTCAGCTTCACTGACTTCTG rv: CTTTGCATGTGAATCAGCTCCAGTC	132
	STATH	fw: TTCATCTTGGCTCTCATGGTTTCCATG rv: GCCATACCCATAACCGAATCTTCCA	93
Semen	SEMG1	fw: GGAAGATGACAGTGATCGT rv: CAACTGACACCTTGATATTGG	121
	PRM1	fw: GAGAGCCATGAGGTGCTGCC rv: AGGCAGGAGTTTGGTGGATGTGC	90
	KLK3	fw: GACGTGGATTGGTGCTGCACC rv: CTTCTCGCACTCCCAGCCTC	64
Vaginal mucosa	MUC4	fw: CTGCTACAATCAAGGCCACTGCTAC rv: AAGGGAAGTTCTAGGTTGACAGTTGG	141
	MYOZ1	fw: CGTGTTCTCCGGTCACAGCAG rv: TGGATTCAGCCGGCTGCTCG	88
	CYP2B7P1	fw: CCTCATGTCGCAGAGAGAGTCTAC rv: CCCATGGGGAGAAGGTCAGCA	146
Menstrual secretion	MMP7	fw: GAACAGGCTCAGGACTATCTC rv: TTAACATTCCAGTTATAGGTAGGCC	127
	MMP10	fw: GCATCTTGCATTCCTTGTGCTGTTG rv: GGTATTGCTGGGCAAGATCCTTGTT	107
	MMP11	fw: CAACCGACAGAAGAGGTTCG rv: GAACCGAAGGATCCTGTAGG	76
Nasal mucosa	BPIFA1	fw: CAAGTGAATACGCCCCTGGTCG rv: GAATGGGTGCAGTCACCAAGGAC	131
Male	RPS4Y1	fw: TGGAAGAGGCAAAGTACAAGTTGTGC rv: GGATTCCCTTCACTCCCACAGTAAT	63
Female	XIST	fw: ATTTTAACTGATCCCATTGAAGATACCACGC rv: TCAGAATGTCCAAGAGGAGCCTAAGG	83
Housekeeping genes	ACTB	fw: CAGAGCCTCGCCTTTGCCGAT rv: CGCGGCGATATCATCATCCATGGT	75
	18S-rRNA	fw: GACTCAACACGGGAAACCTCACC rv: CTCCACCAACTAAGAACGGCCATG	110

The ionic PCR products were separated by size using the 3500xL Genetic Analyzer (3500 Series instrument) (Applied Biosystems™). The dyed fragments were separated through electrokinetic migration through the thin duct of the capillary electrophoresis (CE). The fragments then passed a laser beam near the end of the capillary, which caused the attached dyes of the fragments to emit different fluorescence light that were further separated by the diffraction system and detected by a CCD camera. The limit of detection (LOD) was set at 50 RFU for each allele. The signal from the fragments were then transformed into digital data and analyzed using the GeneMapper® software version 1.6 (Applied Biosystems™).

2.6 Data analysis

2.6.1 Data Collection

The data collected from the Thermo Fisher genetic analyzer, were further sized and genotyped by the GeneMapper™ ID-X Software v1.6 and Peak Scanner™ Software. The peak scanner™ software is a nucleic acid scaling program that performs preliminary sizing and detects peaks and fragment sizes from specific Capillary electrophoresis assays (55). The GeneMapper® software performs an analysis on fsa- files that are a product of a data collection software for fragment analysis. The product of this analysis is a fitting RNA profile containing peaks that represent the detected PCR fragments.

Each peak has an estimated peak-height and size (bp) that is based on the quantity and length of the PCR fragments. The height of the peak provides a good estimate of quantity because it is based on the intensity of the emitted light/relative fluorescence units (RFU). The threshold for a peak to be detected was set to 50 RFU, any peak with a value below this is regarded as noise.

The fragment's length is measured and compared to an internal size standard that contains artificial DNA fragments of known size. The GeneMapper software contains two files: the bin-file and the panels-file, used to recognize peaks that have a value inside a particular range. The panel-file describes the sizes (bp) of the different markers, while the bin-file contains the minimum/maximum values a marker can have to be inside the panel region. If a marker has a size (bp) outside the panel area, it will be labeled: “Outside Marker Region” (“OMR”). However, if the value lies inside the panel area, but outside the bin area, it will be labeled “Off ladder” (“OL”). These alleles must be checked to see how large the deviation (bp) is

from the expected value, if they are not considered reliable, they are therefore removed from the RNA-profile.

All values that lie inside the panel region of a specific mRNA marker, will be labeled accordingly to the representative body fluid/tissue type of that mRNA marker. Figure 2.2 shows a portion of an unprocessed RNA profile. We can see from the figure that the STATH marker is detected with a bp of 7804 and the housekeeping genes, ACTB and 18S-rRNA, are detected with a bp of 15210 and 1973 respectively. There is also a detected marker, in the top window, that is an actual value from the MUC4 bio marker but is marked as “OL” since it has a value outside the threshold/bin area. The “OL” value will in this case be renamed “MUC4” and taken into the RNA-profile. The values detected outside the panel area, the “OMR” values, will be deleted and removed in the processed RNA profile.

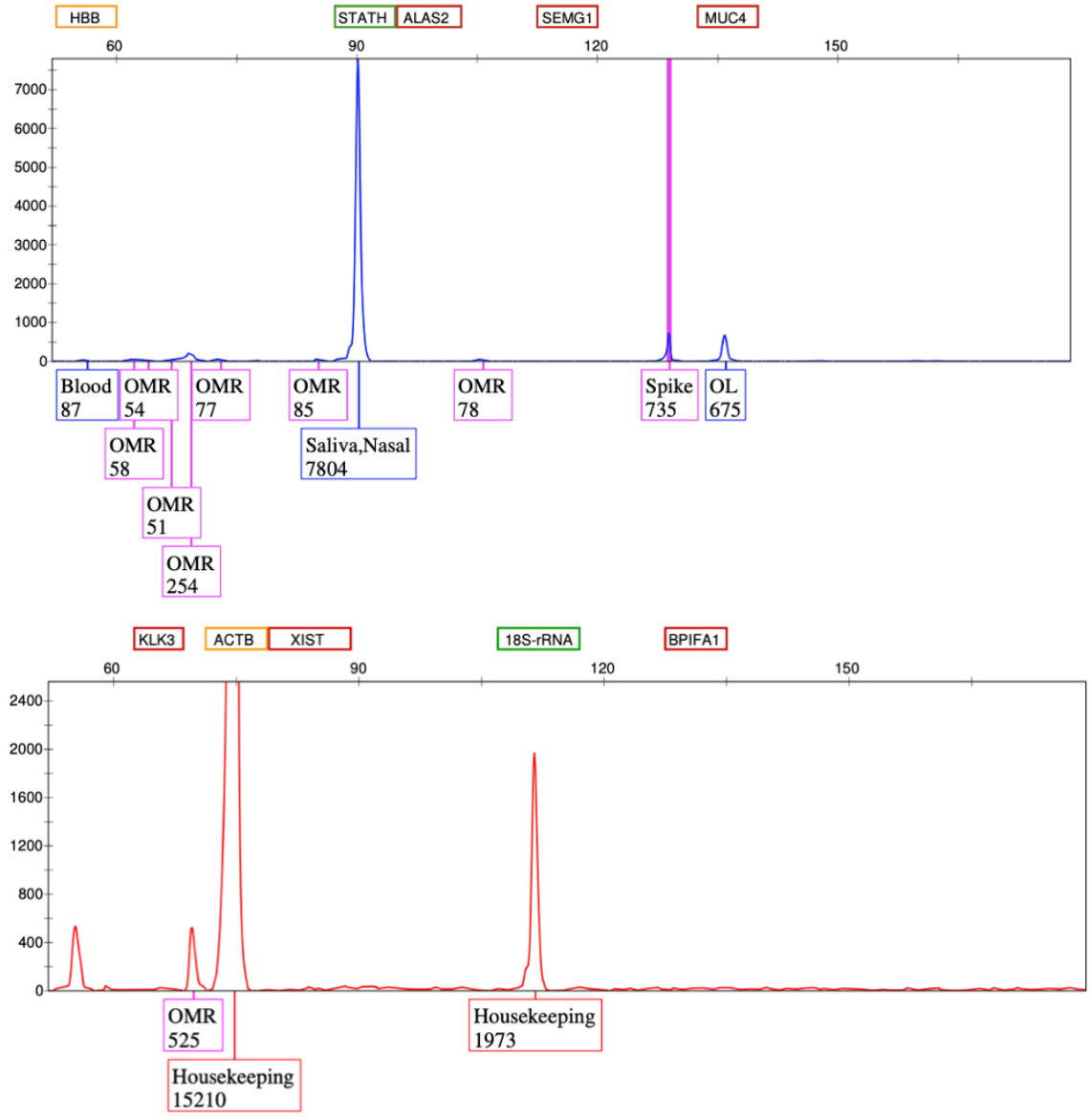


Figure 2.4: *An example of a part of an RNA profile. The x-axis is the projected peak height whereas the y-axis is the estimated size of the PCR fragment (bp). The size increases from left to right along the x-axis. In this study we removed the peaks that were too small and the peaks that were outside the allele ladder (OL) or the marker region (OMR). We were left with an RNA profile containing only reliable peaks that were labeled accordingly to the appropriate body fluid/tissue type. Above the frame, you can see the panels that is named after their representative mRNA marker.*

After the removal of the unreliable peaks, the RNA-profile, detected markers and corresponding peak heights, were converted, and saved as an excel-file in the Microsoft® Excel software (version 16.72) that is listed in Appendix 1. The excel file was then uploaded to the R software and run through the RStudio software version 2022.07.2. All the codes that were performed in the RStudio and the R-library for data analysis and statistical analysis, can be found in Appendix 1.

2.6.2 Data transformation

2.6.2.1 Detection rates of markers

We calculated the detection rates of all 19 markers in the datafile by dividing two matrices. The first matrix contained the number of markers that were detected, and the second matrix contained the total amount of markers that could have been detected, but not necessarily got detected. This allowed us to determine the detection rates for each marker. An illustration of the process is explained in figure 2.3. A table that contains the detection rate for each marker in each body fluid can be found in table 3.2.

Finding the detection rate of each marker made it easier to detect abnormalities in the RNA-profiles. Examples of abnormalities that could show up are: unexpected mRNA markers in body fluid samples, undetected housekeeping genes or wrongly detected sex markers in sex specific body fluids such as: semen, vaginal secretion, or menstruation blood. Finding these irregularities early, is important because it could influence the statistical analysis later. The code that created Table 3.2 can be found in Appendix 1.

Matrix with the number of detected markers

	Marker1	Marker2	Marker3
Sample 1	m_1s_1	m_2s_1	m_3s_1
Sample 2	m_1s_2	m_2s_2	m_3s_2

Matrix with the total number of markers

	Marker1	Marker2	Marker3
Sample 1	M_1S_1	M_2S_1	M_3S_1
Sample 2	M_1S_2	M_2S_2	M_3S_2

Detection rate of markers

	Marker1	Marker2	Marker3
Sample 1	m_1s_1/ M_1S_1	m_2s_1/ M_2S_1	m_3s_1/ M_3S_1
Sample 2	m_1s_2/ M_1S_2	m_2s_2/ M_2S_2	m_3s_2/ M_3S_2

Figure 2.3: *The numerical matrix containing the number of detected markers for each body fluid/tissue is divided by the numerical matrix containing the total amount of markers that could have been detected for each body fluid/tissue. This results in a matrix containing the detection rate of all the markers for each body fluid/tissue type.*

In this study we were interested to see if the different cDNA volumes that were used influenced the detection rate and/or the magnitude of the detected peak. By calculating the detection rate among the markers in the three different volume groups, we were able to see if there was an increment among the 19 mRNA markers. The detection rate was calculated by counting all the detected mRNA markers for each volume group and dividing it by the total number of samples that were in the respective volume group. The calculations and plotting were performed in R with the ggplot2 package. The code used for calculation and plotting can be found in Appendix 1. An analysis of variance (ANOVA) was later performed to see if there was a significant difference between the three volume groups relative to the detection rate.

2.6.2.2 Data transformation before statistical analysis

A transformation of the dataset was necessary before we could perform a logistic regression analysis. The logistic regression models were fit using the glm() function from the R-Stats package. The glm() functions requires at least three inputs: the formula that describes the model we wish to fit, the distribution family and the name of the dataset we want to analyze. The dataset needs to contain the categorical binary variable ($y_1 \ldots y_n$) and the response variables (x_{np}) that are listed as columns in the dataset (see Figure 2.4).

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Figure 2.4: *This figure shows the structure of the dataset that we used in our regression model. The dataset contains (n) data cases and (p) predictor variables. The binary response variable is here (y) and its predictor variables is (x_{np}).*

The original dataset that was uploaded from Microsoft Excel contained 8 columns and 1710 rows. Each row represented information about a marker from a sample taken from a body fluid or tissue from participant (P1, P2, P3 … etc). The 8 columns contain information about the height, what body fluid/tissue the sample is taken from, peak height, volume of PCR and the participant ID.

In the study we modified the dataset to eliminate any mRNA markers that were not detected, leaving just the mRNA markers, their representative peak heights, and the bodily fluid or tissue that the marker was detected in. The transformation of the dataset was done using an R-code that can be found in Appendix 1. An illustration of the transformation is given in figure 2.5 and figure 2.6. The transformed dataset was then saved in R-softwere as “df1”. The dataset can be found in Appendix 1. The transformation made it possible to create various logistic regression models that we later used for prediction.

Participant	S_type	Marker	…	Height
P1	Type1	Marker1	…	Height_Marker1_Type1
P1	Type1	Marker2	…	Height_Marker2_Type1
P2	Type2	Marker3	…	Height_Marker3_Type2
P2	Type2	Marker4	…	Height_Marker4_Type2
P3	Type3	Marker5	…	Height_Marker5_Type3
P3	Type3	Marker6	…	Height_Marker6_Type3

Samples	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6
Sample 1	Height(Marker1-Sample1)	…	…	…	…	Height(Marker6-Sample1)
Sample 2	Height(Marker1- Sample2)	…	…	…	…	Height(Marker6-Sample2)
Sample 3	Height(Marker1- Sample 3)	…	…	…	…	Height(Marker6- Sample3)

Figure 2.5: *The transformation of the original dataset to the dataset that was further used in logistic regression modeling. The transformation created a dataset with the response variable as first column and the predictor variables as the remaining columns in the dataset.*

S_Name	Deltaker	Type	Volum	Marker	Detected	Allele	Height
01-P1MRTPLUS	P1	Menstruasjonsblod	0.5	MYOZ1	J	Vaginal_Mucosa	2329
08-P1SRTPLUS	P1	Spytt	0.5	HTN3	J	Saliva	401
40-P4SERTPLUS	P1	Spytt	3.0	MUC4	J	Mucosa	2689
19-P3MRTPLUS	P3	Menstruasjonsblod	1.0	ALAS2	J	Blood	1248
43-P1SRTPLUS	P2	Spytt	3.0	STATH	J	Saliva_Nasal	31362
09-P4VRTPLUS	P4	Vaginal	0.5	MYOZ1	J	Vaginal_Mucosa	1130
38-P1VRTPLUS	P1	Vaginal	3.0	MYOZ1	J	Vaginal_Mucosa	17624

Type	MYOZ1	HTN3	MUC4	ALAS2	STATH
Menstruasjonsblod	2329	0	0	0	0
Spytt	0	401	0	0	0
Spytt	0	0	2689	0	0
Menstruasjonsblod	0	0	0	1248	0
Spytt	0	0	0	0	31362
Vaginal	1130	0	0	0	0
Vaginal	17624	0	0	0	0

Figure 2.6: *This is a visualization of a section of the original dataset that has been transformed.*

2.6.3 Statistical analysis

2.6.3.1 Analysis of Variance (ANOVA)

An analysis of variance (ANOVA) was used to determine if there were significant differences between samples with different amount of PCR input volume. The mean value of each chosen group is compared using one-way ANOVA to determine whether there is a statistically significant difference between the groups. The hypothesis that is used for this analysis was described as follows:

$$H0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_p$$

H1: means are not all equal

Here μ_k is equal to the mean value of each group k. The independent categorical variable (predictor variables) x_p were in this study the three different volume groups: 0.5µL, 1.0µL and 3.0 µL (p =1, 2, 3), whereas the quantitative dependent variable (the response variable) was

the measured peak height of all the mRNA markers. The One-way ANOVA analysis were done using the R- computer software.

We also performed a two-way ANOVA with two categorical variables (predictor variables). This was to estimate how the mean peak height value changed according to both the cDNA volume and what type of marker that was detected. This was done for all body fluid types and their detected markers.

2.6.3.2 Logistic regression Analysis

A logistic regression analysis was used to model the probability of the binary outcome based on the predictor variables: $x_1, \ldots, x_p$. We chose the detected peak height of all the mRNA markers as the predictor variables in the logistic regression model. For each body fluid/tissue types, we created a univariate and a multiple logistic model with the response variable (y) as the body fluid/tissue type and the predictor variable(s) ($x_1, \ldots, x_p$) as the mRNA markers that were detected in the body or tissue type that we were testing. The response variable y is binary and is set to either 1 or 0 depending on the body fluid or tissue we are analyzing is present or not:

$$P = P(y = 1) = P(\text{"}\textit{the body fluid or tissue is present}\text{"})$$

Hence

$$1 - P = P(y = 0) = P(\text{"}\textit{the body fluid or tissue is not present}\text{"})$$

We use the multiple logistic model:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

That can be reformulated to:

$$P = \frac{e^{\beta_0+\beta_1 x_{1+\cdots+}\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_{1+\cdots+}\beta_p x_p}}$$

We let PI (prognostic index) be:

$$PI = e^{\beta_0+\beta_1 x_{1+\cdots+}\beta_p x_p}$$

Then:

$$P = \frac{e^{PI}}{1 + e^{PI}}$$

where:

- For the univariate logistic model we let the prognostic index (PI) be: $e^{\beta_0 + \beta_1 x_1}$
- P is the probability of the occurrence
- $\beta_0, \beta_1 \ldots, \beta_{np}$, are the estimated coefficients in the regression model.
- $x_1, \ldots x_p$ are the values of the predictor variables

2.6.3.3 The different datasets

After transforming the original dataset, we created an univariate logistic regression model with the target body fluid as response variable and one of its representative RNA marker as predictor variable. Likewise, we created a multivariate logistic regression model with the target body fluid as response variable, but with all the representative RNA markers as predictor variables. This means that for a body fluid with *n* number of markers we created *n+1* different logistic models for the targeted body fluid.

We created three different datasets based on the original dataset collected from the numerous RNA profiles: df1, df2 and df3, described in detail below. Using different datasets, such as quantile data or datasets with varying characteristics, can provide valuable insights into the model's performance and generalizability. Diverse datasets help evaluate the model's robustness by assessing its ability to handle variations in the detection of mRNA markers in the different profiles. It can help us identify factors that contribute to improved or diminished prediction accuracy. By testing the model on multiple datasets, we can gain a better understanding of its reliability and determine the contexts in which it performs optimally.

Df1

The dataset labeled as df1 exhibited minimal alterations, maintaining consistent detected peak height values comparable to those in the original dataset (Figure 2.6).

Df2

The df2 dataset employed a quantile-based representation of the original peak height distribution, with values categorized into distinct quantiles ranging from 1 to 20. The value of the quantile is determined by a distribution based on a range of RFU values, which goes from

a minimum of 0 to a maximum of 32,662. The corresponding peak height ratio for each quantile can be found in table 2.5.

Table 2.5: *Table containing the corresponding peak height ratio for each quantile.*

Quantile (1-10)	**RFU - value**	**Quantile (11-20)**	**RFU - value**
1	0	**11**	16331.0
2	1633.10	**12**	17964.1
3	3266.20	**13**	19597.2
4	4899.30	**14**	21230.3
5	6532.40	**15**	22863.4
6	8165.50	**16**	24496.5
7	9798.60	**17**	26129.6
8	11431.7	**18**	27762.7
9	13064.8	**19**	29395.8
10	14697.9	**20**	31028.9 <

Quantile datasets can help address issues related to data skewness. In logistic regression, an imbalanced distribution of the predictor variable can lead to biased predictions. By using a quantile dataset, which represents the data distribution in a more balanced manner, the model can learn from a more representative sample, reducing the impact of skewed data on the prediction. This can result in a more accurate performance of the logistic regression model.

Df3

The third dataset, denoted as df3, incorporated the same quantile values as the previous dataset (df2). However, df3 introduced an additional penalty mechanism to account for mRNA markers detected in incongruent body fluid types. This penalty system aimed to reflect the discrepancy between the observed marker and its associated body fluid. The penalty gives a negative sign to the quantile value of markers that is detected in an unexpected body fluid.

Consequently, higher detected RFU values gain larger penalties in the same way a larger RFU value gives a higher quantile value.

For instance, if the vaginal secretion marker MUC4 was mistakenly detected in a nasal sample with a peak height (RFU) value of 11 246, it would be assigned a transformed value of – 7. The number 7 is given by looking at the corresponding quantile value (see Table 2.4) and the negative value indicates that the marker is detected in an unexpected body fluid. The quantile value will still be the same as in df2, only with a different

In logistic regression, the choice of penalty or cost function for misclassifications can impact the model's behavior and predictive performance. By varying the penalty for mistaken detection of genes, we can explore the impact the unexpected, detected RNA markers have on the dataset.

Each dataset was used to create the n+1 logistic regression models for all the body fluid types giving a total of 3(n+1) logistic regression models with different properties.

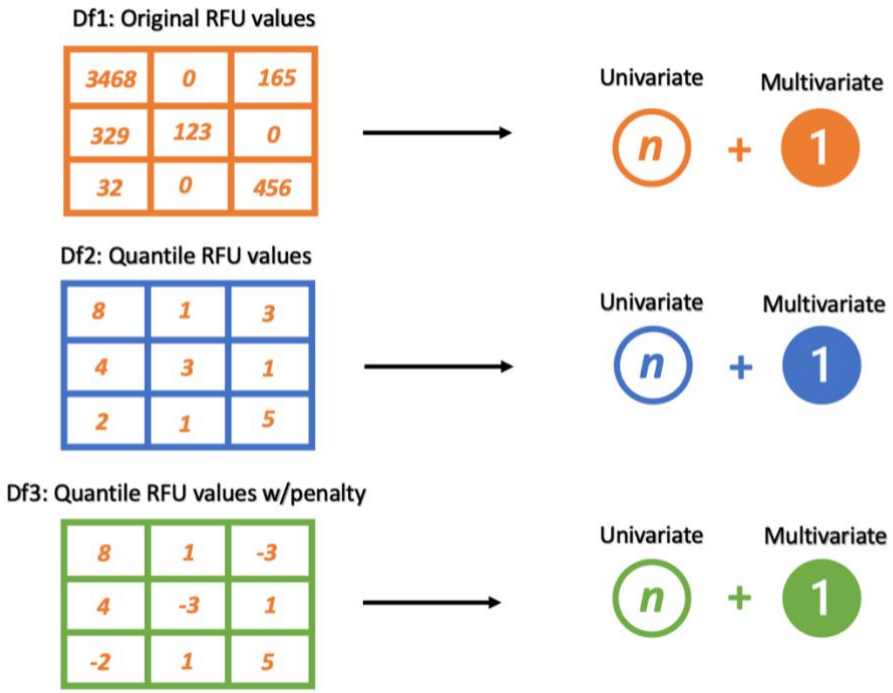


***Figure 2.7**: the distribution of all regression models created based of the three datasets. The values in dataset does not correspond to the correct quantiles, the values are chosen randomly.*

3. Results

All R-codes that were used to create the figures and tables in this section, can be found in Appendix 1. The datasets containing the underlying values used to create the different figures and tables, can be found in Appendix 2. In this section we will first give some general information about the dataset containing data collected from the 90 RNA profiles generated from the GeneMapper software. We will then compare each body fluid based on values collected from its representative RNA-profile. Lastly, we will go through each body fluid independently and provide an informative overview of the qualities each body fluid holds and present some of its statistical properties.

3.1 General information about the dataset

The 38 samples that we collected from the six different body fluid/tissue types, created a total number of 90 RNA-profiles with 615 detected mRNA markers. The presence of the housekeeping genes (18S- rRNA and ACTB) is required for all RNA-profiles to be considered valid. The profiles that do not have these markers detected, were not taken into this study. Each body fluid has two to four representative mRNA markers that is exclusively expressed in that specific body fluid. A summary of the representative mRNA markers for each body fluid can be found in *Table 3.1*.

Table 3.1: *A summary of the various mRNA markers exclusively expressed in a specific body fluid.*

Body fluid/ tissue	**Markers**		
Menstruation blood	MMP7	MMP10	MMP11
Blood	HBB	ALAS2	CD93
Vaginal secretion	MUC4	MYOZ1	CYP2B7P1
Semen	KLK3	PRM1	SEMG1
Nasal secretion	STATH	BPIFA1	
Saliva	HTN3	STATH	
Housekeeping	18S-rRNA	ACTB	
Sex (F/M)	XIST	RPS4Y1	

Table 3.2 contains the detection rate of markers, i.e, the calculated fraction of the detected markers per all possible detections that could have occurred. We can see from the table that the detection rate of markers varies between the different body fluids. As expected, all body fluid types a have a 100% detection rate for the two housekeeping genes. The blood specific markers CD93 and HBB appear in all body fluid types except for the semen samples. The specificity for these two markers is therefore relatively low. We can see that the mRNA markers expressed in semen, KLK3, SEMG1 and PRM1, are among the most specific markers and are only detected in the semen samples.

Table 3.2: *A table that contains the detection rate for each marker in each body fluid. The name of the mRNA markers is listed in the left column and the rate of detection for each body fluid in the right. Because we only kept the samples that had both housekeeping genes in this study. The detection rate for the 18S-rRNA gene and ACTB, marked in green, has therefore a detection rate is 1.0 (100% for every body fluid). The sex specific mRNA markers are marked in light blue.*

mRNA marker	**Blood**	**Menstruation blood**	**Nose secretion**	**Semen**	**Saliva**	**Vaginal secretion**
18S-rRNA	1,00	1,00	1,00	1,00	1,00	1,00
ACTB	1,00	1,00	1,00	1,00	1,00	1,00
ALAS2	0,91	0,33	0,00	0,00	0,03	0,00
CD93	1,00	0,89	0,92	0,00	0,24	0,42
HBB	1,00	1,00	0,33	0,00	0,27	0,42
BPIFA1	0,00	0,22	1,00	0,00	0,06	0,08
CYP2B7P1	0,00	0,44	0,17	0,00	0,00	0,42
HTN3	0,09	0,00	0,00	0,00	0,88	0,00
KLK3	0,00	0,00	0,00	0,33	0,00	0,00
MMP10	0,00	1,00	0,00	0,00	0,03	0,00
MMP11	0,00	0,78	0,00	0,00	0,00	0,25
MMP7	0,00	0,89	0,25	0,00	0,06	0,25
MUC4	0,09	0,89	0,92	0,00	0,77	1,00
MYOZ1	0,00	0,89	0,00	0,25	0,59	1,00
PRM1	0,00	0,00	0,00	0,83	0,00	0,00

SEMG1	0,00	0,00	0,00	1,00	0,00	0,00
STATH	0,09	0,11	1,00	0,00	0,94	0,00
RPS4Y1	0,27	0,11	0,25	0,92	0,24	0,08
XIST	0,73	1,00	0,50	0,00	0,44	0,58

3.1.1 PCR volume: detection rate and peak height

This section will describe how the input volume of cDNA affects the detection rate of the mRNA markers. Twenty-eight samples were amplified using 0.5, 1.0 and 3.0 μL cDNA input, the remaining 10 samples were only amplified with 1.0 μL. First, we will see if an increment of cDNA volume contributes to a higher detection rate among the mRNA markers. Later, we will determine if there is a statistically significant difference of the detected peak heights between the various volume groups by looking at the results from the analysis of variance (ANOVA).

Figure 3.1 describes the detection rate among the 19 different mRNA markers in the three volume groups. From the plot we can see that there is an increment of detection rate among all mRNA markers between volume 0.5 - 1.0 μL and/or 0.5- 3.0μL. The R-code and dataset that were used to create this plot, can be found in Appendix 1 and Appendix 2 respectively.

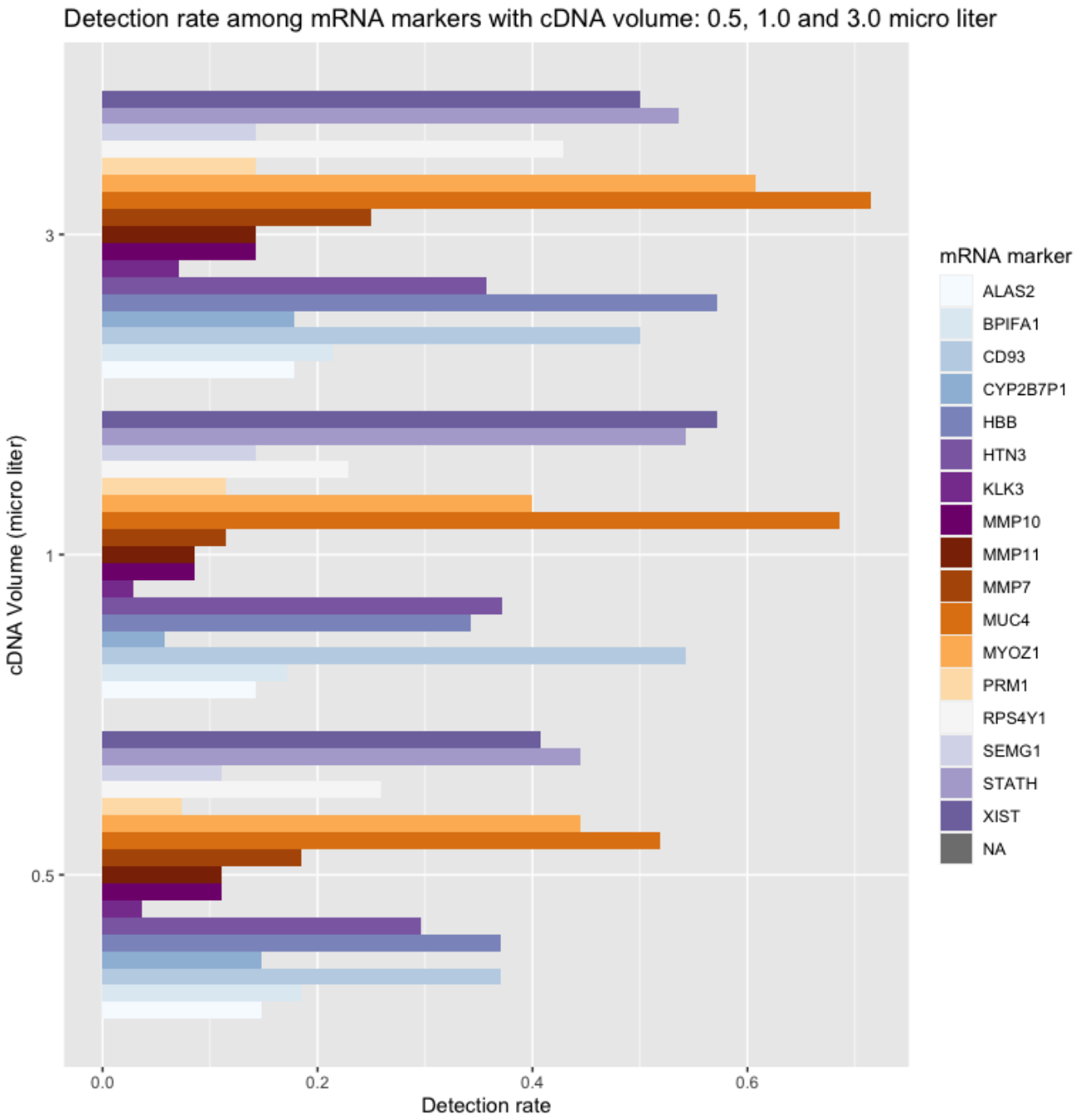


Figure 3.1: *this figure was created in R and plotted with the open-source data visualization package, ggplot2. The detection rate of all markers can be found along the x-axis and the three volume groups: 0.5, 1.0 and 3.0μL, along the y-axis. Each color represents an mRNA marker and is labeled at the right side of figure 3.1. The detection rate increases as the volume increases for all mRNA markers between volume 0.5 and/or 1.0μL or 0.5 and 3.0μL.*

The one-way variance analysis that was performed (one way ANOVA), showed a significant difference of peak height between the three different volume groups (p-value = 0.0000305***). The proportion of variation explained by the model (R^2) was calculated to be 0.972, which means that approximately 97.2% of the variation of the peak height can be explained by the different volume groups. A summary of the ANOVA analysis can be found in *Table 3.3*.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{60.4}{(2097.6 + 60.4)} = 1 - 0.028 = \mathbf{0.972}$$

Table 3.3: *A summary of the one-way ANOVA analysis based on the detected peak height between the different volume groups. We can see that there is a significant difference between the three different groups based on the calculated p-value (0.000305).*

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*1*	*60.4*	*60.41*	*17.65*	*0.0000305****
Residuals	*613*	*2097.6*	*3.42*	-	-

We also performed a two-way analysis of variance to see if the means of the peak height changed according to the levels of the categorical variables: the volume and the mRNA markers. The two-way ANOVA showed that both the cDNA volume and the type of mRNA marker explained the variation in the detected peak height significantly (p value < 0.001). See *Table 3.4* for a summary of the two-way ANOVA.

Table 3.4*: a summary of the two-way ANOVA analysis based on the quantitative variable (peak height) and the two categorical variables (cDNA volume & mRNA marker type). Both categorical variables explain the variation of the peak height significantly with a p-value of 3.84* 10^{-6} and 2.0 *10^{-16} respectively.*

	Df	*Sum Sq*	*Mean Sq*	*F-value*	*P-value*
Volume	*1*	*2.406e+09*	*2.406e+09*	*21.750*	*0.000003.84****
Marker	*18*	*1.479e+10*	*8.216e+08*	*7.429*	*2e-16****
Residuals	*595*	*6.581e+10*	*1.106e+08*	*-*	*-*

3.1.2 The detection of sex specific mRNA markers

In our study, sex-specific mRNA markers, namely RPS4Y1 for males and XIST for females, were consistently detected across various body fluid types. Figure 3.2 illustrates the proportion of these sex markers relative to the total number of detected markers among the different body fluid types. We excluded the housekeeping genes as they were found to be present in all the samples, and their inclusion would not provide additional discriminatory power. One sample from the menstruation sample had to be removed in this section due to a detection of both the female and male sex-specific marker.

The proportion of sex-specific markers, relative to the total number of detected markers, ranged from 13% to 28%. Notably, it was unexpected to find one male-specific marker present in a vaginal secretion sample. In contrast, the semen samples exclusively exhibited detection of the male-specific gene (RPS4Y1), suggesting a higher degree of stability in the detection of this marker within the semen samples.

However, the female biomarker, XIST, showed a higher frequency of occurrence across almost all sample types.

Furthermore, we investigated the occurrence of sex-specific markers for all samples in a specific sample type. Here one of the menstruation blood samples had to be removed since two sex specific mRNA markers were detected in the same sample. In the analysis, we found out that the blood samples and menstruation blood samples had the highest detection rates, with all samples containing a sex-specific marker. In contrast, the saliva and vaginal secretion samples had the lowest detection rates, with approximately two-thirds of their respective samples showing the presence of a sex-specific marker. (*see figure 3.3*)

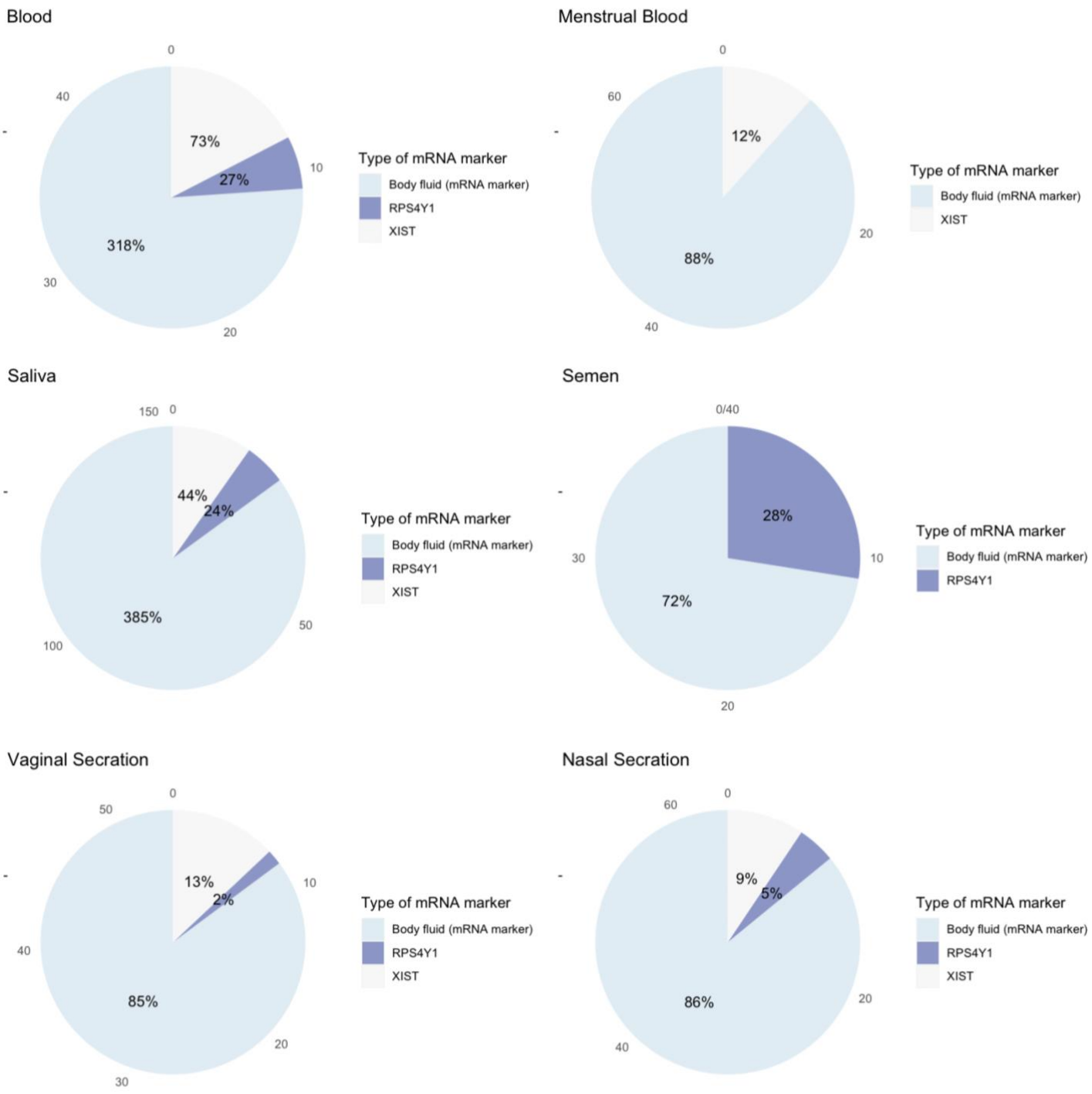


***Figure 3.2**: Shows the proportion of these sex markers relative to the total number of detected markers among the six different body fluid types: blood, menstruation blood, saliva, semen, vaginal secretion, and nasal secretion.*

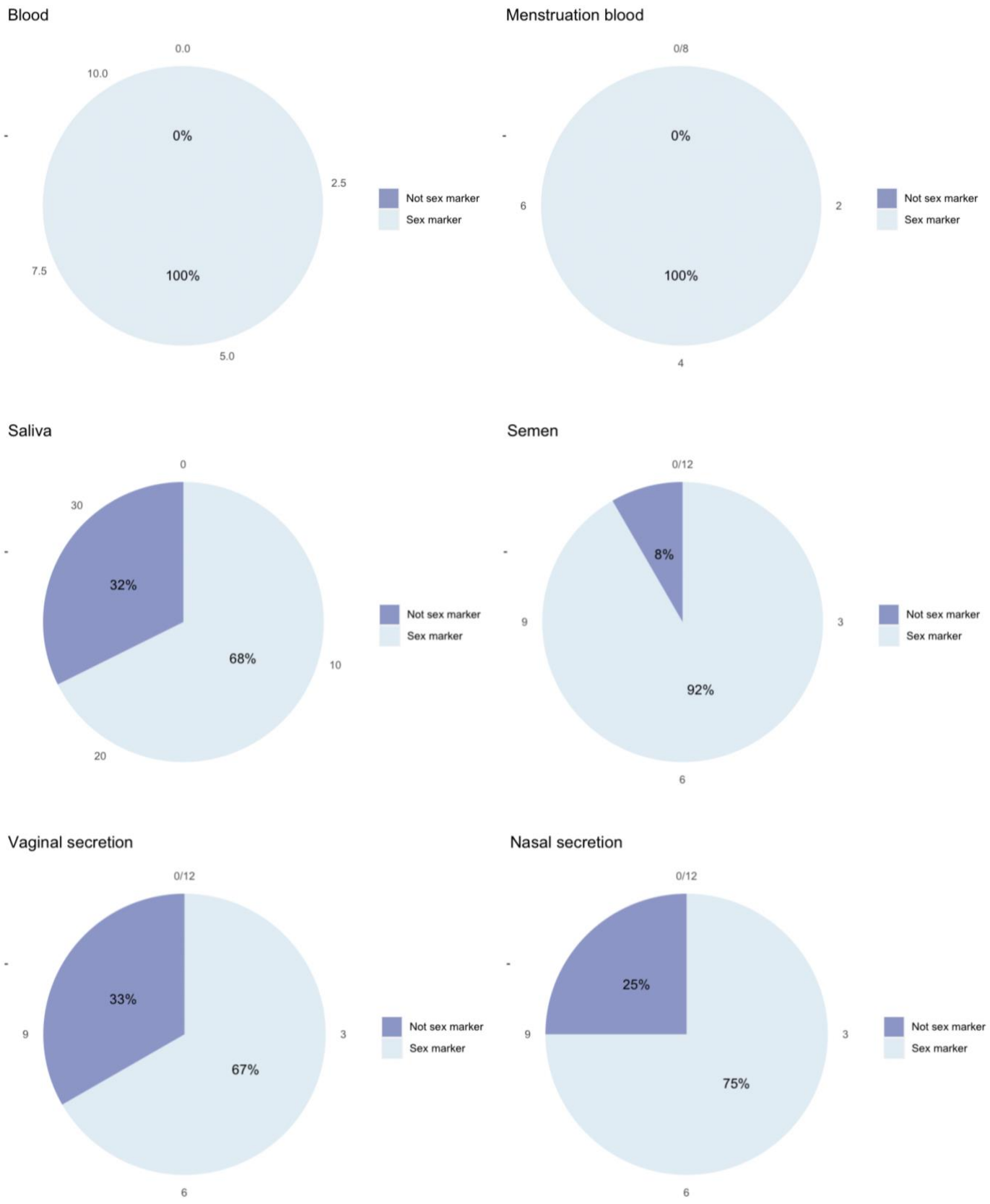


Figure 3.3: *pie chart that illustrate the occurrence of sex-specific markers for all samples in a specific sample type.*

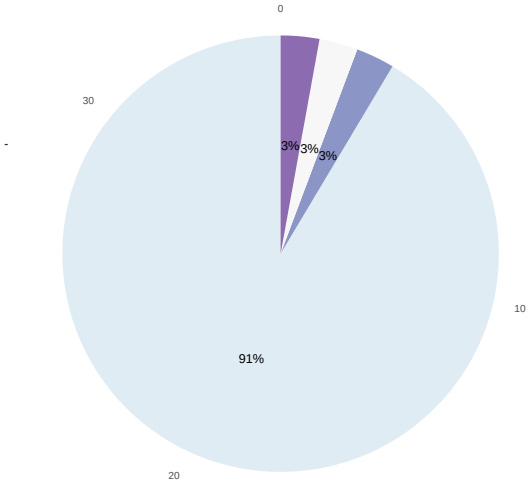
3.1.3 Incorrect detection of mRNA markers

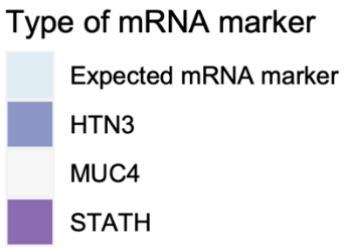
The mRNA markers represent a specific bodily fluid because they are exclusively expressed in a certain body fluid (see Table 1). All markers that are detected in an incorrect body fluid type, are considered incorrectly detected. In this study we chose to make an exception for the mRNA markers detected in the menstruation samples. Since the menstruation blood samples are extracted from the vaginal area and contains blood, will we accept the blood and vaginal mRNA markers as an equal representation for the menstruation samples. This means we consider the detection of MUC4, MYOZ1 & CYP2B7P1 and HBB, ALAS2 & CD93, as normal in the menstruation samples.

The number of incorrect mRNA markers varies between the six different body fluids/tissues. The distribution of the incorrect and correct detected markers for each body fluid are illustrated in figure 3.4. The saliva and nasal secretion samples had in total more incorrect than correct detected markers with only 47% and 44% of the total amount of mRNA markers correctly detected. We can see that the vaginal specific markers, MUC4 and MYOZ1, appear frequently in both the saliva and nasal secretion samples. Surprisingly, the MYOZ1 mRNA marker appears in three of the semen samples and the MUC4 mRNA marker in one in the blood samples. The blood specific mRNA marker HBB, was also incorrectly detected in several other body fluid types like the saliva, vaginal and nasal secretion samples. The body fluid types with the most correct detected mRNA markers, were the samples collected from menstruation blood, blood and semen.

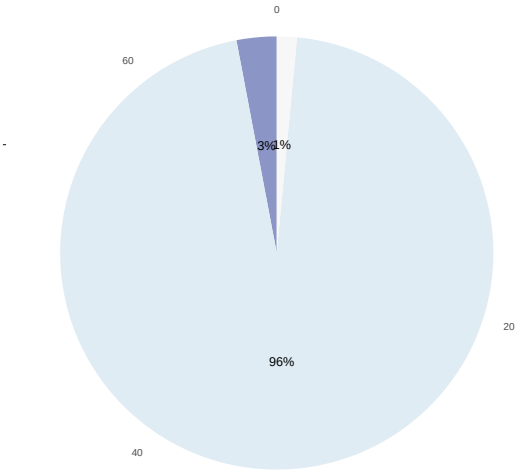
The distribution of incorrectly and correctly detected mRNA markers in each body fluid are summed up in a table that can be found in Appendix 2. The R code used to create the pie charts can be found in Appendix 1.

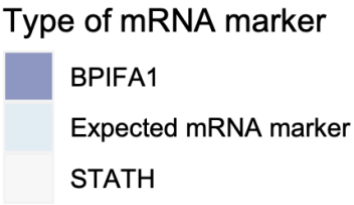
Blood



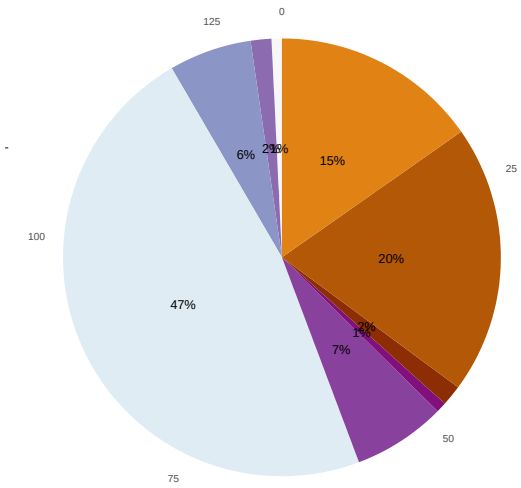


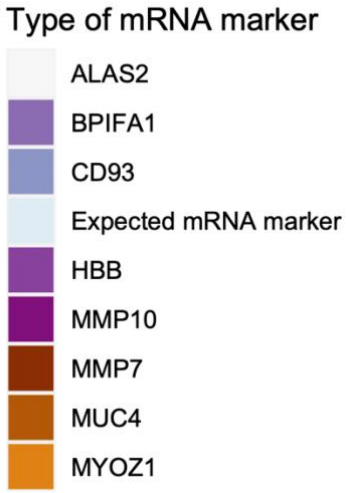
Menstrual Blood



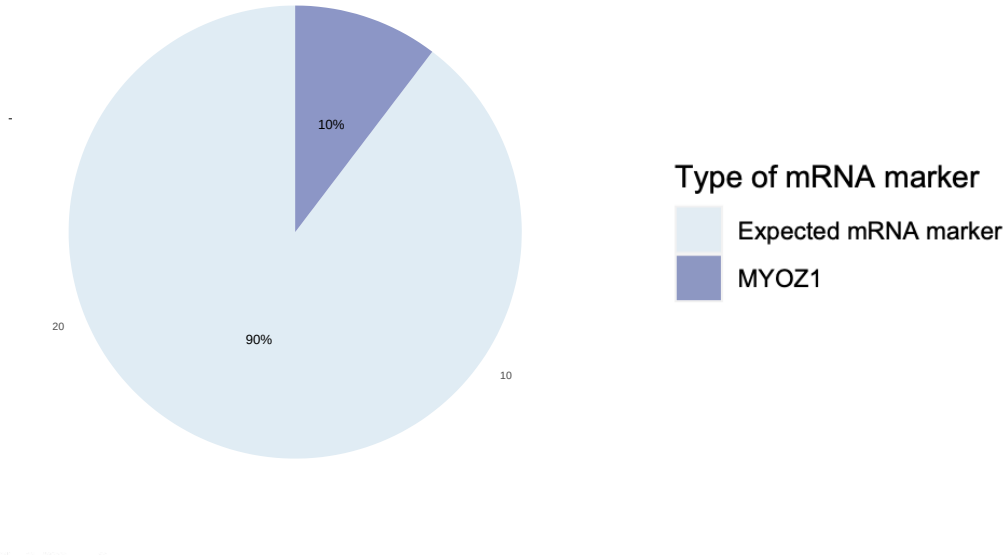


Saliva

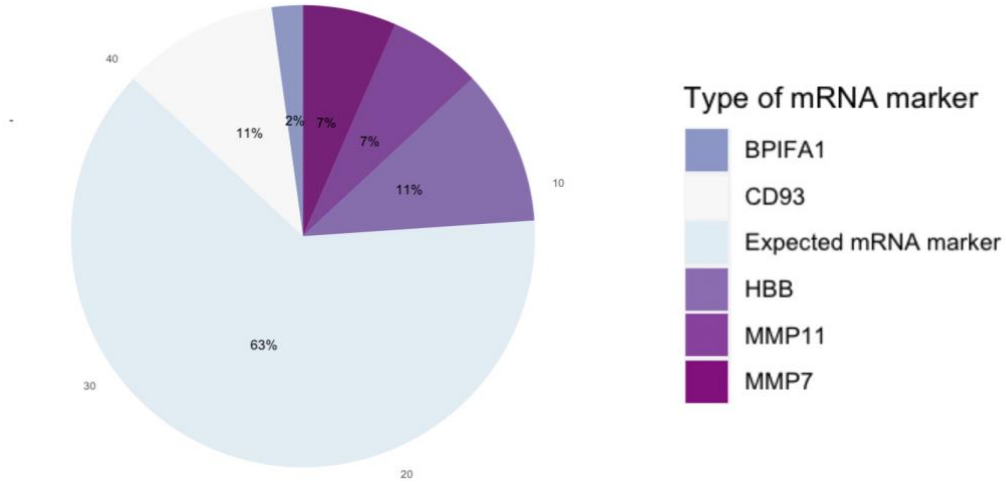




Semen



Vaginal Secretion



Nasal Secretion

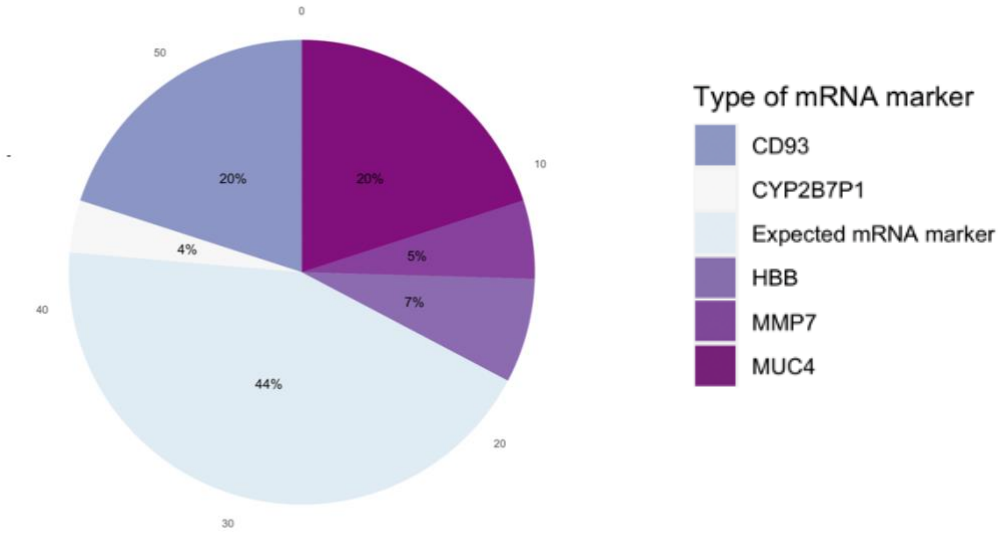


Figure 3.4: *Pie charts that illustrate the distribution of the unexpected and the expected mRNA markers in each body fluid. The unexpected mRNA markers are represented with a specific color that can be found at the right side of the pie chart. The percentage of the expected mRNA markers that were detected are illustrated in a light blue color in each pie chart. The percentage is rounded up to the closest integer.*

3.2 In-depth analysis of each body fluid

In this section we will individually go through each body fluid and look at the traits and qualities for all generated RNA-profiles related to that specific body fluid type. First, we will look at the detection rate and peak height value among the different volume groups and for each individual marker. Second, we will display the correlation between the markers that are specific for the body fluid we are analyzing. Lastly, we will go through the predicted values for each body fluid type and see how well each logistic regression models worked as a predicting module. The housekeeping genes were not taken into this part of the study since they appear in all the 38 samples and are neither body fluid specific nor sex specific markers.

We performed a one-way ANOVA for all body fluids to see if there was a significant difference between the three volume groups and the detection rate in a specific body fluid type. We also performed a two-way ANOVA with the volume and marker as predictor variables to see if either volume or the marker type had an influence on the detected RFU value. For the ANOVA, we set the significance level at 5% (0.05*). A p-value less or equal to 1% (0.01**) is regarded as highly significant. The results from all the ANOVA tests can be found in Appendix 2.

We conducted predictive modeling using fitted logistic regression models to estimate the presence of mRNA markers in different body fluid types. For each body fluid type, we created multiple univariate logistic regression models based on a dataset. In our study, we used three datasets (df1, df2, and df3), resulting in n*3 univariate logistic regression models for each body fluid type. Additionally, we fitted a multivariate model considering all markers for each dataset. Consequently, we obtained a total of n*3 + 3 logistic models for all body fluid types. To assess the performance of these models, we used prediction techniques and evaluated their outcomes using a confusion matrix.

The logistic regression model was tested by dividing the dataset into a training set, consisting of 70% of the samples, and a test set, containing the remaining 30% of the samples. The goal was to categorize each sample as either the target body fluid (binary value: 1) or another body fluid type (binary value: 0). The samples were categorized based on their calculated probability of belonging to the target category. Samples with a probability greater than or equal to 0.5 were classified as the target body fluid, while samples with a probability less than 0.5 were classified as another body fluid type. This approach enabled us to predict and assign categorical labels to the samples in the test set using the logistic regression model's probabilities. We used the same training dataset on all the logistic regression models also the models based of the different datasets. The distribution of the different samples that was tested for prediction can be found in Table 3.5.

Table 3.5*: the distribution of the different samples in the training dataset.*

Body fliud	**Number of samples**
Blood	5 samples
Menstruation blood	3 samples
Semen	3 samples
Saliva	11 samples
Nose secretion	5 samples
Vaginal secretion	3 vaginal

3.2.1 In-depth analysis of the blood samples

This study consisted of 11 blood samples that were used to make 11 RNA-profiles with 68 detected markers. Twenty-two of these markers were housekeeping genes, 11 were sex specific genes and the remaining 35 were specific mRNA biomarkers for body fluid identification. Only 3 out of the 35 mRNA markers were unexpectedly detected.

The peak height registrated in the RNA-profile, can be seen as a representative value of the cDNA quantity for the amplified mRNA marker. The variation in peak height value among the different blood specific markers, can be seen in a box plot in Figure 3.5.

From the plot, we can see that the mRNA marker ALAS2, had the highest variance among the eight different markers. The male specific sex marker, RPS4Y1, had the highest mean peak value (32088 RFU) while the marker CD93, had the lowest (7137 RFU).

After performing a two-way ANOVA, we found a high significant difference in mean peak height for the volume groups and the eight different mRNA markers (*p-value* < 0.01**). (See table 44 in Appendix 2). However, the one-way ANOVA showed no statistically significant difference in detection rate between the three volume groups (*p-value* >0.05). An increment in volume, did not increase the detection rate in the blood samples.

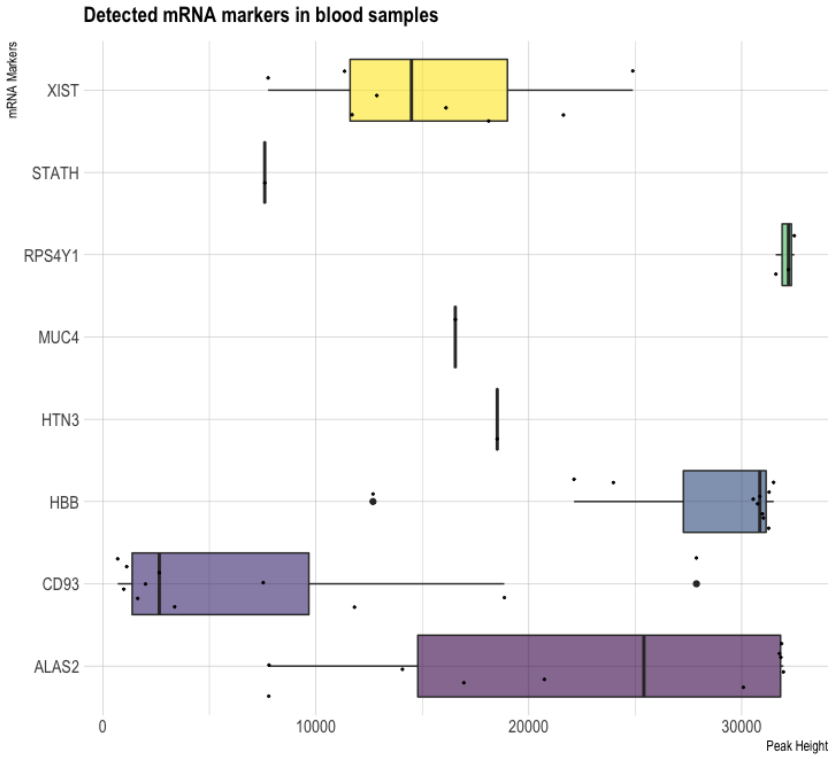


Figure 3.5*: A box plot that shows the detected peak height (RFU) for each mRNA marker in the RNA-profiles created by the 11 blood samples.*

3.2.1.1 Correlation between the mRNA markers in the blood samples

We wanted to see if there was a positive or negative correlation between the three blood specific markers based on their peak height magnitude. A positive correlation between two markers indicates a common change or detection in peak height. A negative correlation can

indicate that a high RFU value of one marker comes with a low RFU value from another marker.

The correlation between the markers: ALAS2, CD93 and HBB, is illustrated in Figure 3.6. A high correlation is illustrated as a larger circle while the color of the circle gives information about the sign of the real number calculated *(red= negative, blue= positive).* The color gradient can be seen on the right side of the plot.

From Figure 3.6, we can see a heigh correlation between the RFU values detected in the HBB marker and the ALAS2 marker (*correlation = 0.67*). The lowest correlation was detected between the HBB and the CD93 marker (*correlation = 0.31*).

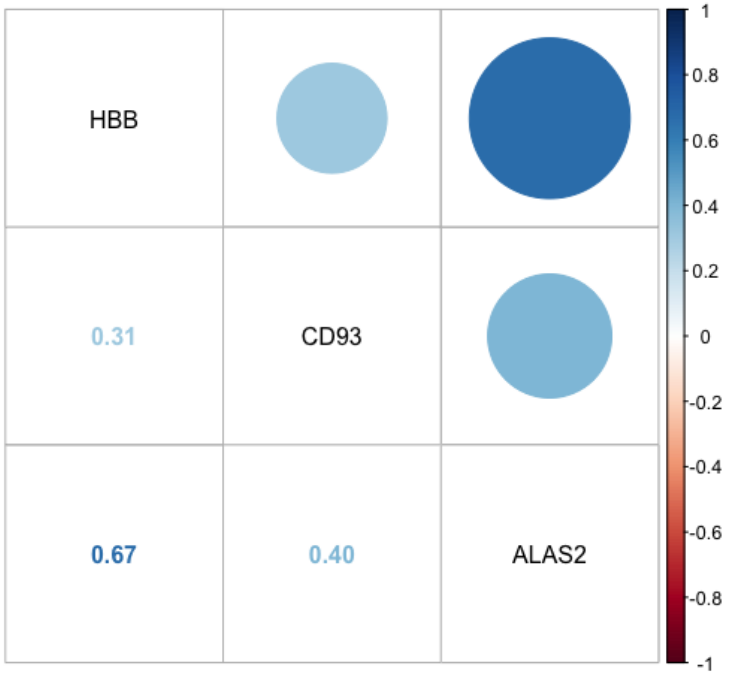


Figure 3.6: *A plot that illustrates the calculated correlation between the three blood specific mRNA markers. The correlation is illustrated as a circle where the size of the circle gives information about the correlation value while the color gives an information about the sign of the real number calculated (red= negative, blue= positive).*

3.2.1.2 Statistical properties and prediction for blood samples

We used logistic regression to create a suitable model for blood sample prediction.

In this section we will first describe the summary of the logistic regression models that we created from the collected data from the blood samples. All summaries of the logistic regression models can be found in Appendix 2 from table 7 to 24. Second, we will present the results from the prediction based on a train and a test dataset created from each logistic regression model.

Twelve logistic regression models were created based on three different datasets: df1, df2 and df3. The characteristic for each dataset is described in the method section 2.6.3.3, the datasets can also be found in Appendix 1. One multivariate model was created with each blood specific marker, HBB, ALAS2, CD93, as a predictor variable to the fitted model. Three univariate models were also generated with each marker as a predictor variable. The models were created as training dataset based of 70% of the dataset and tested with the remaining 30%.

From the predictions, we could conclude that the multivariate logistic regression models were the best fitted models for prediction. Also, the univariate model based of the ALAS2 marker gave an equally good prediction for all datasets. There was no noticeable difference between the three datasets, consequently we will only mention the result from the multivariate model created from the dataset df1. The results from the prediction can be summed up in a confusion matrix as in Figure 3.7.

The confusion matrix shows the true positives (*left top square*), the false positives (*right top square*), the false positives (*left bottom square*) and the false negatives (*right bottom square*).

The test dataset consisted of five blood samples. Three out of the five blood samples in the test dataset were detected and categorized as blood. This means that 60% of the blood samples was correctly categorized as blood. Two samples in the test dataset were predicted unexpectedly as false negatives among the 30 samples. The main difference between the expected and unexpected predicted blood samples was a lower detection in peak height from the ALAS2 mRNA marker.

We can see that there were no false positive predictions, which contributed to a heigh specificity (*1.0*). Like the rest of the predicted models, the accuracy was calculated to be 93.33%.

Summaries for all the predicted models can be found in Table 25-28 in Appendix 2. The R code that fitted the logistic regression models and preformed the predictions, can be found in Appendix 1.

Table 3.6: The accuracy of the prediction is calculated to be 0.93 (93%) with a sensitivity of 0.6 and specificity of 1.0.

Accuracy	**0.933**
Sensitivity	0.600
Specificity	1.000

Figure 3.7: a table consisting of the calculated values from the prediction of the blood samples. From the prediction we got 3 true positives, 25 true negatives and 2 false negatives. From all 5 blood samples the model managed to find and categorize 3 of them. This and all other confusion matrixes is plotted in R-studio with the function confusionMatrix () from the caret package.

3.2.2 In-depth analysis of the menstruation blood samples

We had a total number of 9 menstruation samples in this study. The 9 samples created 9 RNA-profiles consisting of 95 detected mRNA markers. 18 markers were housekeeping genes, 10 detected markers were sex specific markers and 67 markers body fluid specific mRNA markers. There were only 3 markers that were unexpectedly detected among the body fluid specific mRNA markers.

Figure 3.8 illustrates a boxplot based of the peak height for each detected marker in the menstruation blood samples. We can see from the plot that the 4 menstruation blood specific markers, HBB, MMP7, MMP10 and MMP11, have a relatively heigh mean peak height except for the MMP11 mRNA marker. Surprisingly, we can also see a low detection in the male specific sex marker, RPS4Y1 (RFU= 165). All markers that are considered unexpectedly detected, have a very low mean peak height value.

We performed a two-way ANOVA to see if there was a significant difference in mean peak height between the three volume groups (0.5, 1.0 and 3.0 μL) and between the detected mRNA markers. We found out that there was a significant difference in peak height between the three volume groups (p-value <0.05), including a significant difference between the mRNA markers. Also, the detection rate did not significantly increase due to an incensement in volume. However, by performing a one-way ANOVA with the volume groups as the only response variable, we found out that there was no significant difference between the volume groups.

***Figure 3.8**: A box plot showing the detected peak height for each marker found in the menstruation blood samples.*

3.2.2.1 Correlation between the mRNA markers in the menstrual blood samples

A correlation matrix was created to see if there was a common change between the four menstruation blood markers (see Figure 3.9). There was no heigh correlation between most of the mRNA markers. The only two markers that had a significantly heigh correlation was the MMP10 and the MMP7 marker (*correlation = 0.88).* The MMP7 and MMP11 mRNA markers had a relatively low correlation compared with the others. This might be due to the low RFU detection in of the MMP11 marker (*correlation = 0.16*).

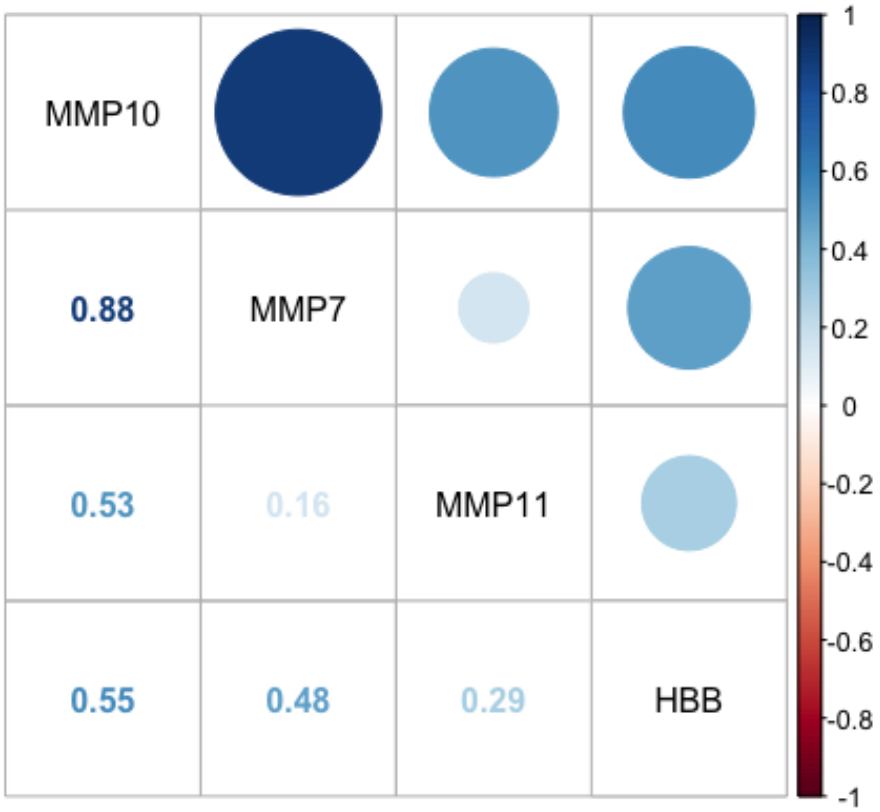


Figure 3.9: *a correlation matrix showing the correlation values between the four menstruation mRNA markers. The size of the circle corresponds to the value of the number while the color represents the sign of the value.*

3.2.2.2 Statistical properties and prediction for menstrual blood samples

We created fifteen different logistic regression models based on each datasets (df1, df2 & df3). Like the blood samples, we fitted a univariate model for each marker that were

exclusively expressed in the menstruation blood samples and a multivariate model for all the menstruation blood specific markers combined.

We saw no significant difference in prediction between the three datasets. There was on the other hand a difference in prediction between the four univariate models and the multivariate model. The model that gave the best prediction was the two univariate logistic regression models that had the MMP10 and the MMP11 marker as the response variable. These two models had a 100% accuracy in almost all the three datasets. The predicted values from the two univariate models can be illustrated in Figure 3.10. The other predictions can be found in table 29-31 in Appendix 2.

All three menstruation samples from the training dataset got detected and categorized correctly, giving the model a faultless prediction.

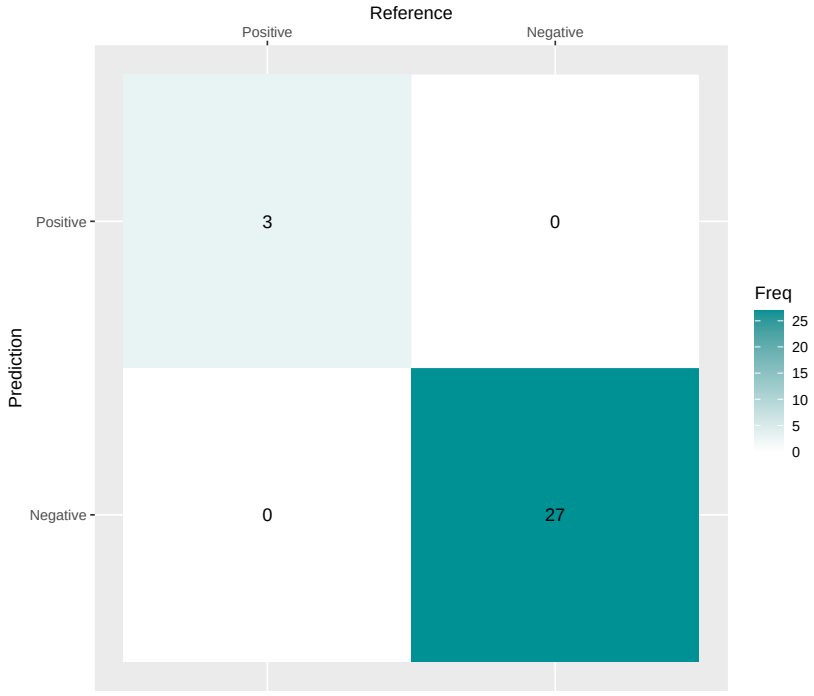


Table 3.7: The accuracy of the prediction is calculated to be 1.0 (100%) with a sensitivity of 1 and specificity of 0.6.

Accuracy	1.000
Sensitivity	1.000
Specificity	1.000

Figure 3.10: *All three menstruation samples got categorized correctly with 3 true positives and 25 true negatives. No false positives nor false negatives were predicted.*

3.2.3 In-depth analysis of the saliva samples

There were a total number of 34 saliva samples in this study. The 34 samples created 34 RNA-profiles with 222 detected mRNA markers. 68 of these markers were housekeeping genes, 23 markers were sex specific markers and the remaining 131 detected were body fluid specific mRNA markers.

Figure 3.11 shows a boxplot based on the detected peak height among the 12 different mRNA markers that were detected in the saliva samples. From the plot we can see that the saliva mRNA marker; HTN3, has the highest mean peak value of all the detected mRNA markers *(mean RFU value = 18067).* The other saliva specific marker, STATH, was also detected in 94% of all saliva samples and had a mean peak height value of 9612 RFU. Surprisingly, both the vaginal secretion and menstruation blood markers were detected in many of the samples. The MUC4 marker and MYOZ1, that is specific for vaginal secretion, got detected in 76% and 59% of the samples.

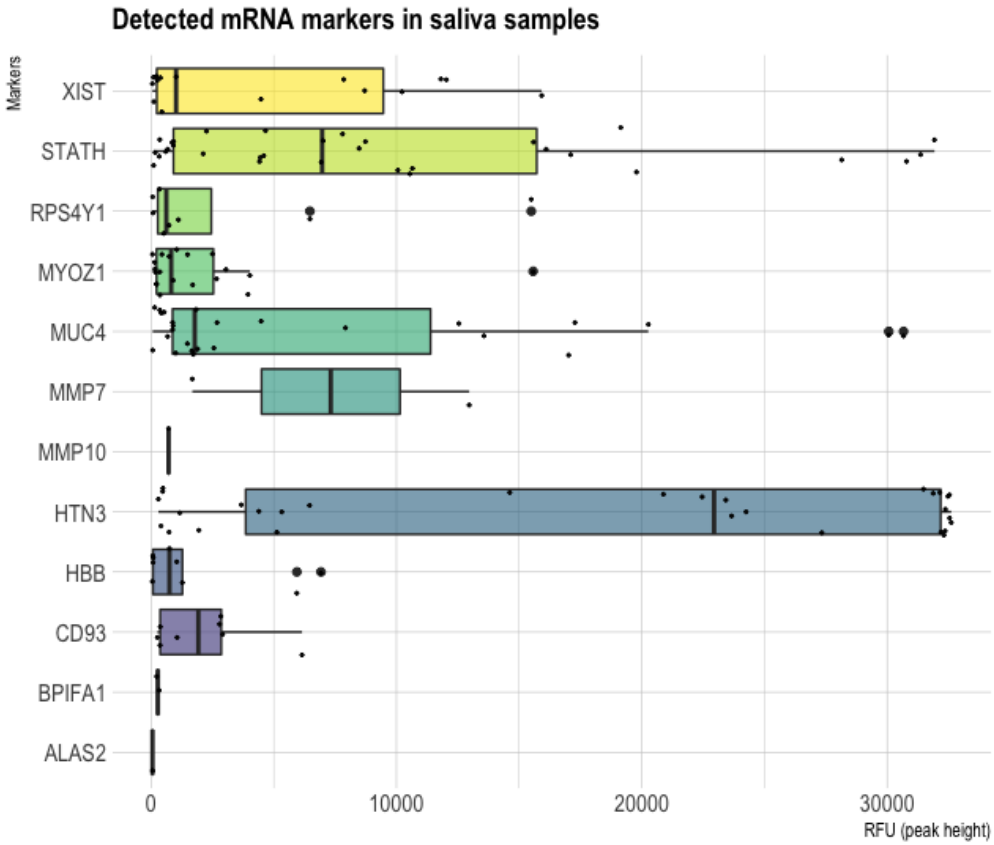


Figure 3.11: *A boxplot showing the distributed peak height for all the 12 detected mRNA markers in the saliva samples.*

We performed a two-way ANOVA to see if there was a significant difference in mean peak height between the volume groups and the 12 detected markers. from the performed ANOVA we found a significant difference between the mRNA markers *(p-value < 0.01)*, but no significant difference between the volume groups *(p-value>0.05).* For the detection rate, we could conclude, by looking at the one-way ANOVA values, that there was no significant increasement in detection rate due to an increasement in volume *(p-value >0.05).*

3.2.3.1 Correlation between the mRNA markers in the saliva samples

The correlation between the two saliva samples HTN3 and STATH was calculated to be relatively low with a value of 0.37. A correlation matrix of the two mRNA markers is shown in Figure 3.12. The size and color of the circle indicates a low, but positive correlation between the saliva specific markers.

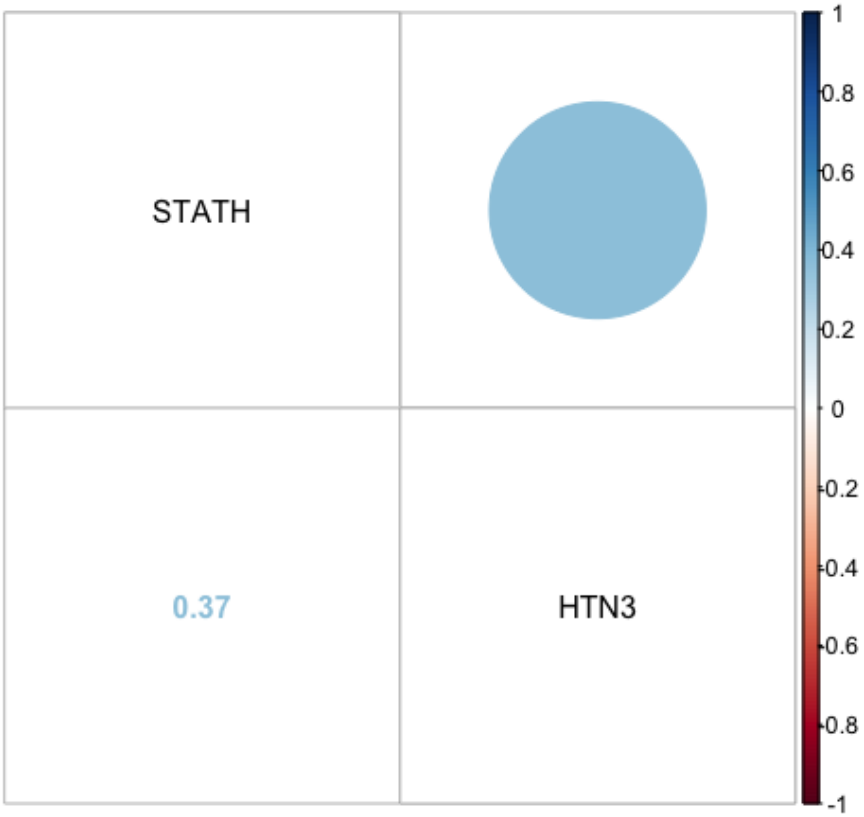


Figure 3.12: *A correlation matrix showing the correlation values between the two mRNA markers that is specific for saliva. High positive correlation gives a larger dark blue circle, while a low negative correlation gives a smaller red circle.*

3.2.3.2 Statistical properties and prediction for saliva samples

We created 9 logistic regression models for the saliva samples, two univariate models for each mRNA marker related to saliva and one multivariate model created based on both mRNA markers. We created the univariate and multivariate models for each of the three datasets.

We used the same proportion of training and test data as the other body fluids in this study. We tried predicting the test dataset based on the model created from the train dataset. From the prediction we could conclude that the multivariate model created from the df3 dataset was the model that predicted the best. This model had an 96.7% accuracy with only one false negative predicted. This one false negative marker had a two out of three wrong detected mRNA markers. The summary of the prediction can be found in a confusion matrix in figure 3.13. From the other models, we found out that the HTN3 marker was a better predictor variable for prediction than the STATH mRNA marker. The lg3.b model had a 100% sensitivity and a 91% specificity.

Table 3.8: The accuracy of the prediction is calculated to be 0.97 (97%) with a sensitivity of 0.91 and specificity of 1.0.

Accuracy	0.966
Sensitivity	0.909
Specificity	1.000

Figure 3.13: *In this prediction all except one saliva samples got detected and categorized correctly. One saliva marker did not get detected giving a false negative value.*

3.2.4 In-depth analysis of the semen samples

We created 12 RNA-profiles from the 12 semen samples that were collected. We detected in total 64 mRNA markers, 24 markers were housekeeping genes, 11 were sex specific genes and the remaining 29 were mRNA markers specific for a body fluid. There were three unexpectedly detected mRNA markers in the semen samples.

The detection in peak height varied among the five different detected mRNA markers. For all markers, the mean peak heights were generally quite low. The male specific sex marker had the highest mean peak height of 9443.364 RFU. The lowest mean peak height among the semes specific markers was the detected among the KLK3 markers (RFU= 120.5). The unexpectedly detected marker, MYOZ1, had the lowest mean peak height of all markers. A summary of the peak heights for all markers detected in the semen samples can be found as a boxplot in figure 3.14.

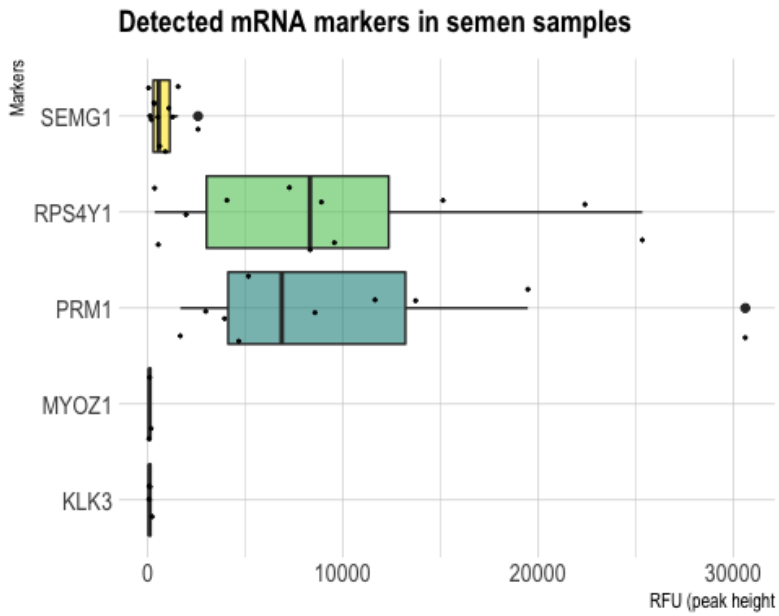


Figure 3.14: *a box plot showing the distribution of peak heights among the detected markers in semen samples.*

after performing a one-way ANOVA, we found a slight significant difference in peak height between the volume groups in the semen samples after performing a one-way ANOVA (P value= 0.022). We did not, on the other hand, find any significant difference in detection rate among the different volume groups (P-value>0.05).

3.2.4.1 Correlation between the mRNA markers in the semen samples

We found the correlation between the three semen markers, SEMG1, KLK3 and PRM1, and found a slight high positive correlation between the SEMG1 and PRM1 mRNA marker. (corr= 0.58). The correlation between the other semen specific markers were quite low, but positive. The correlation between the three markers can be summed up in a correlation matrix (Figure 3.15).

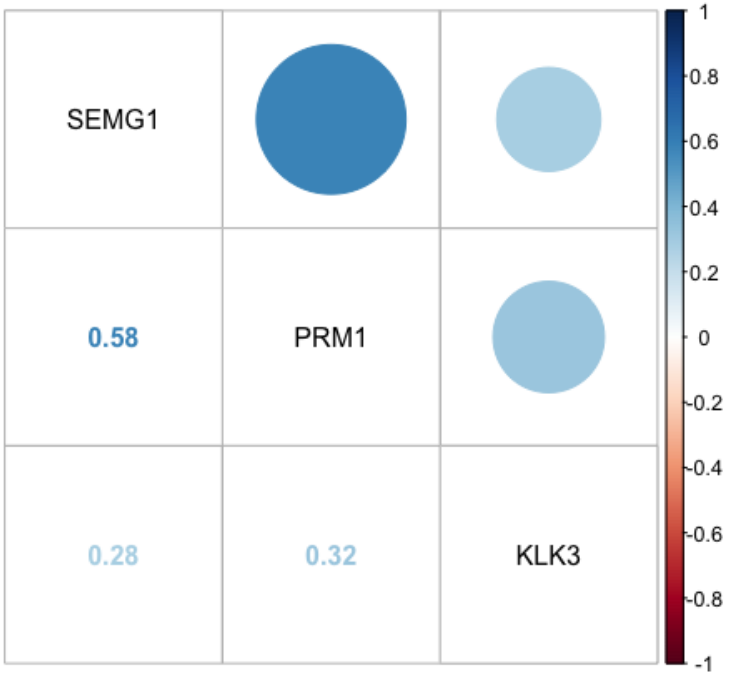


Figure 3.15: *a correlation matrix that shows the correlation values between each pair of semen specific mRNA markers. A high positive correlation gives a larger dark blue circle, while a low negative correlation gives a smaller red circle.*

3.2.4.2 Statistical properties and prediction for semen samples

For the semen samples we created 12 logistic regression models, 9 univariate models based of on each mRNA specific marker for each dataset and three multivariate model for all three markers combined. The 12 models were created by a training dataset and tested by a test dataset. From the prediction we found out that all models, except from the multivariate model created from dataset df2 (lg2.se), had a perfect prediction with 100% accuracy with highest sensitivity and specificity.

In our test dataset we had, like the vaginal secretion samples, only three semen samples. Unlike the prediction form vaginal secretion that only had one out of three correct detected samples, the logistic regression models for prediction of semen samples managed to categorize all the samples correctly.

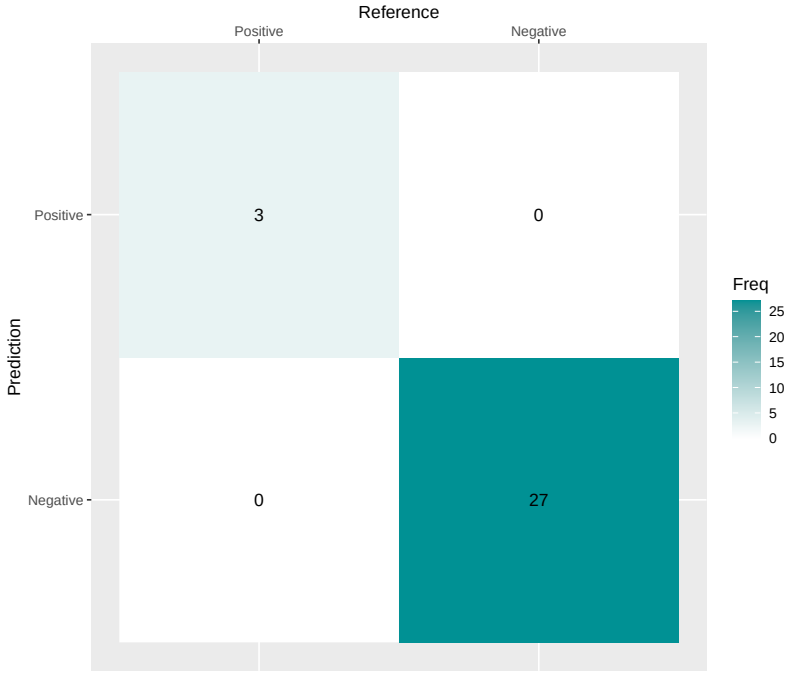


Table 3.9: The accuracy of the prediction is calculated to be 1.0 (100%) with a sensitivity of 1.0 and specificity of 1.0.

Accuracy	1.000
Sensitivity	1.000
Specificity	1.000

Figure 3.16: *The prediction of the semen samples gave a perfect prediction with no false negative nor false positives. All three semen samples from the training dataset got predicted correctly.*

3.2.5 In-depth analysis of the vaginal secretion samples

This study consisted of 12 vaginal secretion samples given by voluntary participants. The 12 RNA profile created by these samples consisted of 78 detected RNA-markers. 24 markers were housekeeping genes, 8 were sex specific mRNA markers and the remaining 46 were markers exclusively expressed in a specific body fluid type. The predictions and statistical analysis were calculated without the detected values from the housekeeping genes.

A one-way ANOVA was performed to see if there was a significant difference in RFU values between the three volume groups. Another one-way ANOVA was performed to see if there was a difference in detection rate among the three different volume groups. We found no significant difference in the RFU values nor the detection rate between the three volume groups ($p > 0.05$). There was, on the other hand, a significant difference between the RFU

values and the detection rate between the 10 different detected markers ($p < 0.05$). A summary of the detected peak height for each mRNA marker is shown in a box plot in Figure 3.17.

The MUC4 mRNA marker had a much higher RFU value than the other markers. The MUC4 marker had a mean value of 25713.08 RFU, while the MYOZ1 markers, that had the second highest mean value, had a mean value of only 3645.58 RFU. The unexpectedly detected mRNA markers have a rather low detection in peak height.

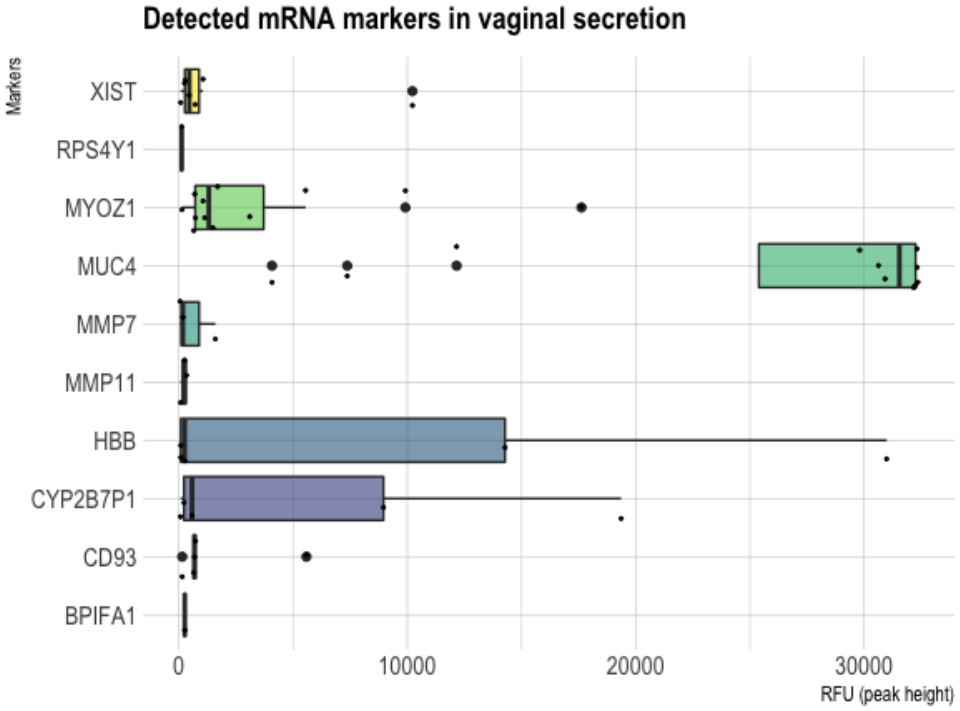


Figure 3.17: *a box plot showing all detected peak heights for each marker. The detected mRNA markers can be found along the Y-axis, while the corresponding RFU value for each detected sample can be found along the X-axis.*

3.2.5.1 Correlation between the mRNA markers in the vaginal secretion samples

We used the R-programing software to calculate the correlation between the three markers that were specific for vaginal secretion: MUC4, MYOZ1 and CYP2B7P1. The result is summed up in a correlation matrix (figure 3.18). From the figure we can see that there a high positive correlation between the MYOZ1 and the CYP2B7P1 marker (correlation = 0.65), but a rather low positive correlation between the MUC4 and the other two markers (correlation <0.35).

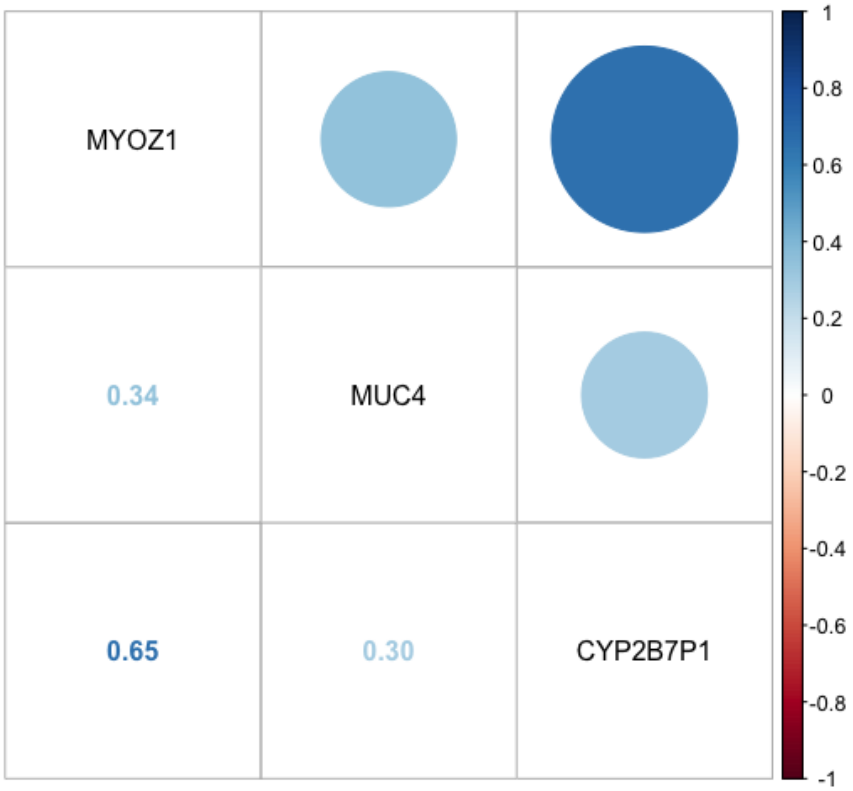


Figure 3.18: *a correlation matrix containing the calculated correlation between the vaginal specific mRNA markers. A high positive correlation is illustrated as a larger dark blue circle, while a negative correlation is illustrated as a smaller red circle.*

3.2.5.2 Statistical properties and prediction for the vaginal samples

12 logistic regression models were created by a train dataset and tested by a test dataset. 9 of the models were univariate model based on each vaginal marker created from df1, df2 and df3. Three multivariate models based on a combination of all markers were created from each dataset and tested with the same random selected samples.

The prediction among the univariate and the multivariate models were mostly similar between all models. The model that gave the best outcome was created from the df3 dataset with the multivariate and the univariable for MUC4. The other models did not identify any of the three vaginal samples from the other samples and had therefore no specificity. The vaginal

samples that did not get detected had no unexpected detected markers, while the vaginal sample that was detected in the multivariate model had two unexpected detected markers. Since the df3 dataset had penalty for the unexpectedly detected markers it could have contributed to a better prediction.

Since we only had three vaginal secretion samples in our test dataset, gave this us difficulty in getting a good overview of the models ability to distinguish the vaginal samples from the other body fluid types.

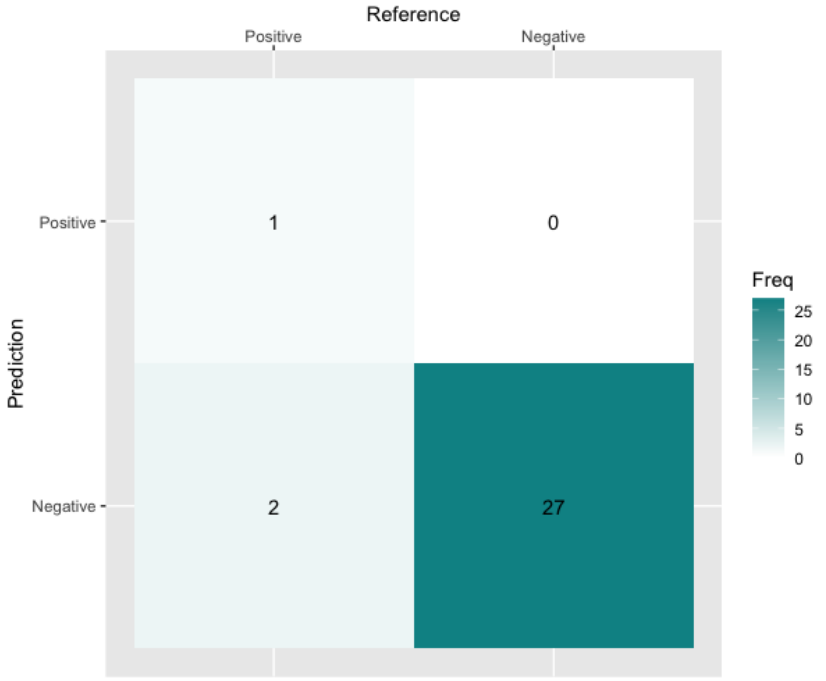


Table 3.10: The accuracy of the prediction is calculated to be 0.93 (93%) with a sensitivity of 0.33 and specificity of 0.6.

Accuracy	0.930
Sensitivity	0.333
Specificity	1.000

Figure 3.19: *From the prediction of vaginal secretion, we only found 1 out of the 3 vaginal secretion samples. Two of the samples did not get detected and ended up as false negative values.*

3.2.6 In-depth analysis of the nasal secretion samples

In this study there was a total number of 12 RNA-profiles, created from the 12 nose secretion samples. 24 housekeeping genes were detected along with 6 female and 3 male sex markers. The remaining 55 markers were markers associated to a specific body fluid where 56% of these markers were unexpectedly detected.

There was no significant difference between the detected mean RFU value or detection rate between the three volume groups (p-value > 0.05). There was, on the other hand, a strong

significant difference between the RFU values and detection rate between each RNA marker that was detected in the nose secretion samples (p-value< 0.01).

The detected RFU values varied between the 9 different detected markers that were found in the nose secretion samples. The highest detected mean values were found among the MUC4, RPS4Y1 and STATH markers with a mean value of 23939.18, 26110.5 and 20571 RFU. Most of the wrong detected markers, except the MUC4 marker, have a relatively low mean RFU value. Figure 3.20 shows a boxplot of the detected peak height for each of the 9 detected markers.

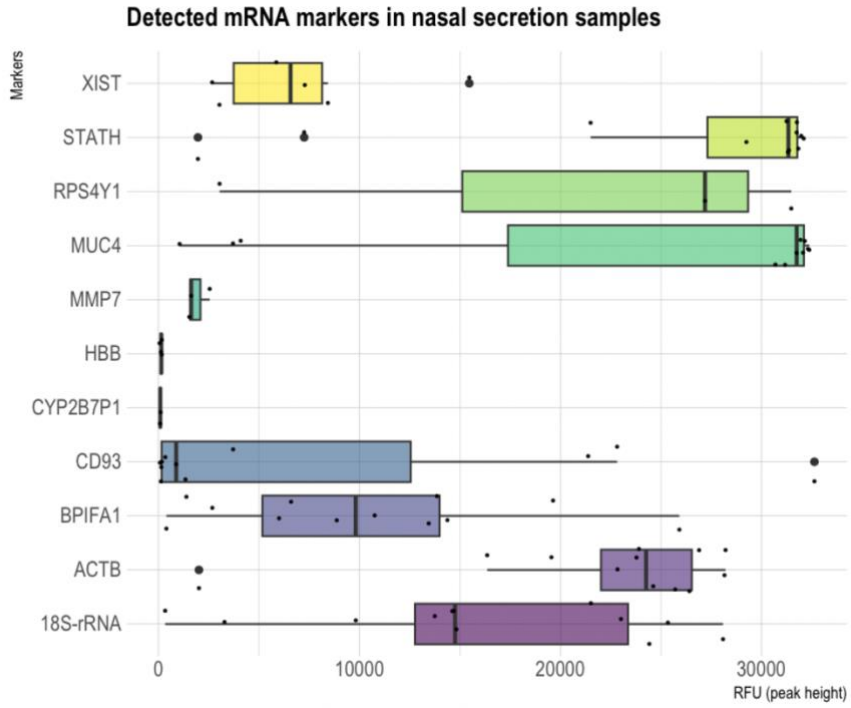


figure 3.20: *a boxplot showing the detected peak height value for all the detected mRNA markers in the nose secretion samples. Each marker can be found along the y-axis, while the RFU value is listed along the x-axis. This plot was created in the R-studio software*.

3.2.6.1 Correlation between the mRNA markers in the nasal secretion samples

We calculated the correlation values between the mRNA markers: STATH and the BPIFA1. The calculations were done and plotted in the R-studio software. From the correlation matrix (Figure 3.21), we can see that the correlation between the two nasal specific markers is quite high with a positive value of 0.67.

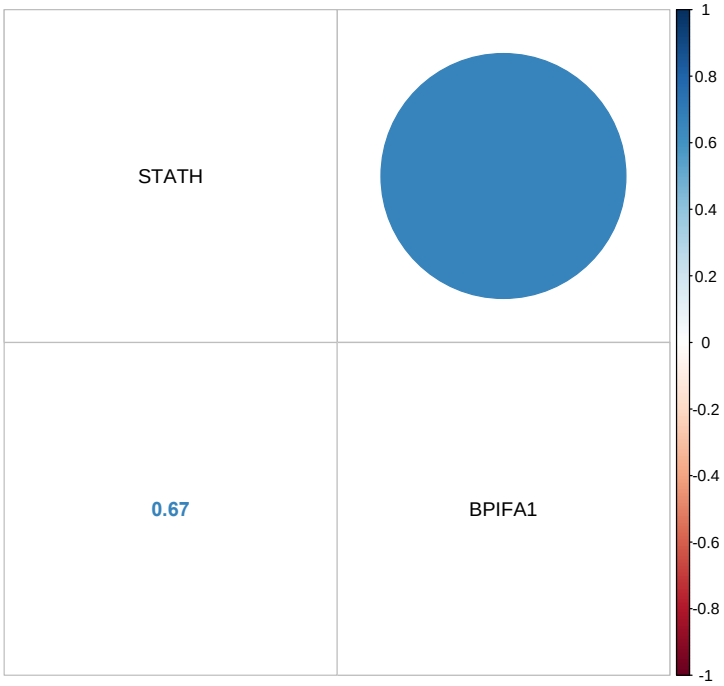


Figure 3.21: *a correlation matrix with the calculated correlation between the STATH and BPIFA1 RFU values. A high value gives a larger circle. Negative correlation gives a red color that increase in color when the value increases. A positive value gives a blue color that gets darker when the value increases.*

3.2.6.2 Statistical properties and prediction for nasal secretion samples

Three multivariate logistic regression models were created from the df1, df2 and df3 datasets. Six univariate models were created with each marker as a predictor variable. A prediction was performed based on a training and a test dataset. Both the model based of the BPIFA1 marker *(lg1.n2, lg2.n2 & lg3.n2)* and the multivariate model *(lg1.n, lg2.n & lg3.n)* were equally good in all the different datasets.

Out of the five nasal samples from the test dataset, only four got detected and categorized as the targeted body fluid. Although this gives a considerably good prediction with 80% correct prediction.

Table 3.11: The accuracy of the prediction is calculated to be 0.80 (80%) with a sensitivity of 0.8 and specificity of 1.0.

Accuracy	0.80
Sensitivity	0.80
Specificity	1.00

Figure 3.22: *The prediction of the nasal samples gave a almost perfect prediction with only one false negative prediction. This gives the prediction an accuracy of 80%.*

3.3 Summary of ANOVA - Detection rate and cDNA volume

3.3.1 cDNA Volume

The six one-ANOVA analysis we performed based on the detected peak height value showed us that there was no significant difference between the volume groups in most of the body fluid types. The only body fluid that showed any significant difference between the mean RFU values in the three volume groups, was values form the blood samples. However, by running a two way-ANOVA with both the volume and marker type as predictor variables, we found out that both the menstruation and blood samples had significant difference in mean RFU value between the different volume groups.

Table 3.12: *A table consisting of the calculated p-value from the one-way ANOVA analysis with the three volume groups as predictor variables and the peak height values as response variable. A p-value less than 0.05 shows that that it is a significant difference between the three volume groups. A value above this threshold indicates no significant difference.*

Body fluid	**Volume group (p-value)**	**Significant**
Blood	0.0379	Yes
Menstruation blood	0.0578	No
Saliva	0.122	No
Semen	0.209	No
Vaginal secretion	0.68	No
Nasal secretion	0.199	No
Total dataset	0.00596	yes

3.3.1 Detection rate

Based on the results of the six one-way ANOVA analyses *performed*, it can be *concluded* that there was no statistically significant difference observed among the volume groups in terms of the detection rate. (See Table 3.1)

Table 3.13: *A table consisting of the calculated p-value from the one-way ANOVA analysis with the volume group as predictor variable and the detection rate as response variable. A p-value less than 0.05 shows that that it is a significant difference between the three volume groups. A value above this threshold indicates no significant difference.*

Body fluid	***Detection rate (p-value)***	***Significant***
Blood	0.328	No
Menstruation blood	0.909	No
Saliva	0.671	No
Semen	0.743	No
Vaginal secretion	0.471	No
Nasal secretion	0.913	No
Total dataset	0.476	No

3.4 Summary of the predicted values

3.4.1 Multivariate logistic regression models

For each dataset (df1, df2, and df3), a separate multivariate logistic regression model was fitted for every body fluid type. The accuracy of the models exhibited variations ranging from 0% to 100% across different body fluids and from 33% to 100% across the various datasets. (See Table 3.14)

Table 3.14: the accuracy calculated from the fitted multivariant models for all three datasets in all six body fluid types.

Sample type	**Df1**	**Df2**	**Df3**
Blood	60%	60%	60%
Menstruation blood	33%	66%	33%
Saliva	64%	64%	91%
Semen	100%	33%	100%
Vaginal secretion	0%	0%	33%
Nasal secretion	80%	80%	80%

3.4.2 The best fitted model for the univariate logistic regression models

The univariate models are fitted based on only one specific marker associated with the target body fluid. We found the mRNA marker that gave the best accuracy in their representative univariate model for each dataset. We can see form table 3.15 that there was no difference in accuracy between the best fitted models among the univariate logistic regression models. There was not much difference between the accuracy of the multivariate models and the best predicted univariate models. The only exception was the multivariate model from dataset df3, that was able to detect one of the vaginal secretion samples, which the other models couldn't do. However, we can see that all the univariate models from the semen samples had an accuracy of 100% while the multivariate model from df2 only had 33% detection accuracy.

Table 3.15: The table displays the univariate models with the highest accuracy among all three datasets. The accuracy is presented as a percentage, and the marker used as the predictor variable in each univariate model is indicated within parentheses. For cases where multiple univariate models show comparable levels of prediction performance, these models are listed within the same set of parentheses.

Sample type	**Df1**	**Df2**	**Df3**
Blood	60% (ALAS2)	60% (ALAS2)	60% (ALAS2/ HBB)
Menstruation blood	100% (MMP10/MMP11)	100% (MMP11)	100% (MMP10/MMP11)
Saliva	64% (HTN3)	64% (HTN3)	64% (HTN3)
Semen	100% (KLK3/PRM1/SEMG1)	100% (KLK3/PRM1/SEMG1)	100% (KLK3/PRM1/SEMG1)
Vaginal secretion	0%	0%	0%
Nasal secretion	80% (BPIFA1)	80% (BPIFA1)	80% (BPIFA1)

4. Discussion

4.1 The datasett

The dataset used in this study consisted of samples from six different body fluids: blood, menstrual blood, saliva, semen, vaginal secretion, and nasal secretion. It comprised a total of 90 samples, with varying sample numbers for each body fluid type, with nasal secretion as the largest sample group with 34 samples. It is worth mentioning that the sample distribution across the different body fluid types was different, with saliva having the largest number of samples.

The dataset size was determined based on the availability of samples within the study's reach. While a larger dataset could provide more statistical power and generalizability, the sample size in this study was determined to be sufficient for the specific research objectives. (56).
It is important to acknowledge that the varying sample numbers across body fluid types can lead to some limitations regarding statistical comparisons and the robustness of the predictive models. The relatively smaller sample sizes for some body fluid types, such as menstrual blood and blood samples, should be considered when interpreting the results.

Future studies could consider expanding the dataset to include a larger number of samples for each body fluid type, enabling more robust statistical analyses and improving the accuracy of the predictive models. Additionally, one can later ensure a more balanced distribution of samples across the different body fluid types, facilitating more comprehensive comparisons and evaluations.

4.2 Volume and RFU values

In this study, we looked at the cDNA volume's effect on marker detection and peak height for different body fluids. We performed a one-way ANOVA to study the relationship between the volume of cDNA (predictor variable) and the peak height (response variable).

When the ANOVA was performed on the entire dataset, the volume of cDNA showed a significant effect on peak height, with a p-value below 0.01, which indicates a strong significant variation among the different volume groups. However, upon further analysis by separating the dataset based on each body fluid type, the significance of the volume as a

predictor was increased to a value above the p-value threshold indicating no significant difference between the groups. The individual ANOVA tests conducted within each body fluid type, except for the blood samples, revealed p-values higher than 0.05, indicating that the volume was not a significant factor in determining peak height within each group. This suggests that the volume of cDNA may have a more pronounced effect on peak height in blood samples than other body fluid types (See table 3.12).

We also performed a two-way ANOVA with both cDNA volume and marker type as predictors to see if there were potential interactions between these variables. The results indicated that in addition to the volume of cDNA, the marker type also played a significant role in determining peak height among the menstruation blood and blood samples. In the menstruation and blood samples the p-value for significant differences among the volume groups decreased to a p-value below 0.05 (see table 44 and 45 in Appendix 2). This suggests that the combination of marker type and cDNA volume may have a joint effect on the observed peak height in the menstruation and blood samples.

In forensic genetics, the cell type and quantity of biological material are often unknown before analysis. To address this, we tested different volumes added to the PCR reaction. We observed variations in peak heights, which depended on the specific cell types. Consequently, recommending an optimal volume for PCR addition is challenging when the cell type is still being determined.

4.3 Volume and Detection Rate

In addition to examining the impact of cDNA volume on peak height, we also investigated its effect on the detection rate of mRNA markers among the samples. We performed a one-way ANOVA to see if the volume of cDNA influenced the detection rate. Surprisingly, our results revealed that the volume of cDNA did not significantly affect the detection rate across all body fluid types, as indicated by a p-value greater than 0.05.

This finding suggests that variations in cDNA volume, within the range tested in our study, did not significantly impact the ability to detect mRNA markers in the samples. These results indicate that the volume of cDNA may not be a critical factor affecting our study's overall detection rate of mRNA markers. However, the samples included in this study were mainly of

good quality, and a larger effect is expected on samples containing lower amounts of degraded RNA.

Another factor one may consider is the detection threshold. The threshold for detection is 50 RFU, which could be considered low. Raising this threshold could lead to a more pronounced influence on marker detection. An increased threshold could also contribute to less detection of the unexpected mRNA markers in different body fluids. This is because most of the mRNA markers that have been in another expected body fluid have a relatively low RFU value. Additionally, other factors, such as RNA quality and extraction efficiency, still play a role in the overall detection rate and should be considered in future studies.

4.4 Detection of sex-specific mRNA markers

Our study aimed to analyze the detection of sex-specific mRNA markers, XIST for females and RPS4Y1 for males, in various body fluid types. We made several interesting observations among the 19 different markers analyzed in our study, including two housekeeping genes, two sex-specific genes, and the remaining mRNA markers exclusively expressed in specific body fluids.

Firstly, we found that both the menstrual and blood samples showed the presence of a sex marker in all their respective samples. This observation suggests a high level of reliability in detecting sex markers within these body fluid types.

The detection of sex markers in other body fluid types varied. The saliva samples showed a detection rate of a sex marker in only 68% of the samples, indicating a lower frequency of occurrence. On the other hand, the semen samples showed a higher detection rate of 92%, suggesting a more reliable identification of male-specific markers in this body fluid type. Similarly, the vaginal secretion samples showed a detection rate of 67%, while the nasal secretion samples showed a detection rate of 75%.

It is worth noting that one vaginal secretion sample demonstrated the unexpected detection of a male-specific marker. This finding highlights the potential for cross-contamination or other sexual activity before sample collection. On the other hand, there were some restrictions regarding the time since sexual activity before collection samples to avoid mixtures.

However, semen could have persevered for a longer time and been detected in one of the samples.

Furthermore, our analysis revealed that more female sex markers were detected compared to male-specific markers. This discrepancy can be attributed to the larger number of female participants in our study, leading to a higher prevalence of female-specific markers across the body fluid types analyzed. This finding underscores the influence of participant demographics on the detection patterns of sex markers.

4.5 The influence of the STATH mRNA marker in saliva and nasal secretion samples

In this study, we investigated the detection of specific mRNA markers in saliva and nasal secretion samples. It was previously reported in a study by Sakurada et al. (30) that the marker STATH, which is commonly associated with saliva samples, can also be detected in nasal secretion samples. So, we explored the possibility of improving the prediction model for nasal samples by including both the BPIFA1 marker, specific to nasal secretion and the STATH marker.

Upon analyzing the results, we observed that the logistic model using only the BPIFA1 marker achieved a relatively high prediction accuracy, correctly identifying 4 out of 5 nasal samples. This suggests that BPIFA1 alone exhibits a strong association with nasal secretion samples and can effectively predict nasal fluid identification.

Surprisingly, when we incorporated the STATH marker into the model, we noticed a decrease in the prediction accuracy, with 3 out of 5 nasal samples correctly identified. This result was unexpected since the inclusion of STATH, which is commonly associated with saliva, was anticipated to enhance the prediction for nasal samples. However, the introduction of STATH led to two false positives and two false negatives, indicating a less accurate prediction compared to the model using only BPIFA1.

We can see from Figure 3.4 that the STATH marker often appear in various body fluids, especially nasal samples. The presence of STATH in nasal samples might represent a contamination or crossover effect from the adjacent oral cavity, leading to false-positive predictions. Additionally, the differential expression patterns or regulation mechanisms of

STATH and BPIFA1 in nasal secretion samples could potentially conflict, resulting in the misclassification of some nasal samples.

These findings highlight the importance of thoroughly investigating marker specificity and considering potential cross-reactivity or overlapping expression patterns when designing prediction models.

4.6 Correlation and prediction

In examining the influence of marker correlation on the logistic regression models, we aimed to determine if markers with high correlation could impact the predictive performance. Within the blood samples, the ALAS2 and HBB markers exhibited the highest correlation. Interestingly, these two markers demonstrated the most accurate predictions among the three fitted univariate models. This suggests that the strong correlation between ALAS2 and HBB may contribute to their combined effectiveness as predictors in the logistic regression model for blood samples.

However, in the menstruation samples, the MMP10 and MMP7 markers displayed the highest correlation. Surprisingly, the univariate logistic regression models based of the MMP10 and MMP11 markers gave better prediction than the multivariate model based of all markers and the univariate model based on MMP7 marker. This might indicate that it could be other factors that can affect the prediction, such as different patterns in the detection of the markers in the menstruation samples.

Furthermore, we investigated the correlation between the STATH and BPIFA1 markers in saliva and nasal secretion samples. Despite a relatively low correlation, we explored whether incorporating both markers would improve the prediction. However, our findings revealed that the inclusion of STATH did not enhance the prediction more than was achieved by solely using the BPIFA1 marker.
Other factors, such as marker specificity, may have a more significant impact on the accuracy of the logistic regression model since the BPIFA1 marker is more specific than the STATH marker.

Our study demonstrates that marker correlation does not consistently dictate the predictive power of logistic regression models. While markers with high correlation, such as ALAS2 and

HBB, may yield improved predictions, other factors, including marker specificity and differential expression patterns, should also be considered. Further studies are necessary to understand the relationships between markers and their impact on the predictive modeling for body fluid identification.

4.7 Specificity among mRNA markers and prediction

The presence of unexpected markers in various body fluid types gives an interesting view of our study. One such marker is MUC4, which is typically expressed exclusively in vaginal secretion but was unexpectedly detected in blood (3%), nasal secretion (20%), and saliva (20%) samples. Similarly, the STATH marker, associated with saliva, appeared in both blood (3%) and menstruation blood (1%) samples. The HBB marker appeared in 11% of vaginal secretion samples, 7% of saliva samples, and 7% of nasal secretion samples. This suggests that HBB, STATH, and MUC4 markers have low specificity and are, therefore, more unreliable in a prediction model.

Detecting these markers in unexpected body fluids highlights potential cross-contamination or trace amounts of these fluids in the samples. For instance, can traces of blood be present in the nose or oral cavities due to the breakage of small blood vessels.

Analyzing the prevalence of unexpectedly detected markers across body fluid types, we observed that vaginal secretion samples had the highest proportion (37%) of unexpected detected markers, followed by saliva samples (43%) and nasal secretion samples (46%). The impact of these unexpected markers on prediction accuracy is notable, as reflected in the performance of the logistic regression models. For instance, the vaginal secretion samples showed the poorest prediction performance, identifying only 1 out of 3 samples from the best-fitted model multivariate model based on the df3 dataset (lg3.v). Similarly, the blood samples displayed suboptimal prediction accuracy, with only 3 out of 5 samples correctly identified. These discrepancies can be attributed to unanticipated markers, which likely influence the performance of the logistic regression models.

Overall, the presence of unexpected markers highlights the complexity of mRNA profiling and the challenges associated with accurately predicting body fluid types. Cross-contamination can contribute to the unexpected detection of markers in various body fluids.

4.8. Multivariate vs. univariate modeling

Since the univariate models only considered a single predictor variable, they were less successful than the multivariate models. Each univariate model focused on a single feature or marker without considering the combined impact of several markers. The multivariate models, on the other hand, included several markers while accounting for their connections and interactions. As a result, the data could be analyzed more thoroughly.

5. Sources

(1) Panneerchelvam, S. & Norazmi, M. N. (2003, July). *Forensic DNA profiling and database*. The Malaysian journal of medical sciences: MJMS.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3561883/#:~:text=DNA%20fingerprinting%20was%20first%20used,had%20not%20committed%20the%20crimes.

(2) Doi et al., M. (2014, May 16). *A simple identification method for vaginal secretions using relative quantification of lactobacillus DNA*. Forensic science international. Genetics. https://pubmed.ncbi.nlm.nih.gov/24905338/

(3) Alexander Lindenbergh et.al. (2012, February 21). *A multiplex (M)RNA-profiling system for the forensic identification of body fluids and contact traces*. Forensic Science International: Genetics. https://www.sciencedirect.com/science/article/pii/S1872497312000336

(4) Minchin, S., & Lodge, J. (2019, October 16). *Understanding biochemistry: Structure and function of Nucleic Acids*. Essays in biochemistry. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6822018/

(5) Lindenbergh et.al, 2. Alexander. (2012, October 1). *Implementation of RNA profiling in forensic casework*. Forensic Science International: Genetics. https://www.sciencedirect.com/science/article/abs/pii/S1872497312001986

(6) Hellen Johannessen et al. (2022, July 21). *Transfer, persistence and recovery of DNA and mrna vaginal mucosa markers after intimate and social contact with bayesian network analysis for activity level reporting*. Forensic Science International: Genetics. https://www.sciencedirect.com/science/article/pii/S1872497322000916

(7) ABC-CLIO. (2006). Introduction . I K. A. William J. Tilstone, *Forensic Science: An Encyclopedia of History, Methods, and Techniques* (ss. 1-2). 130 Cremona Drive Box 1911: ABC-CLIO.

(8) Needham, L. G. (1988). *A HISTORY OF FORENSIC MEDICINE IN CHINA.* Hentet fra Cambridge University Press: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/5D623DAD505D275FFB1AD8529DF2A408/S0025727300048511a.pdf/history_of_forensic_medicine_in_china.pdf

(9) Panneerchelvam, S., & Norazmi, M. N. (2003a, July). *Forensic DNA profiling and database*. The Malaysian journal of medical sciences : MJMS. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3561883/

(10) Pankaj Shrivastava, H. R. (2020). STR Typing and Available Multiplex Kits Including Validation Methods. I *Forensic DNA Typing: Principles, Applications and Advancements* (ss. 27-29). This Springer imprint.

(11) Bioteknologirådet. (2022, November 10). *DNA-analyser I Etterforskning og Rettsvesen*. Bioteknologirådet: https://www.bioteknologiradet.no/temaer/dna-analyser/

(12) Saad, R. (2005, April). *Discovery, development, and current applications of DNA identity testing*. Proceedings (Baylor University. Medical Center). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200713/

(13) Sijen, T., & Harbison, S. (2021, October 28). *On the identification of body fluids and tissues: A crucial link in the investigation and solution of Crime*. Genes.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8617621/#:~:text=Body%20fluid%20and%20tissue%20identification%20can%20add%20evidence%20in%20criminal,and%20the%20activities%20that%20occurred.

(14) Samie, L. (2022, January 4). *Use of bayesian networks for the investigation of the nature of biological material in casework*. Forensic Science International. https://www.sciencedirect.com/science/article/pii/S0379073822000044

(15) Butler, J. M. (2011). Sample Characterization. I J. M. Butler, *Advanced Topics in Forensic DNA Typing: Methodology* (ss. 14-17). National Institute of Standards and Technology.

(16) Vennemann, M. K., & Koppelkamm, A. (2010, August 17). *MRNA profiling in forensic genetics I: Possibilities and limitations*. Forensic Science International. https://www.sciencedirect.com/science/article/pii/S037907381000335X

(17) Cosgrove, M. (1998, November 23). *Nucleotides*. Nutrition. https://www.sciencedirect.com/science/article/pii/S0899900798000756

(18) Houseley, J. & Tollervey, D. (2009, February 21). *The many pathways of RNA degradation*. Cell. https://www.sciencedirect.com/science/article/pii/S0092867409000671

(19) Panni et al., S. (2019, September 4). *Non-coding RNA regulatory networks*. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. https://www.sciencedirect.com/science/article/pii/S1874939919302160

(20) Marcus Gry et al. (2009, August 7). *Correlations between RNA and protein expression profiles in 23 human cell lines - BMC genomics*. BioMed Central. https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-365

(21) WHO. (1993). *Biomarkers and Risk Assessment: Concepts and Principles (1. INTRODUCTION 1.1 Biomarkers - concepts).* Hentet fra INTERNATIONAL PROGRAMME ON CHEMICAL SAFETY: https://www.inchem.org/documents/ehc/ehc/ehc155.htm

(22) Strimbu, K & Tavel, J. A. (2010, November). *What are biomarkers?*. Current opinion in HIV and AIDS. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078627/

(23) Walsh, M. F., Nathanson, K. L., Couch, F. J., & Offit, K. (2016). *Genomic biomarkers for breast cancer risk*. Advances in experimental medicine and biology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5016023/

(24) Ladd-Acosta, C., & Fallin, M. D. (2015, October 27). *The role of epigenetics in genetic and environmental epidemiology* ... The role of epigenetics in genetic and environmental epidemiology. https://www.futuremedicine.com/doi/10.2217/epi.15.102

(25) Julkunen, H., Cichońska, A., Slagboom, P. E., Würtz, P., & Nightingale Health UK Biobank Initiative. (2021, May 4). *Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and covid-19 in the general population*. eLife. https://elifesciences.org/articles/63033

(26) Han, W., & Ye, Y. (2019, March 14). *A repository of microbial marker genes related to human health and diseases for host phenotype prediction using Microbiome Data*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417824/

(27) GeneCards- The Human gene Database. (2023, May 21). *CD93 gene - genecards | C1QR1 protein | C1QR1 antibody*. CD93 Gene - CD93 Molecule. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD93

(28) GeneCards- The Human gene Database. (2023, May 21). *Alas2 gene - genecards | Hem0 protein | Hem0 antibody*. ALAS2 Gene - 5'-Aminolevulinate Synthase 2. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ALAS2

(29) GeneCards- The Human gene Database. (2023c, May 21). *HTN3 gene - genecards | HIS3 protein | HIS3 antibody*. HTN3 Gene - Histatin 3. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HTN3

(30) Sakurada et al., K. (2011, September 20). *Expression of statherin mrna and protein in nasal and vaginal secretions*. Legal Medicine. https://www.sciencedirect.com/science/article/abs/pii/S1344622311000848

(31) GeneCards- The Human gene Database. (2023d, May 22). *Stath gene - genecards | stat protein | stat antibody*. STATH Gene - Statherin. https://www.genecards.org/cgi-bin/carddisp.pl?gene=STATH

(32) GeneCards- The Human gene Database. (2023b, May 21). *BPIFA1 gene - genecards | BPIA1 protein | BPIA1 antibody*. BPIFA1 Gene - BPI Fold Containing Family A Member 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=BPIFA1

(33) GeneCards- The Human gene Database. (2023e, May 22). *SEMG1 gene - genecards | SEMG1 protein | SEMG1 antibody*. SEMG1 Gene - Semenogelin 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SEMG1

(34) GeneCards- The Human gene Database. (2023e, May 22). *PRM1 gene - genecards | HSP1 protein | HSP1 antibody*. PRM1 Gene - Protamine 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PRM1

(35) GeneCards- The Human gene Database. (2023e, May 22). *Klk3 gene - genecards | Klk3 protein | Klk3 antibody*. KLK3 Gene - Kallikrein Related Peptidase 3. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KLK3

(36) Bauer, M., & Patzelt, D. (2007, June 1). *Identification of menstrual blood by real time RT-PCR: Technical improvements and the practical value of negative test results*. Forensic Science International. https://www.sciencedirect.com/science/article/abs/pii/S037907380700161 2?via%3Dihub

(37) Cabral-Pacheco et al., G. A. (2020, December 20). *The roles of matrix metalloproteinases and their inhibitors in human diseases*. International journal of molecular sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7767220/

(38) Akutsu, T. et al. (2017, September 18). *Quantitative evaluation of candidate genes and development of a multiplex RT-PCR assay for the forensic identification of vaginal fluid*. Forensic Science International: Genetics Supplement Series. https://www.sciencedirect.com/science/article/abs/pii/S1875176817301919

(39) GeneCards- The Human gene Database. (2023f, May 22). *Muc4 gene - genecards | Muc4 protein | Muc4 antibody*. MUC4 Gene - Mucin 4, Cell Surface Associated. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MUC4

(40) GeneCards- The Human Gene Database. (2023, May 22). *Myoz1 gene - genecards | MYOZ1 protein | Myoz1 antibody*. MYOZ1 Gene - Myozenin 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYOZ1

(41) Turabelidze, A., Guo, S., & DiPietro, L. A. (2010, August 19). *Importance of housekeeping gene selection for accurate reverse transcription-quantitative polymerase chain reaction in a wound healing model*. Wound repair and regeneration : official publication of the Wound Healing Society [and] the European Tissue Repair Society. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939911/

(42) Thorrez, L., Van Deun, K., Tranchevent, L.-C., Van Lommel, L., Engelen, K., Marchal, K., Moreau, Y., Van Mechelen, I., & Schuit, F. (2008, March 26). *Using ribosomal protein genes as reference: A tale of caution*. PloS one. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2267211/

(43) GeneCards- The Human gene Database. (2023i, May 22). *RNA18S1 gene - genecards | RNA18S1 RNA gene*. RNA18S1 Gene - RNA, 18S Ribosomal 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RNA18S1

(44) GeneCards- The Human gene Database. (2023a, May 21). *ACTB gene - genecards | ACTB protein | ACTB antibody*. ACTB Gene - Actin Beta. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ACTB

(45) GeneCards- The Human gene Database. (2023m, May 22). *Xist gene - genecards | XIST RNA gene*. XIST Gene - X Inactive Specific Transcript. https://www.genecards.org/cgi-bin/carddisp.pl?gene=XIST

(46) GeneCards- The Human gene Database. (2023k, May 22). *RPS4Y1 gene - genecards | RS4Y1 protein | RS4Y1 antibody*. RPS4Y1 Gene - Ribosomal Protein S4 Y-Linked 1. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RPS4Y1

(47) Butler, J. M. (2011). Sample Characterization. I J. M. Butler, *Advanced Topics in Forensic DNA Typing: Methodology* (ss. 14-17). National Institute of Standards and Technology.

(48) Butler, J. M. (2011). Capillary Electrohoresis: Principles and Instrumentation. I J. M. Butler, *Advanced Topics in Forensic DNA Typing: Methodology* (ss. 141-165). National Institute of Standards and Technology.

(49) Salzmann et al., A. P. P. (2021, May 4). *Degradation of human mrna transcripts over time as an indicator of the time since deposition (TSD) in biological crime scene traces*. Forensic Science International: Genetics. https://www.sciencedirect.com/science/article/pii/S1872497321000624#:~:text=In%20summary%2C%20we%20found%20degradation,samples%20were%20difficult%20to%20interpret.

(50) Butler, J. M. (2012). Chapter 4 - PCR Amplification: Capabilities and Cautions. In *Advanced topics in forensic DNA typing: Methodology* (pp. 69–97). essay, Academic Press is an imprint of Elsevier.

(51) Mo, Y., Wan, R., & Zhang, Q. (2012, October 31). *Application of reverse transcription-PCR and real-time PCR in nanotoxicity research*. Methods in molecular biology (Clifton, N.J.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5087796/

(52) Biolabs, N. E. (n.d.). *RT-PCR & cDNA synthesis*. NEB. https://international.neb.com/applications/dna-amplification-pcr-and-qpcr/rt-pcr-and-cdna-synthesis

(53) Butler, J. M. (2011). Capillary Electrohoresis: Principles and Instrumentation. I J. M. Butler, *Advanced Topics in Forensic DNA Typing: Methodology* (ss. 141-165). National Institute of Standards and Technology.

(54) Thermo Fisher Scientific Inc. (2018, June 29). Turbo DNA free kit. Thermo Fisher Scientific Inc. https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2F1907M_turbodnafree_UG.pdf

(55) Thermo Fisher Scientific Inc (2014, April 05). DNA Fragment Analysis by Capillary Electrophoresis. Thermo Fisher inc. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4474504.pdf

(56) Faber, J., & Fonseca, L. M. (2014). *How sample size influences research outcomes*. Dental press journal of orthodontics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296634/

Appendix 1

During the programming in R-studio and R-software, we used the packages: Tidyverse, Corrplot, Ggplot2, Stats and Caret.

1. **The Df1 datsett**

Type	HBB	MYOZ1	MMP10	MMP7	MMP11	CD93	MUC4	CYP2B7P1
Menstruas	31763	2329	32546	32662	942	1888	0	0
Menstruas	17251	192	6416	14520	497	1444	1433	0
Menstruas	17864	1923	2343	0	1010	0	5005	297
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Spytt	0	733	0	0	0	0	875	0
Spytt	0	361	0	0	0	0	0	0
Spytt	0	1692	0	0	0	0	895	0
Spytt	52	0	0	0	0	0	0	0
Spytt	0	0	0	0	0	374	1001	0
Spytt	0	358	0	0	0	0	0	0
Menstruas	31707	3710	31961	32555	514	1781	31761	0
Menstruas	1086	0	542	1490	0	218	137	0
Menstruas	29460	29665	20373	1504	8293	1600	32406	2103
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Spytt	0	1487	0	0	0	0	1720	0
Spytt	0	194	0	0	0	0	146	0
Spytt	0	3947	0	0	0	0	1657	0
Spytt	0	144	0	0	0	0	534	0
Spytt	0	55	0	0	0	0	64	0
Spytt	0	438	0	0	0	0	0	0
Menstruas	31427	3144	32511	32626	1620	3629	31852	0
Menstruas	31722	499	32510	32569	0	17232	16662	135
Menstruas	28762	32534	32355	8982	24406	2759	31932	14399
Sæd	0	171	0	0	0	0	0	0
Sæd	0	66	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	96	0	0	0	0	0	0
Spytt	0	2497	714	0	0	0	2689	0
Spytt	73	225	0	0	0	0	415	0
Spytt	0	15571	0	0	0	0	4486	0
Spytt	0	0	0	0	0	0	0	0
Spytt	0	0	0	0	0	0	1479	0
Spytt	0	0	0	0	0	0	354	0
Blod	31281	0	0	0	0	2002	0	0
Blod	30545	0	0	0	0	696	0	0
Blod	31502	0	0	0	0	1632	0	0
Spytt	0	0	0	0	0	257	2560	0
Spytt	87	4025	0	0	0	0	675	0
Spytt	0	0	0	0	0	0	0	0
Blod	31020	0	0	0	0	7529	0	0
Blod	23977	0	0	0	0	1117	0	0
Blod	22124	0	0	0	0	2647	0	0
Spytt	77	213	0	0	0	0	1837	0
Spytt	0	0	0	0	0	0	0	0
Blod	30848	0	0	0	0	18850	0	0

SEMG1	PRM1	KLK3	STATH	HTN3	BPIFA1	ALAS2
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
340	3942	117	0	0	0	0
50	0	0	0	0	0	0
190	5164	0	0	0	0	0
0	0	0	103	401	0	0
0	0	0	167	297	229	0
0	0	0	0	482	0	0
0	0	0	8739	32541	0	0
0	0	0	4664	23426	0	57
0	0	0	8475	32610	0	0
0	0	0	0	0	434	0
0	0	0	0	0	0	0
0	0	0	0	0	0	1248
1085	4667	0	0	0	0	0
323	2972	67	0	0	0	0
116	1674	0	0	0	0	0
0	0	0	677	1170	0	0
0	0	0	914	0	0	0
0	0	0	330	729	0	0
0	0	0	4420	20883	0	0
0	0	0	6936	32314	0	0
0	0	0	7007	32144	0	0
0	0	0	403	0	723	0
0	0	0	0	0	0	1178
0	0	0	0	0	0	9129
1560	19483	77	0	0	0	0
611	8576	221	0	0	0	0
906	30627	0	0	0	0	0
519	11648	0	0	0	0	0
0	0	0	2123	3678	0	0
0	0	0	2257	5327	0	0
0	0	0	4446	4394	0	0
0	0	0	31362	32194	0	0
0	0	0	19145	32477	0	0
0	0	0	15589	32539	0	0
0	0	0	0	0	0	16948
0	0	0	0	0	0	20734
0	0	0	0	0	0	7802
0	0	0	599	1946	0	0
0	0	0	7804	32372	0	0
0	0	0	907	0	0	0
0	0	0	0	0	0	31957
0	0	0	0	0	0	14068
0	0	0	0	0	0	7790
0	0	0	355	473	0	0
0	0	0	4590	14618	0	0
0	0	0	0	0	0	31771

2. Df2 datasett

Type	HBB	MYOZ1	MMP10	MMP7	MMP11	CD93	MUC4	CYP2B7P1
Menstruas	20	2	20	20	1	2	1	1
Menstruas	11	1	4	9	1	1	1	1
Menstruas	12	2	2	1	1	1	4	1
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	2	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Menstruas	20	3	20	20	1	2	20	1
Menstruas	1	1	1	1	1	1	1	1
Menstruas	19	19	13	1	7	1	20	3
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	2	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	3	1	1	1	1	2	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Menstruas	20	2	20	20	2	3	20	1
Menstruas	20	1	20	20	1	11	11	1
Menstruas	19	20	20	6	20	2	20	15
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Sæd	1	1	1	1	1	1	1	1
Spytt	1	2	1	1	1	1	2	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	10	1	1	1	1	3	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Blod	20	1	1	1	1	2	1	1
Blod	20	1	1	1	1	1	1	1
Blod	20	1	1	1	1	2	1	1
Spytt	1	1	1	1	1	1	2	1
Spytt	1	3	1	1	1	1	1	1
Spytt	1	1	1	1	1	1	1	1
Blod	20	1	1	1	1	5	1	1
Blod	16	1	1	1	1	1	1	1
Blod	14	1	1	1	1	2	1	1
Spytt	1	1	1	1	1	1	2	1
Spytt	1	1	1	1	1	1	1	1
Blod	20	1	1	1	1	12	1	1

SEMG1	PRM1	KLK3	STATH	HTN3	BPIFA1	ALAS2
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
3	3	11	1	1	1	1
1	1	1	1	1	1	1
2	4	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	6	20	1	1
1	1	1	3	15	1	1
1	1	1	6	20	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
9	4	1	1	1	1	1
3	2	7	1	1	1	1
1	2	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	3	13	1	1
1	1	1	5	20	1	1
1	1	1	5	20	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	6
13	13	7	1	1	1	1
5	6	20	1	1	1	1
8	20	1	1	1	1	1
5	8	1	1	1	1	1
1	1	1	2	3	1	1
1	1	1	2	4	1	1
1	1	1	3	3	1	1
1	1	1	20	20	1	1
1	1	1	12	20	1	1
1	1	1	10	20	1	1
1	1	1	1	1	1	11
1	1	1	1	1	1	13
1	1	1	1	1	1	5
1	1	1	1	2	1	1
1	1	1	5	20	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	20
1	1	1	1	1	1	9
1	1	1	1	1	1	5
1	1	1	1	1	1	1
1	1	1	3	9	1	1
1	1	1	1	1	1	20

3. Df3

Type	HBB	MYOZ1	MMP10	MMP7	MMP11	CD93	MUC4	CYP2B7P1
Menstruas	20	2	20	20	1	2	0	0
Menstruas	11	1	4	9	1	1	1	0
Menstruas	12	2	2	0	1	0	4	1
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Spytt	0	-1	0	0	0	0	-1	0
Spytt	0	-1	0	0	0	0	0	0
Spytt	0	-2	0	0	0	0	-1	0
Spytt	-1	0	0	0	0	0	0	0
Spytt	0	0	0	0	0	-1	-1	0
Spytt	0	-1	0	0	0	0	0	0
Menstruas	20	3	20	20	1	2	20	0
Menstruas	1	0	1	1	0	1	1	0
Menstruas	19	19	13	1	7	1	20	3
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Spytt	0	-1	0	0	0	0	-2	0
Spytt	0	-1	0	0	0	0	-1	0
Spytt	0	-3	0	0	0	0	-2	0
Spytt	0	-1	0	0	0	0	-1	0
Spytt	0	-1	0	0	0	0	-1	0
Spytt	0	-1	0	0	0	0	0	0
Menstruas	20	2	20	20	2	3	20	0
Menstruas	20	1	20	20	0	11	11	1
Menstruas	19	20	20	6	20	2	20	15
Sæd	0	-1	0	0	0	0	0	0
Sæd	0	-1	0	0	0	0	0	0
Sæd	0	0	0	0	0	0	0	0
Sæd	0	-1	0	0	0	0	0	0
Spytt	0	-2	-1	0	0	0	-2	0
Spytt	-1	-1	0	0	0	0	-1	0
Spytt	0	-10	0	0	0	0	-3	0
Spytt	0	0	0	0	0	0	0	0
Spytt	0	0	0	0	0	0	-1	0
Spytt	0	0	0	0	0	0	-1	0
Blod	20	0	0	0	0	2	0	0
Blod	20	0	0	0	0	1	0	0
Blod	20	0	0	0	0	2	0	0
Spytt	0	0	0	0	0	-1	-2	0
Spytt	-1	-3	0	0	0	0	-1	0
Spytt	0	0	0	0	0	0	0	0
Blod	20	0	0	0	0	5	0	0
Blod	16	0	0	0	0	1	0	0
Blod	14	0	0	0	0	2	0	0
Spytt	-1	-1	0	0	0	0	-2	0
Spytt	0	0	0	0	0	0	0	0
Blod	20	0	0	0	0	12	0	0

SEMG1	PRM1	KLK3	STATH	HTN3	BPIFA1	ALAS2
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
3	3	11	0	0	0	0
1	0	0	0	0	0	0
2	4	0	0	0	0	0
0	0	0	1	1	0	0
0	0	0	1	1	-1	0
0	0	0	0	1	0	0
0	0	0	6	20	0	0
0	0	0	3	15	0	-1
0	0	0	6	20	0	0
0	0	0	0	0	-1	0
0	0	0	0	0	0	0
0	0	0	0	0	0	1
9	4	0	0	0	0	0
3	2	7	0	0	0	0
1	2	0	0	0	0	0
0	0	0	1	1	0	0
0	0	0	1	0	0	0
0	0	0	1	1	0	0
0	0	0	3	13	0	0
0	0	0	5	20	0	0
0	0	0	5	20	0	0
0	0	0	-1	0	-1	0
0	0	0	0	0	0	1
0	0	0	0	0	0	6
13	13	7	0	0	0	0
5	6	20	0	0	0	0
8	20	0	0	0	0	0
5	8	0	0	0	0	0
0	0	0	2	3	0	0
0	0	0	2	4	0	0
0	0	0	3	3	0	0
0	0	0	20	20	0	0
0	0	0	12	20	0	0
0	0	0	10	20	0	0
0	0	0	0	0	0	11
0	0	0	0	0	0	13
0	0	0	0	0	0	5
0	0	0	1	2	0	0
0	0	0	5	20	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	20
0	0	0	0	0	0	9
0	0	0	0	0	0	5
0	0	0	1	1	0	0
0	0	0	3	9	0	0
0	0	0	0	0	0	20

4. Code for transformation of dataset

```
table.rmarker<- function(data.file){
  data<- data.file %>%
    na.omit() %>%
    filter(Allele != "Housekeeping") %>%
    filter(Marker!="XIST") %>%
    filter(Marker!="RPS4Y1")

  samples =unique(data$S_Name)
  markers = unique(data$Marker)
  Type=c()
  dat = NULL
  for(sample in samples){
    sub= subset(data, S_Name==sample, select= c(Type, Height, Marker))
    newrow= sub$Height[match(markers,sub$Marker)]
    Type= c(sub$Type[1],Type)
    dat=rbind(dat,newrow)
  }

  rownames(dat)<-samples
  Type= data$Type[match(samples,data$S_Name)]
  dat <- cbind(Type,dat)
  colnames(dat) = c("Type",markers)
  dat[is.na(dat)] <- 0
  dat <- as.data.frame(dat)
  dat[,2:ncol(dat)] <-as.data.frame(sapply(dat[,2:ncol(dat)], as.numeric))
  return(dat)
}
```

5. Code for calculation of detection rate

```
marker.dist= function(datafile){

  df1<- datafile

  df2 <- datafile %>%

    na.omit()


  num.marker.tot <- as.data.frame.matrix(table(df1$S_type,df1$Marker)) # table of the total amount of all the different markers in the datafile

  num.marker <- as.data.frame.matrix(table(df2$S_type,df2$Marker)) # table of the total amount of all the different markers that were detected


  data.frame.marker <- round(as.data.frame.matrix(num.marker/num.marker.tot),3)


  return(data.frame.marker)


}
```

6. Code figure 3.1

```
data <- RNA.data %>%

  filter(`Allele` != "Housekeeping") %>%

  na.omit()


t.data <- data %>%

  dplyr::select(Marker, Volum)


tbl.t.data <- as.data.frame(table(t.data))


row_sub = apply(tbl.t.data, 1, function(row) all(row !=0 ))

tbl.t.data <- tbl.t.data[row_sub,]




tbl.t.data <- within(tbl.t.data, tbl.t.data$per  <- ifelse(Volum=="0.5", tbl.t.data$Freq/27,
```

```
                                    ifelse(Volum=="1", tbl.t.data$Freq/34,
                                        ifelse(Volum=="3", tbl.t.data$Freq/28, 0))))
tbl.t.data <- tbl.t.data$tbl.t.data
tbl.t.data <- tbl.t.data %>%
  mutate(x_pct= per*100)



# plot


my_colors <- c("#f7fbff", "#e0ecf4", "#bfd3e6", "#9ebcda", "#8c96c6", "#8c6bb1", "#88419d",
"#810f7c",
            "#8c2d04", "#b35806", "#e08214", "#fdb863", "#fee0b6", "#f7f7f7", "#d8daeb", "#b2abd2",
            "#8073ac", "#542788", "#2d004b")


data.bar <- ggplot(tbl.t.data, aes(x = as.character(Volum), y = per, fill = Marker)) +
  ggtitle("Percentage of detected mRNA markers in volume group: 0.5mL, 1.0 mL & 3.0mL") +
  xlab("Volume of cDNA (micro L.)") +
  ylab("Percentage of detected mRNA markers") +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = setNames(my_colors, unique(tbl.t.data$Marker))) +
  scale_y_continuous(labels = scales::percent, name = "Percentage of detected mRNA markers") +
  theme(axis.text.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  coord_flip()
```

7. Code for figure 3.2 and 3.3 (we use blood as example) (3.2)

```
data <- RNA.data %>%
  filter(`Allele` != "Housekeeping") %>%
  filter(S_Name != "03-P3MRTPLUS") %>%
  na.omit()
```

```
#Blood
b.data <- data %>%
  filter(Type=="Blod") %>%
   dplyr::select(Marker)

b.data$sex.marker <- ifelse(b.data$Marker=="XIST" | b.data$Marker=="RPS4Y1", b.data$Marker,
"No sex-marker")

b.data <- as.data.frame(table(b.data$sex.marker))

b.sex.chart <- ggplot(b.data, aes(x="", y=Freq, fill=Var1)) +
 geom_bar(stat="identity", width=1) +
 coord_polar("y", start=0) +
 theme(panel.background = element_rect(fill = "white"))+
 scale_fill_manual(values = c("#e0ecf4", "#8c96c6","#f7f7f7")) +
 labs(title= "Blood",x= NULL, y= NULL, fill= "Type of mRNA marker") +
 geom_text(aes(label = paste0(round((Freq/sum(b.data$Freq))*100), "%")),
           position = position_stack(vjust = 0.5))
```

(3.3)

```
b.data <- data %>%
 filter(Type == "Blod") %>%
 dplyr::select(Marker)

b.data$sex.marker <- ifelse(b.data$Marker == "XIST" | b.data$Marker == "RPS4Y1", "Sex-marker",
"No sex-marker")

b.data <- as.data.frame(table(b.data$sex.marker))
b.data <- b.data[2,]

b.data <- b.data %>%
```

```
  add_row(Var1 = "Total", Freq = 11)

new_value <- b.data$Freq[2] - b.data$Freq[1]

b.data <- b.data %>%
  add_row(Var1 = "nonsex", Freq = new_value) %>%
  filter(Var1 != "Total")

bp.sex.chart <- ggplot(b.data, aes(x = "", y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  theme(panel.background = element_rect(fill = "white")) +
  scale_fill_manual(values = c("#8c96c6", "#e0ecf4"),
                    labels = c("Not sex marker","Sex marker")) +
  labs(title = "Blood", x = NULL, y = NULL, fill = NULL) +
  geom_text(aes(label = paste0(round((Freq / sum(Freq)) * 100), "%")),
            position = position_stack(vjust = 0.5))
```

8. Code for figure 3.4

```
my_colors <- c("#f7fbff", "#e0ecf4", "#bfd3e6", "#9ebcda", "#8c96c6", "#8c6bb1",
               "#88419d", "#810f7c", "#8c2d04", "#b35806", "#e08214", "#fdb863",
               "#fee0b6", "#f7f7f7", "#d8daeb", "#b2abd2","#8073ac", "#542788", "#2d004b")


data <- RNA.data %>%
  filter(`Allele` != "Housekeeping") %>%
  filter(`Allele` != "Female") %>%
```

```
  filter(`Allele` != "Male") %>%
  na.omit()



#Blood
b.data <- data %>%
  filter(Type=="Blod") %>%
  dplyr::select(Marker)


b.data$corr.marker <- ifelse(b.data$Marker=="HBB" |
b.data$Marker=="ALAS2"|b.data$Marker=="CD93", "Correct", b.data$Marker)


b.data <- as.data.frame(table(b.data$corr.marker))


b.chart <- ggplot(b.data, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme(panel.background = element_rect(fill = "white"))+
  scale_fill_manual(values = c("#e0ecf4", "#8c96c6","#f7f7f7","#8c6bb1")) +
  labs(title= "Blood",x= NULL, y= NULL, fill= "Detection of mRNA marker") +
  geom_text(aes(label = paste0(round((Freq/sum(b.data$Freq))*100), "%")),
            position = position_stack(vjust = 0.5))
```

9. Code for box plots (results) – we use blood as example

```
B.data <- data %>%
  filter(S_type=="Blod") %>%
  dplyr::select(Height, Marker, Volum)


B.data.Box <- B.data %>%
  ggplot( aes(x= Marker, y= Height, fill= Marker)) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6) +
```

```
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=14)
  ) +
  ggtitle("Detected mRNA markers in blood samples") +
  xlab("mRNA Markers") +
  ylab("Peak Height") +
  coord_flip()
```

10. Code for correlation plots (results) – we use blood as example

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
 usr <- par("usr")
 on.exit(par(usr))
 par(usr = c(0, 1, 0, 1))
 Cor <- abs(cor(x, y))
 txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])
 if(missing(cex.cor)) {
  cex.cor <- 0.4 / strwidth(txt)
 }

 text(0.5, 0.5, txt,
      cex = 1 + cex.cor * Cor)
}

corrplot.mixed(cor(data[,c(2,7,16)]),  # collumns with the blood markers
               lower = "number",
               upper = "circle",
               tl.col = "black")
```

11. Code for prediction of body fluids (blood as example with dataset 1 df1)

```
org.data<- table.rmarker(RNA.data)

# model fitting
df1$Type <- ifelse(df1$Type=="Blod", 1,0)
df1$Type <- as.factor(df1$Type)

df2$Type <- ifelse(df2$Type=="Blod", 1,0)
df2$Type <- as.factor(df2$Type)

df3$Type <- ifelse(df3$Type=="Blod", 1,0)
df3$Type <- as.factor(df3$Type)



#make this example reproducible by defining a seed number
set.seed(123)

#create train/test data with 2/3 train and 1/3 test data
df1$id <- 1:nrow(df1)
df2$id <- 1:nrow(df2)
df3$id <- 1:nrow(df3)

#train dataset
train.df1 <- df1 %>% dplyr::sample_frac(0.66)
idx.train <- list(train.df1$id)
idx.train <- idx.train[[1]]
train.df2 <- df2[idx.train,]
train.df3 <- df3[idx.train,]
```

```
#test dataset
test.df1  <- dplyr::anti_join(df1, train.df1, by = 'id')
idx.test <- list(test.df1$id)
idx.test <- idx.test[[1]]
test.df2  <- df2[idx.test,]
test.df3  <- df3[idx.test,]


#we find the response
idx.response <- match(rownames(test.df1), rownames(df1))


B.types <- as.data.frame(org.data[idx.response,1])
B.types$Classification <- ifelse(B.types$`org.data[idx.response, 1]` =="Blod",1,0)
B.types$Classification <- as.factor(B.types$Classification)


test.df1 <- test.df1 %>%
  select(-c(Type,id))
test.df2 <- test.df2 %>%
  select(-c(Type,id))
test.df3 <- test.df3 %>%
  select(-c(Type,id))

train.df1 <- train.df1 %>%
  select(-id)
train.df2 <- train.df2 %>%
  select(-id)
train.df3 <- train.df3 %>%
  select(-id)
```

```
#model1- df1
lg1.b1<- glm(Type~HBB, data= train.df1, family = binomial)
lg1.b2<- glm(Type~ALAS2, data= train.df1, family = binomial)
lg1.b3<- glm(Type~CD93, data= train.df1, family = binomial)
lg1.b<- glm(Type~HBB+ALAS2+CD93, data= train.df1, family = binomial)


#model2- df2
lg2.b1<- glm(Type~HBB, data= train.df2, family = binomial)
lg2.b2<- glm(Type~ALAS2, data= train.df2, family = binomial)
lg2.b3<- glm(Type~CD93, data= train.df2, family = binomial)
lg2.b<- glm(Type~HBB+ALAS2+CD93, data= train.df2, family = binomial)


#model3- df3
lg3.b1<- glm(Type~HBB, data= train.df3, family = binomial)
lg3.b2<- glm(Type~ALAS2, data= train.df3, family = binomial)
lg3.b3<- glm(Type~CD93, data= train.df3, family = binomial)
lg3.b<- glm(Type~HBB+ALAS2+CD93, data= train.df3, family = binomial)


#prediction model1- df1
pred.lg1.b1 <- predict(lg1.b1,test.df1, type="response")
pred.lg1.b2 <- predict(lg1.b2,test.df1, type="response")
pred.lg1.b3 <- predict(lg1.b3,test.df1, type="response")
pred.lg1.b <- predict(lg1.b,test.df1, type="response")

#prediction model2- df2
pred.lg2.b1 <- predict(lg2.b1,test.df2, type="response")
```

```
pred.lg2.b2 <- predict(lg2.b2,test.df2, type="response")
pred.lg2.b3 <- predict(lg2.b3,test.df2, type="response")
pred.lg2.b <- predict(lg2.b,test.df2, type="response")

#prediction model3- df3
pred.lg3.b1 <- predict(lg3.b1,test.df3, type="response")
pred.lg3.b2 <- predict(lg3.b2,test.df3, type="response")
pred.lg3.b3 <- predict(lg3.b3,test.df3, type="response")
pred.lg3.b <- predict(lg3.b,test.df3, type="response")


# we categorize the different values in df1

df1.b1.pred.tbl <- data.frame(Blood = pred.lg1.b1 ,
                        not.Blood = 1 - pred.lg1.b1,
                        Classification = if_else(pred.lg1.b1  > 0.5, 1, 0)) %>%
  mutate(Classification = factor(Classification, levels = c(0, 1)),
         org.data= B.types$`org.data[idx.response, 1]`)



df1.b2.pred.tbl <- data.frame(Blood = pred.lg1.b2 ,
                        not.Blood = 1 - pred.lg1.b2,
                        Classification = if_else(pred.lg1.b2  > 0.5, 1, 0)) %>%
  mutate(Classification = factor(Classification, levels = c(0, 1)),
         org.data= B.types$`org.data[idx.response, 1]`)

df1.b3.pred.tbl <- data.frame(Blood = pred.lg1.b3 ,
                        not.Blood = 1 - pred.lg1.b3,
                        Classification = if_else(pred.lg1.b3  > 0.5, 1, 0)) %>%
  mutate(Classification = factor(Classification, levels = c(0, 1)),
         org.data= B.types$`org.data[idx.response, 1]`)
```

```
df1.b.pred.tbl <- data.frame(Blood = pred.lg1.b ,
                             not.Blood = 1 - pred.lg1.b,
                             Classification = if_else(pred.lg1.b  > 0.5, 1, 0)) %>%
  mutate(Classification = factor(Classification, levels = c(0, 1)),
         org.data= B.types$`org.data[idx.response, 1]`)
```

```
lg1.b1.conf <- confusionMatrix( data =df1.b1.pred.tbl$Classification,
reference=B.types$Classification)

lg1.b2.conf <- confusionMatrix( data =df1.b2.pred.tbl$Classification,
reference=B.types$Classification)

lg1.b3.conf <- confusionMatrix( data =df1.b3.pred.tbl$Classification,
reference=B.types$Classification)

lg1.b.conf <- confusionMatrix( data =df1.b.pred.tbl$Classification, reference=B.types$Classification)
```

Plotting confusion matrix:

```
cm <- confusionMatrix(factor(B.types$Classification), factor(df1.b.pred.tbl$Classification ), dnn =
c("Prediction", "Reference"))

plt <- as.data.frame(cm$table)

plt$Prediction <- factor(plt$Prediction, levels=rev(levels(plt$Prediction)))

pt <- ggplot(plt, aes(Prediction,Reference, fill= Freq)) +
  geom_tile() + geom_text(aes(label=Freq)) +
  scale_fill_gradient(low="white", high="#009194") +
  labs(x = "Reference",y = "Prediction") +
  scale_x_discrete(labels=c("Positive","Negative"),position="top") +
```

```
  scale_y_discrete(labels=c("Negative","Positive"))


plot(pt)
```

APPENDIX 2

Table 1-6: *a summary of the model system structure for the univariable and multivariable logistic models fitted with data frame df1, df2 and df3 for marker1 +…+ marker4. The respective model names can be found in the R codes in Appendix 1 in file 'logistic regression models.R'. The same names will be listed in the summary for the logistic regression models bellow.*

Table 1: *Names for the logistic regression models for blood.*

	Univariable			**multivariable**
	HBB	ALAS2	CD93	HBB + …+ CD93
Original data (df1)	lg1.b1	lg1.b2	lg1.b3	lg1.b
Quantile data (df2)	lg2.b1	lg2.b2	lg2.b3	lg2.b
Quantile data (*penalty for unexpected detection)* (df3)	lg3.b1	lg3.b2	lg3.b3	lg3.b

Table 2: *Names for the logistic regression models for the menstruation blood samples*

	Univariable				**multivariable**
	HBB	MMP7	MMP10	MMP11	HBB + …+ MMP11
Original data (df1)	lg1.m1	lg1.m2	lg1.m3	lg1.m4	lg1.m
Quantile data (df2)	lg2.m1	lg2.m2	lg2.m3	lg2.m4	lg2.m
Quantile data (*penalty for unexpected detection)* (df3)	lg3.m1	lg3.m2	lg3.m3	lg3.m4	lg3.m

Table 3: *Names for the logistic regression models for the saliva samples*

	Univariable		**multivariable**
	HTN3	STATH	HTN3 + STATH
Original data (df1)	lg1.s1	lg1.s2	lg1.s
Quantile data (df2)	lg2.s1	lg2.s2	lg2.s
Quantile data (*penalty for unexpected detection)* (df3)	lg3.s1	lg3.s2	lg3.s

***Table 4**: Names for the logistic regression models for the semen samples*

	Univariable			**multivariable**
	KLK3	PRM1	SEMG1	KLK3 +…+ SEMG1
Original data (df1)	lg1.se1	lg1.se2	lg1.se3	lg1.se
Quantile data (df2)	lg2.se1	lg2.se2	lg2.se3	lg2.se
Quantile data (*penalty for unexpected detection)* (df3)	lg3.se1	lg3.se2	lg3.se3	lg3.se

***Table 5**: Names for the logistic regression models for the vaginal secretion samples*

	Univariable			**multivariable**
	MUC4	MYOZ1	CYP2B7P1	MUC4 + …+ CYP2B7P1
Original data (df1)	lg1.v1	lg1.v2	lg1.v3	lg1.v
Quantile data (df2)	lg2.v1	lg2.v2	lg2.v3	lg2.v
Quantile data (*penalty for unexpected detection)* (df3)	lg3.v1	lg3.v2	lg3.v3	lg3.v

***Table 6:** Names for the nasal secretion samples*

	Univariable		**multivariable**
	STATH	BPIFA1	STATH+ BPIFA1
Original data (df1)	lg1.n1	lg1.n2	lg1.n
Quantile data (df2)	lg2.n1	lg2.n2	lg2.n
Quantile data (*penalty for unexpected detection)* (df3)	lg3.n1	lg3.n2	lg3.n

Logistic regression models

The blood samples

***Table 7:** Logistic regression summary of the original data from the blood samples (df1)*

	Univariate		**Multivariate**	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HBB *(lg1.b1)*	1.635e-04	3.56e-05	5.771e-05	0.414118

ALAS *(lg1.b2)*	0.0005862	0.00116	4.533e-04	0.034822
CD93 *(lg1.b3)*	9.022e-05	0.0198	-7.509e-06	0.967379

Table 8: *logistic regression summary of the original data from the blood samples (df2)*

	Univariate		**Multivariate**	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HBB *(lg2.b1)*	1.3122	0.001303	0.827448	0.19826
ALAS *(lg2.b2)*	2.896	0.00495	2.011363	0.06512
CD93 *(lg2.b3)*	0.6967	0.001505	-0.007026	0.99038

Table 9: *logistic regression summary of models created from the blood samples. the quantile data set with registrated penalty for unexpected detection from the blood samples (df3)*

	Univariate		**Multivariate**	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HBB *(lg3.b1)*	0.6846	0.0185	0.1567	0.7359
ALAS *(lg3.b2)*	0.6008	7.82e-06	0.3254	0.0317
CD93 *(lg3.b3)*	0.6657	0.00340	0.2982	0.5184

The menstruation blood samples

Table 10: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		**Multivariant**	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
MMP7 *(lg1.m1)*	0.0003432	0.00492	3.429e-04	0.3961
MMP10 *(lg1.m2)*	0.008951	0.03440	9.936e-03	0.1006
MMP11 *(lg1.m3)*	0.010390	0.00219	-8.103e-03	0.9074
HBB *(lg1.m4)*	1.145e-04	0.000265	2.149e-05	0.9354

Table 11: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		**Multivariant**	

Variable (x,..xn)	b-value	p-value	b-value	p-value
MMP7 *(lg2.m1)*	0.7398	0.003650	6.194e-02	1.000
MMP10 *(lg2.m2)*	22.86	0.997	5.914e+01	0.997
MMP11 *(lg2.m3)*	0.5400	0.006320	-7.320e+00	0.998
HBB *(lg2.m4)*	0.9966	0.00316	6.289e+00	0.999

Table 12: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
MMP7 *(lg3.m1)*	2.580	0.996	8.546e-02	1.000
MMP10 *(lg3.m2)*	5.594	1.000	5.517e+00	1.000
MMP11 *(lg3.m3)*	2.3316	0.996	-1.456e-01	1.000
HBB *(lg3.m4)*	0.5518	0.00861	1.723e-02	1.000

The Saliva samples

Table 13: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HTN3 *(lg1.s1)*	0.0003004	0.0201	3.066e-04	0.0149
STATH *(lg1.s2)*	2.805e-05	0.14815	-4.118e-05	0.3282

Table 14: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HTN3 *(lg2.s1)*	1.1631	2.84e-06	1.17677	4.04e-06
STATH *(lg2.s2)*	0.4119	4.31e-05	-0.03674	0.826

Table 15: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HTN3 *(lg3.s1)*	3.0412	0.993	3.1945	0.995525
STATH *(lg3.s2)*	0.38579	5.47e-07	0.2683	0.071573

The semen samples

Table 16: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
SEMG1 *(lg1.se1)*	0.844	0.997	8.899e-01	0.998
PRM1 *(lg1.se2)*	0.01275	0.996	-2.391e-02	0.999
KLK3 *(lg1.se3)*	0.2700	0.991	-1.326e+00	0.998

Table 17: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
SEMG1 *(lg2.se1)*	20.68	0.996	22.4379	0.998
PRM1 *(lg2.se2)*	0.9879	0.000742	-0.2332	0.918
KLK3 *(lg2.se3)*	0.06696	0.54254	-0.6466	0.221

Table 18: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
SEMG1 *(lg3.se1)*	5.764	1.000	22.4379	0.998
PRM1 *(lg3.se2)*	2.4239	0.994	-0.2332	0.918

KLK3 *(lg3.se3)*	1.9830	0.992	-0.6466	0.221

The vaginal secretion samples

Table 19: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
MUC4 *(lg1.v1)*	9.807e-05	0.000381	9.740e-05	0.000711
MYOZ1 *(lg1.v2)*	5.364e-05	0.219	-7.955e-05	0.329567
CYP2B7P1 *(lg1.v3)*	1.812e-04	0.0512	1.825e-04	0.208061

Table 20: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
MUC4 *(lg2.v1)*	0.16240	0.000393	0.16175	0.04765
MYOZ1 *(lg2.v2)*	0.07300	0.333	-0.16039	0.278040
CYP2B7P1 *(lg2.v3)*	0.18051	0.0531	0.20009	0.194948

Table 21: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
MUC4 *(lg3.v1)*	0.21827	4.19e-06	0.28756	1.81e-05
MYOZ1 *(lg3.v2)*	0.13944	0.047	-0.26516	0.0696
CYP2B7P1 *(lg3.v3)*	0.18916	0.0405	0.13410	0.3735

The nasal secretion samples

Table 22: *logistic regression summary of the original data from the blood samples (df1)*

Univariant	Multivariant

Variable (x,..xn)	b-value	p-value	b-value	p-value
HTN3 *(lg1.n1)*	-0.04908	0.993	-0.046644	0.99670
BPIFA1 *(lg1.n2)*	0.007109	0.035793	0.006411	0.04884

Table 23: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HTN3 *(lg2.n1)*	-14.99	0.995	-14.65	0.998
BPIFA1 *(lg2.n2)*	21.69	0.995	23.31	0.998

Table 24: *logistic regression summary of the original data from the blood samples (df1)*

	Univariant		Multivariant	
Variable (x,..xn)	**b-value**	**p-value**	**b-value**	**p-value**
HTN3 *(lg3.n1)*	-0.15835	0.104	-0.1075	1.000
BPIFA1 *(lg3.n2)*	44.24	0.002	44.0505	0.998

Prediction of the models

The blood samples

Table 25: Results from the prediction based of the Df1 dataset

	Univariate			**Multivariate**
	HBB	ALAS2	CD93	HBB+ALAS2+CD93
True positives	1	3	0	3
True Negatives	24	25	23	25
False positives	1	0	2	0
False Negatives	4	2	5	2
Accuracy	0.8333	0.9333	0.7667	0.9333
Sensitivity	0.9600	1.000	0.9200	1.0000
Specificity	0.2000	0.6000	0.0000	0.6000

Table 26: Results from the prediction based of the Df2 dataset

	Univariate			**Multivariate**
	HBB	ALAS2	CD93	HBB+ALAS2+CD93
True positives	0	3	0	3
True Negatives	25	25	23	25
False positives	0	0	2	0
False Negatives	5	2	5	2
Accuracy	0.8333	0.9333	0.7667	0.9333
Sensitivity	1.000	1.000	0.9200	1.000
Specificity	0.000	0.6000	0.0000	0.6000

Table 27: Results from the prediction based of the Df3 dataset

	Univariate			**Multivariate**
	HBB	ALAS2	CD93	HBB+ALAS2+CD93
True positives	3	3	1	3
True Negatives	24	25	23	25

False positives	1	0	0	0
False Negatives	2	2	4	2
Accuracy	0.9	0.9333	0.8667	0.9333
Sensitivity	0.9600	1.0000	1.0000	1.0000
Specificity	0.6000	0.6000	0.2000	0.6000

The menstruation blood samples

Table 28: Results from the prediction based of the Df1 dataset

	Univariate				**Multivariate**
	HBB	MMP7	MMP10	MMP11	HBB +…+MMP11
True positives	0	2	3	3	1
True Negatives	27	27	27	27	27
False positives	0	0	0	0	0
False Negatives	3	1	0	0	2
Accuracy	0.9	0.9667	1.000	1.000	0.9333
Sensitivity	1.000	1.000	1.000	1.000	1.000
Specificity	0.000	0.6667	1.000	1.000	0.333

Table 29: Results from the prediction based of the Df2 dataset

	Univariate				**Multivariate**
	HBB	MMP7	MMP10	MMP11	HBB+ … +MMP11
True positives	0	2	2	3	2
True Negatives	27	27	27	27	27
False positives	0	0	0	0	0
False Negatives	3	1	1	0	1
Accuracy	0.9	0.9667	0.9667	0.9	0.9667
Sensitivity	1.000	1.000	1.000	1.000	1.000
Specificity	0.000	0.6667	1.000	1.000	0.333

Table 30: Results from the prediction based of the Df3 dataset

	Univariate				**Multivariate**
	HBB	MMP7	MMP10	MMP11	HBB+ … +MMP11
True positives	1	2	3	3	1
True Negatives	24	27	27	27	27
False positives	3	0	0	0	0
False Negatives	2	1	0	0	2
Accuracy	0.8333	0.9667	0.9	0.9	0.9333
Sensitivity	0.8889	1.000	1.000	1.000	1.000
Specificity	0.3333	0.6667	0.000	0.000	0.3333

The Saliva samples

Table 31: Results from the prediction based of the Df1 dataset

	Univariate		**Multivariate**
	HTN3	STATH	HTN3+STATH
True positives	7	2	7
True Negatives	19	16	19
False positives	0	3	0
False Negatives	4	9	4
Accuracy	0.8667	0.6000	0.8667
Sensitivity	1.0000	0.8421	1.0000
Specificity	0.6364	0.1818	0.6364

Table 32: Results from the prediction based of the Df2 dataset

	Univariate		**Multivariate**
	HTN3	STATH	HTN3+STATH
True positives	7	1	7
True Negatives	19	16	19
False positives	0	3	0

False Negatives	4	10	4
Accuracy	0.8667	0.5667	0.8667
Sensitivity	1.0000	0.84211	1.0000
Specificity	0.6364	0.09091	0.6364

Table 33: Results from the prediction based of the Df3 dataset

	Univariate		**Multivariate**
	HTN3	STATH	HTN3+STATH
True positives	7	2	10
True Negatives	19	15	19
False positives	0	4	0
False Negatives	4	9	1
Accuracy	0.8667	0.5667	0.9667
Sensitivity	1.0000	0.7895	1.0000
Specificity	0.6364	0. 1818	0.9091

The semen samples

Table 34: Results from the prediction based of the Df1 dataset

	Univariate			**Multivariate**
	KLK3	PRM1	SEMG1	KLK3+PRM1+SEMG1
True positives	3	3	3	3
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	0	0	0	0
Accuracy	1.000	1.000	1.000	1.000
Sensitivity	1.000	1.000	1.000	1.000
Specificity	1.000	1.000	1.000	1.000

Table 35: Results from the prediction based of the Df2 dataset

	Univariate			**Multivariate**
	KLK3	PRM1	SEMG1	KLK3+PRM1+SEMG1
True positives	3	3	3	1
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	0	0	0	2
Accuracy	1.000	1.000	1.000	0.9333
Sensitivity	1.000	1.000	1.000	1.000
Specificity	1.000	1.000	1.000	0.3333

Table 36: Results from the prediction based of the Df3 dataset

	Univariate			**Multivariate**
	KLK3	PRM1	SEMG1	KLK3+PRM1+SEMG1
True positives	3	3	3	3
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	0	0	0	0
Accuracy	1.000	1.000	1.000	1.000
Sensitivity	1.000	1.000	1.000	1.000
Specificity	1.000	1.000	1.000	1.000

The vaginal secretion samples

Table 37: Results from the prediction based of the Df1 dataset

	Univariate			Multivariate
	MUC4	MYOZ1	CYP2B7P1	MUC4+ MYOZ1+CYP2B7P1
True positives	0	0	0	0
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	3	3	3	3
Accuracy	0.900	0.900	0.900	0.900
Sensitivity	1.000	1.000	1.000	1.000
Specificity	0.000	0.000	0.000	0.000

Table 38: Results from the prediction based of the Df2 dataset

	Univariate			Multivariate
	MUC4	MYOZ1	CYP2B7P1	MUC4+ MYOZ1+CYP2B7P1
True positives	0	0	0	0
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	3	3	3	3
Accuracy	0.900	0.900	0.900	0.900
Sensitivity	1.000	1.000	1.000	1.000
Specificity	0.000	0.000	0.000	0.000

Table 39: Results from the prediction based of the Df3 dataset

	Univariate			Multivariate
	MUC4	MYOZ1	CYP2B7P1	MUC4+ MYOZ1+CYP2B7P1
True positives	1	0	0	1
True Negatives	27	27	27	27
False positives	0	0	0	0
False Negatives	2	3	3	2
Accuracy	0.9333	0.900	0.900	0.9333
Sensitivity	1.000	1.000	1.000	1.000

Specificity	0.3333	0.000	0.000	0.3333

Nasal secretion samples

Table 40: Results from the prediction based of the Df1 dataset

	Univariate		**Multivariate**
	HTN3	BPIFA1	HTN3+ BPIFA1
True positives	3	4	4
True Negatives	25	25	25
False positives	2	0	0
False Negatives	0	1	1
Accuracy	0. 8333	0.9667	0.9667
Sensitivity	1.0000	1.0000	1.0000
Specificity	0.0000	0.8000	0.8000

Table 41: Results from the prediction based of the Df2 dataset

	Univariate		**Multivariate**
	HTN3	BPIFA1	HTN3+ BPIFA1
True positives	3	4	4
True Negatives	25	25	25
False positives	2	0	0
False Negatives	0	1	1
Accuracy	0. 8333	0.9667	0.9667
Sensitivity	1.0000	1.0000	1.0000
Specificity	0.0000	0.8000	0.8000

Table 42: Results from the prediction based of the Df3 dataset

	Univariate		**Multivariate**
	STATH	BPIFA1	STATH + BPIFA1
True positives	3	4	4
True Negatives	23	25	25

False positives	2	0	0
False Negatives	2	1	1
Accuracy	0. 8333	0.9667	0.9667
Sensitivity	1.0000	1.0000	1.0000
Specificity	0.0000	0.8000	0.8000

Summary of the unexpected detection of markers. The prediction from the five models can be illustrated in figure 3.7. Both the code and the dataset that were used to create this result, can be found in Appendix 1 and Appendix 2 respectfully.

Table 43: *Summary of the counted number of the unexpectedly detected mRNA markers in each body fluid. The percentage of the expected and unexpected detected markers are shown in brackets.*

	Blood	**Semen**	**Saliva**	**Menstruation blood**	**Nose secretion**	**Vaginal secretion**
Unexpected number of markers (%)	3 (8.6%)	3 (10.3%)	69 (52.7%)	3 (4.5%)	31 (56.4%)	17 (37%)
Expected number of Markers (%)	32 (91.4%)	26 (89.7%)	62 (47.3%)	64 (95.5%)	24 (43.6%)	29 (63%)
Total number of detected markers (%)	35 (100%)	29 (100%)	131 (100%)	67 (100%)	55 (100%)	46 (100%)

Summary of the one- and two-way ANOVA values

Blood samples

Table 44: *Two way ANOVA:* Height ~ Volum + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***

Volume	*2*	*7.935e+08*	*396772503*	*8.566*	*0.000906****
Marker	*7*	*3.158e+09*	*451181700*	*9.741*	*9.54e-07****
Residuals	*36*	*1.667e+09*	*46317443*	*-*	*-*

Menstruation blood samples

Table 45: Two way ANOVA: Height ~ Volume + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.001e+09*	*500358283*	*4.281*	*0.018129**
Marker	*12*	*5.247e+09*	*437267404*	*3.741*	*0.000287 ****
Residuals	*62*	*7.246e+09*	*116873312*	*-*	*-*

Salvia samples

Table 46: Two way ANOVA: Height ~ Volume + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*4.585e+08*	*229247907*	*2.903*	*0.0582*
Marker	*11*	*5.197e+09*	*472499023*	*5.984*	*5.98e-08****
Residuals	*140*	*1.106e+10*	*78965516*	*-*	*-*

Semen samples

Table 47: Two way ANOVA: Height ~ Volume + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.879e+08*	*93968945*	*2.567*	*0.092038*
Marker	*4*	*9.215e+08*	*230369975*	*6.292*	*0.000706****
Residuals	*33*	*1.208e+09*	*36611968*	*-*	*-*

Vaginal samples

Table 48: Two way ANOVA: Height ~ Volume + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.155e+08*	*57735033*	*0.994*	*0.378*
Marker	*9*	*5.150e+09*	*572246997*	*9.856*	*6.39e-08****
Residuals	*42*	*2.439e+09*	*58061023*	-	-

Nasal samples

Table 49: Two way ANOVA: Height ~ Volume + Marker

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*5.746e+08*	*287315071*	*2.887*	*0.0645*
Marker	*8*	*5.308e+09*	*663558930*	*6.668*	*5.25e-06 ****
Residuals	*53*	*5.274e+09*	*99517161*	-	-

One way – ANOVA
Blood samples

Table 50: *One way ANOVA:* Height ~ Volum

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*7.935e+08*	*396772503*	*3.535*	*0.0379*
Residuals	*43*	*4.826e+09*	*112225578*	-	-

Menstruation blood samples

Table 51: One way ANOVA: Height ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.001e+09*	*500358283*	*2.964*	*0.0578*
Residuals	*74*	*1.249e+10*	*168829111*	-	-

Salvia samples

Table 52: One way ANOVA: Height ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*4.585e+08*	*229247907*	*2.13*	*0.122*
Residuals	*151*	*1.625e+10*	*107633520*	*-*	*-*

Semen samples

Table 53: one way ANOVA: Height ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.879e+08*	*93968945*	*1.633*	*0.209*
Residuals	*37*	*2.130e+09*	*57558780*	*-*	*-*

Vaginal secretion

Table 54: one way ANOVA: Height ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*1.155e+08*	*57735033*	*0.388*	*0.68*
Residuals	*51*	*7.589e+09*	*148799724*	*-*	*-*

Nose secretion

Table 55: one way ANOVA: Height ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Volume	*2*	*5.746e+08*	*287315071*	*1.656*	*0.199*
Residuals	*61*	*1.058e+10*	*173489852*	*-*	*-*

One- way ANOVA (Detection rate)

Blood samples

Table 56: *One way ANOVA:* Per ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Detection rate	*2*	*0.2778*	*0.1389*	*1.203*	*0.328*
Residuals	*15*	*1.7311*	*0.1154*	*-*	*-*

Menstruation blood samples

Table 57: One way ANOVA: Per ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Detection rate	*2*	*0.015*	*0.00749*	*0.095*	*0.909*
Residuals	*30*	*2.355*	*0.07851*	*-*	*-*

Salvia samples

Table 58: One way ANOVA: Per ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Detection rate	*2*	*0.095*	*0.04775*	*0.402*	*0.671*
Residuals	*48*	*5.700*	*0.11876*	*-*	*-*

Semen samples

Table 59: one way ANOVA: Per ~ Volume

	Df	***Sum Sq***	***Mean Sq***	***F-value***	***P-value***
Detection rate	*2*	*0.0531*	*0.02654*	*0.307*	*0.743*
Residuals	*10*	*0.8656*	*0.08656*	*-*	*-*

Vaginal secretion

Table 60: one way ANOVA: Per ~ Volume

	Df	*Sum Sq*	*Mean Sq*	*F-value*	*P-value*
Detection rate	*2*	*0.1687*	*0.08434*	*0.783*	*0.471*
Residuals	*19*	*2.0472*	*0.10775*	-	-

Nose secretion

Table 61: one way ANOVA: Per ~ Volume

	Df	*Sum Sq*	*Mean Sq*	*F-value*	*P-value*
Detection rate	*2*	*0.0228*	*0.01138*	*0.091*	*0.913*
Residuals	*23*	*2.8811*	*0.12526*	-	-

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway