UNIVERSITY OF TURKU
Faculty of Science

# Assessing the effects of *ANO7* mutations on mRNA splicing using a minigene splicing assay

Nea Tulonen

Physiology and Genetics
Master's thesis
Credits: 30 op

Supervisor(s):
Gudrun Wahlström
Christina Nokkala

7.12.2021
Turku

Master's thesis

**Subject**: Biology, Physiology and Genetics
**Author**: Nea Tulonen
**Title**: Assessing the effects of *ANO7* mutations on mRNA splicing using a minigene splicing assay
**Supervisor(s)**: Gudrun Wahlström, Christina Nokkala
**Number of pages**: 59 pages + 6 appendix pages
**Date**: 7.12.2021

Prostate cancer (PrCa) is the most common cancer type in men. Dysregulated splicing is considered a hallmark of cancer, and PrCa has its own characteristic splicing landscape. Alternative splicing (AS) enables production of multiple protein isoforms from a single gene by altering splicing of exons and introns. AS outcomes can be studied with reverse transcription PCR (RT-PCR) or a minigene splicing assay, which allows comparison of wild type and mutant sequences in identical conditions.

This MSc study focused on two splicing mutations in *ANO7*, a PrCa susceptibility gene. Variant rs77559646 in the splice donor site of intron 4 leads to exon skipping and increased intronic expression presumably by disrupting binding of an essential splicing factor. The variant was corrected by CRISPR-Cas9 base-editing, also producing a new unwanted mutation upstream. This study aimed to determine whether normal splicing would be restored regardless. Ultimately, RT-PCR analysis revealed that intron 3 expression was not reduced. Additionally, the region exhibited deletions related to *Alu* repeat elements. Splicing assay showed that exon 4 splicing was improved, but only slightly.

Variant rs78154103 in intron 7 causes cryptic splice site selection and partial intron retention. The variant's effect on splicing was studied with a minigene splicing assay using a longer construct than in a previous study, allowing formation of endogenous secondary structures, suspected of contributing to dysregulated splicing. As anticipated, more exon skipping was detected compared to reference, but inclusion of cryptic exon did not differ between the constructs, contrary to what was expected.

**Key words**: *ANO7*, prostate cancer, SNP, mutation, alternative splicing, RT-PCR, minigene splicing assay

Pro gradu -tutkielma

**Pääaine**: Biologia, Fysiologian ja genetiikan linja
**Tekijä**: Nea Tulonen
**Otsikko**: Assessing the effects of *ANO7* mutations on mRNA splicing using a minigene splicing assay
**Ohjaaja(t)**: Gudrun Wahlström, Christina Nokkala
**Sivumäärä**: 59 sivua + liitteet 6 sivua
**Päivämäärä**: 7.12.2021

Eturauhassyöpä on miesten yleisin syöpä. Eturauhassyövässä, kuten muissakin syövissä, havaitaan häiriintynyttä esiaste-RNA:n silmukointia. Vaihtoehtoinen silmukointi mahdollistaa useamman erilaisen proteiinituotteen tuoton yhdestä geenistä vaihtoehtoisilla eksoni-introni-yhdistelmillä. Vaihtoehtoista silmukointia voidaan tutkia esimerkiksi RT-PCR-tekniikalla (reverse transcriptase - polymerase chain reaction) tai minigeenien silmukointianalyysillä (engl. minigene splicing assay), joka mahdollistaa villityypin- ja mutanttisekvenssin vertailun identtisissä olosuhteissa.

Tässä tutkielmassa tutkin kahta *ANO7*-eturauhassyöpäalttiusgeenissä ilmenevää silmukointimutaatiota. Variantti rs77559646 sijaitsee introni 4:n luovuttaja-alueella, josta silmukointi tavallisesti alkaa. Variantti johtaa eksoni 4:n poissilmukointiin ja lisääntyneeseen introni 3:n ekspressioon oletettavasti estämällä silmukointiin tarvittavan ribonukleoproteiinin sitoutumisen. Mutaatio on korjattu CRISPR-Cas9 base-editing -tekniikalla, mutta samanaikaisesti syntyi uusi mutaatio yläjuosteessa. Tutkimuksen tavoitteena oli selvittää, palautuuko asianmukainen silmukointi tahattomasta muutoksesta huolimatta. Asiaa tutkittiin RT-PCR-tekniikalla ja minigeenien silmukointianalyysillä. RT-PCR-analyysi osoitti, että introni 3:n ekspressio ei eronnut muokatun ja varianttisekvenssin välillä. Lisäksi tutkitulla introni 3 -alueella havaittiin deleetioita, jotka olivat yhteydessä *Alu*-toistojaksoihin. Silmukointianalyysistä kävi ilmi, että eksoni 4:n silmukointi palautui, mutta odotettua heikommin.

Introni 7:ssä sijaitseva variantti rs78154103 aiheuttaa kryptisen silmukointikohdan aktivaation, johtaen kryptisen eksonin (eksoni + osa intronia) muodostumiseen. Variantin vaikutusta silmukointiin tutkittiin minigeenien silmukointianalyysillä käyttäen pidempää minigeenikonstruktia, kuin aiemmassa tutkimuksessa. Pidempi sekvenssi mahdollistaa RNA:n sekundaarirakenteiden muodostumisen, joiden rakennemuutosten epäillään aiheuttavan häiriintynyttä silmukointia kyseisellä geenialueella. Odotusten mukaisesti variantti aiheutti lisääntynyttä eksonin poistoa verrattuna villityyppiin. Kryptinen eksoni kuitenkin ekspressoitui yhtä vahvasti molemmilla konstrukteilla, toisin kuin odotettiin.

**Avainsanat:** *ANO7*, eturauhassyöpä, SNP, mutaatio, vaihtoehtoinen silmukointi, RT-PCR

# List of abbreviations

| | |
|---|---|
| 3’ss | 3’ splice site; splice acceptor |
| 5’ss | 5’ splice site; splice donor |
| AAMR | Alu/Alu-mediated rearrangement |
| ADT | androgen deprivation therapy |
| ANO | anoctamin |
| ANO7 | anoctamin 7 |
| ANOVA | analysis of variance |
| AR | androgen receptor |
| AR-V | androgen receptor splice variant |
| AS | alternative splicing |
| ASO | antisense oligonucleotide |
| BLAST | Basic Local Alignment Search Tool |
| BPS | branch point sequence |
| CRPC | castration-resistant prostate cancer |
| DHT | 5α-dihydrotestosterone |
| dPSI | delta percent spliced in |
| DRE | digital rectal exam |
| D-TMPP | Dresden-transmembrane protein of the prostate |
| ESE | exonic splicing enhancer |
| ESS | exonic splicing silencer |
| GWAS | genome-wide association studies |
| ISE | intronic splicing enhancer |
| ISS | intronic splicing silencer |
| KLF6 | Kruppel-like Factor 6 |
| LTR | long terminal repeats |

| | |
|---|---|
| mAb | monoclonal antibody |
| MAXENT | maximum entropy model |
| MMBIR | microhomology-mediated break-induced replication |
| MMEJ | microhomology-mediated end joining |
| mRNA-Seq | mRNA-sequencing |
| NAHR | nonallelic homologous recombination |
| NGEP | new gene expressed in prostate |
| NMD | nonsense-mediated decay |
| PrCa | prostate cancer |
| Pre-mRNA | precursor messenger RNA |
| PSA | prostate-specific antigen |
| PSAP | prostatic acid phosphatase |
| RBP | RNA-binding protein |
| RNAi | RNA interference |
| RNP | ribonucleoprotein |
| RT-PCR | reverse transcription polymerase chain reaction |
| RT-qPCR | quantitative RT-PCR |
| SINE | short interspersed elements |
| snRNA | small nuclear RNAs |
| snRNP | small nuclear RNP |
| ss | splice site |
| TMEM16 | transmembrane protein 16 |
| $\Psi$ | pseudo-uridine nucleotide |

# Contents

1.1 Prostate cancer (PrCa) ..... 1
1.1.1 Epidemiology ..... 1
1.1.2 Development and progression ..... 1
1.1.3 Diagnostics and treatment ..... 2
1.2 PrCa genetics ..... 4
1.2.1 Heritability ..... 4
1.2.2. Genetic biomarkers ..... 5
1.3 Alternative splicing and cancer ..... 7
1.3.1 Pre-mRNA splicing ..... 7
1.3.2 Alternative splicing (AS) ..... 8
1.3.3 AS, mutagenesis, and cancer ..... 9
1.3.4 AS and PrCa ..... 11
1.3.5 Studying AS ..... 12
1.4 *ANO7* ..... 14
1.4.1 *ANO7* and the TMEM16 protein family ..... 14
1.4.2. *ANO7* and cancer ..... 16
1.4.3 *ANO7* variants and PrCa ..... 18
1.5 Aims of the study ..... 20
2. Materials and methods ..... 22
2.1 Intron 3 expression level determination ..... 22
2.2 Plasmid construction ..... 24
2.2.1 Exon 4 constructs ..... 24
2.2.2 Intron 7 constructs ..... 26
2.3 Minigene splicing assay ..... 27
2.3.1 Intron 7-specific RT-PCR ..... 29
3. Results ..... 30
3.1 Intron 3 expression level determination ..... 30
3.1.1. Defining sequence deletions ..... 33
3.2 Minigene splicing assay ..... 35
3.2.1 Intron 7 -specific ..... 38
4. Discussion ..... 41
4.1 rs77559646 ..... 42
4.1.1 Intron 3 sequence deletions ..... 42
4.1.2 Minigene splicing assay ..... 45
4.2 rs78154103 ..... 47
Acknowledgments ..... 51
References ..... 52
Appendices .....

# 1. Introduction

## 1.1 Prostate cancer (PrCa)

### 1.1.1 Epidemiology

Prostate cancer (PrCa) is the most common cancer type in men, with a bit over 5,000 cases reported annually in Finland (Pitkäniemi et al. 2020). Incidence rates began increasing in the 1990s, and consequently so did PrCa mortality. This increase in reported cases is, however, most likely due to increased screening for PSA (prostate-specific antigen), which is an enzyme secreted by the prostate gland (Duodecim 2021). Secretion of PSA is elevated in patients with PrCa but can also be caused by other conditions (Käypä hoito 2021). In 2018, PrCa was the most diagnosed cancer in Finnish men, as well as the second leading cause of cancer related deaths in men (> 900 deaths per year) (Pitkäniemi et al. 2020). Fatal cases in PrCa are mainly due to metastatic disease (Wang et al. 2018). However, mortality has decreased drastically and prognosis for PrCa has improved, with 5-year survival rate being well over 90 % (Pitkäniemi et al. 2020; Duodecim 2021). Prognosis depends on how progressed the tumor is, its histologic grade, age and general health of the patient, as well as PSA level (Käypä hoito 2021). It is worthwhile mentioning that many PrCa cases are clinically insignificant, and screening regularly leads to overdiagnosis and overtreatment (National cancer institute 2021).

### 1.1.2 Development and progression

The prostate gland is comprised of luminal cells, basal cells, and neuroendocrine cells (Wang et al. 2018). Most primary prostate cancers are adenocarcinomas (National cancer institute 2021), which develop through a multistep process, as reviewed in Wang et al. (2018). Tumor formation begins with prostatic neoplasia and has been shown to originate from both luminal and basal cells. A small subset of PrCa cases comprise of other forms of carcinoma (National cancer institute 2021). Prostate tissue is known to be quite heterogeneous, consisting of multiple different cell types, and prostate tumors are a mixture of both malignant and nonmalignant cells (Stuart et al. 2004). Overall, PrCa exhibits large heterogeneity between patients and intratumorally, in terms of pathology, function, and genomic features. This heterogeneity complicates detection and development of reliable markers and therapy options (Wang et al. 2018).

Risk factors of PrCa include increasing age, ethnic background, family history, as well as dietary factors and alcohol use (National cancer institute 2021). Risk is higher in men with advanced age, and African-American men have an elevated risk compared to European and Asian men. Some risk-increasing dietary factors include high consumption of fats, processed and red meats, and excessive use of multivitamins (Käypä hoito 2021). Major cellular processes that have been proposed to contribute to prostate carcinogenesis include chronic inflammation, oxidative stress and resulting DNA damage, telomere shortening, senescence, and genetic factors (Shen & Abate-Shen 2010).

A major contributor to PrCa development and progression is the increased expression and/or activation of androgen receptor (AR) and consequently the AR-signaling pathway (Wang et al. 2018). The most common AR-binding androgens are testosterone and its more active metabolite 5α-dihydrotestosterone (DHT). AR-mediated signaling is essential for normal prostate development as well, but is emphasized in PrCa (McCrea et al. 2016). Increased AR activity leads to upregulation and downregulation of various genes, such as PSA, and prostatic AR has long been thought to function as a proliferator (Heinlein & Chang 2004). In 2008, Niu et al. suggested that AR works via proliferator as well as suppressor activities, depending on the cell-type (epithelial basal, epithelial luminal, or differentiating stromal) and the stage of PrCa progression. AR's role as a proliferator, however, remains to be the prevailing view (McCrea et al. 2016). PrCa is strongly hormone responsive and hormonal therapy is common. Unfortunately, so is development of resistance (see more below). Additional signaling pathways dysregulated in PrCa include the PI3K pathway and several DNA repair pathways, among more (Wang et al. 2018).

### 1.1.3 Diagnostics and treatment

PrCa symptoms include changes in urination, such as increased need to urinate, difficulty in emptying the bladder entirely, weak urine flow, and blood in urine (Duodecim 2021). If cancer has spread, the range of symptoms increases. For instance, in the metastasized disease form, bone pain is a common symptom. Despite the high prevalence of PrCa, symptomless men are not screened. Overdiagnosis often leads to overtreatment, which, in worst case scenario leads to permanent side effects and other complications (National cancer institute 2021). If a man does, however, exhibit symptoms, the first diagnostic measure is a digital rectal exam (DRE). In DRE, the doctor feels the prostate for lumps or other abnormalities through the rectum with his/her finger. In addition to DRE, a blood sample will be taken to test for increased levels of PSA. As mentioned above, higher PSA levels do not necessarily imply cancer, since this increase can also be caused by old age, infection, or enlargement of the prostate gland (benign prostatic hyperplasia). Thus, PSA testing should not be used as an only measure. Additionally, whether PSA testing improves prognosis remains debated (National cancer institute 2021). Nevertheless, its great value in screening comes from simplicity and low-risk, compared to other diagnostic measures. Another clinical standard biomarker for PrCa is prostatic acid phosphatase (PSAP) (Guo et al. 2021). Like PSA, PSAP levels are also affected by other factors than just PrCa. If the probability of PrCa remains high after initial testing, a biopsy is performed to diagnose PrCa. Biopsy does, however, pose risks, and performing a biopsy should be carefully evaluated (National Cancer Institute 2021).

Following biopsy, a Gleason score is devised to grade the cancer and the stage of the cancer is determined (staging) (National Cancer Institute 2021). The Gleason score ranges from 6 to 10, where a Gleason score of >8 presents high-risk and worse prognosis. During staging, it is also determined whether the cancer has spread. Staging is important for choosing the right treatment option. In stages I and II, cancer cells are found locally only in the prostate gland. In stage III and IV prostate cancers, the tumor has spread to surrounding tissues and to other parts of the body, respectively.

There are many treatment options, and the choice depends on the grade and stage of the cancer, size of the tumor, patient's age, preexisting medical problems, and whether the cancer is newly discovered or relapsed (Duodecim 2021; National cancer institute 2021). These options include active surveillance, surgery (radical prostatectomy), radiation therapy, and hormone therapy. Careful consideration is required to find the most effective treatment or combination of treatments for each patient and tumor type separately. Low-risk patients are often subjected to surveillance only to avoid overtreatment (National cancer institute 2021), but patients with intermediate-risk PrCa present a challenge regarding treatment decisions (Wang et al. 2018). As with some screening methods, some treatment options pose risks and adverse effects. Radical prostatectomy, for instance, can cause incontinence, impotence, and penile shortening, whereas radiation therapy is often associated with bowel dysfunction (National cancer institute 2021).

Several hormonal treatment options are available, of which androgen deprivation therapy (ADT) is the most common treatment method. Androgen ablation is achieved by surgical or chemical castration (Heinlein & Chan 2004; Shen & Abate-Shen 2010) and is often coupled with chemotherapy or radiation (Wang et al. 2018). ADT, however, often results in resistance and recurrence. The recurrent disease is called castration-resistant prostate cancer (CRPC). Niu et al. (2008) suggested the reason for recurrence is the possible dual-function of AR in proliferation and suppression, where depletion of its suppressor properties promotes tumor progression. The more conventional view, however, is that development of castration resistance is, in most cases, an adaptive response to ADT (Shen & Abate-Shen 2010). In CRPC, androgen signaling is sustained by many mechanisms, such as AR overexpression. Overexpression is achieved by, for instance, copy number amplification, gain-of-function mutations, increased intracellular androgen synthesis, and alternative splicing (see more below) (McCrea et al. 2016). Other forms of resistance have been identified as well (Wang et al. 2018).

In addition to traditional treatment options, development of cancer immunotherapy treatments is intensively pursued. Especially the capacity of PrCa to develop resistance to different therapies, as well as the large intratumoral heterogeneity in PrCa brings about obstacles in finding effective treatments (Wang et al. 2018). New treatment options are being tested in clinical trials constantly (Wang et al. 2018; Duodecim 2021; National Cancer Institute 2021).

## 1.2 PrCa genetics

### 1.2.1 Heritability

PrCa is a complex disease, meaning that genetic and environmental factors are at play. Risk of PrCa increases with age, and incidences are highest at 80 years of age (Pitkäniemi et al. 2020). Elevated risk is also associated with family cancer history and ethnic background: men with affected first-degree relatives are at much greater risk, as are African-American men (National Cancer Institute 2021). The heritability of PrCa is relatively high, and genetic factors are estimated to explain as much as 57 % of the risk, according to Hjelmborg et al. (2014).

Originally in 1993, Carter et al. described PrCa as having three subtypes: familial, hereditary, and sporadic (reviewed in Raghallaigh & Eeles 2021). Both familial and hereditary types are due to increased risk in men with family history of PrCa. Familial PrCa is simply characterized by having a history of PrCa in the family, whereas hereditary PrCa is a more specific type and requires one of three criteria to be fulfilled, which the familial type need not fulfill. Therefore, hereditary PrCa can be thought of as a subtype of familial PrCa, with at least one of the following occurring, according to Raghallaigh & Eeles (2021): PrCa must have been reported in at least three generations (1), onset of PrCa must have happened at the age of <55 years in at least two family members (2), or the disease has been found in at least three first-degree relatives (3). Out of all PrCa cases, the hereditary subtype accounts for 3—5 % of cases, whereas familial PrCa accounts for 10—15 % of cases. Sporadic PrCa covers the remaining 85 % of all PrCa cases (Raghallaigh & Eeles 2021).

Rodríquez et al. (1997) found that family history increases risk of fatal PrCa by 60 % compared to those without affected family members. They also found that risk of PrCa increases proportionally to the number of affected relatives, and the family-degree of an affected relative contributes to the risk as well. Similar results were obtained by Zeeger et al. (2003). Additionally, they found that the age of onset contributes to risk of recurrence, with lower age increasing risk. Familial PrCa is extensively studied using twin studies. For instance, a Nordic twin study conducted by Helmborg et al. (2014) found that the risk to develop PrCa was approximately three times higher for men with an affected monozygotic twin, than for those without. Risk was significantly higher (two-fold) for men with dizygotic twins as well, compared to the overall population. It has been found that, even though age of onset is lower in hereditary PrCa, the sporadic and hereditary forms do not otherwise differ clinically (Bratt 2002). Earlier onset does, however, lead to increased mortality in hereditary PrCa compared to sporadic.

### 1.2.2. Genetic biomarkers

Identifying genetic biomarkers of PrCa is pivotal for disease diagnosis, defining disease subtypes, and development of new therapeutic strategies. The search for effective prognostic biomarkers is intense. Search methods include gene expression profiling, miRNA expression profiling, serum proteomics and metabolomics (Shen & Abate-Shen 2010). PrCa susceptibility has been associated with several genes and hundreds of SNPs through linkage analyses and genome-wide association studies (GWAS) (Raghallaigh & Eeles 2021). Frequencies of PrCa associated variants differ between different ethnic groups, according to Olender & Lee (2019). SNPs associated with PrCa risk are multiplicative, meaning that their effects are cumulative. This makes alleles associated with PrCa with only low-to-moderate risk, yet are common at the population-level, a considerable contributor to PrCa initiation and progression (Dadaev et al. 2018).

Some PrCa susceptibility genes include *HOXB13*, *BRCA1/2*, *ATM*, *CHEK2* and *TP53*, to mention a few (National Cancer Institute 2021; Raghallaigh & Eeles 2021). The tumor suppressor *HOXB13* was the first gene identified as a hereditary prostate cancer gene (National cancer institute 2021). *HOXB13* contributes to PrCa risk by interacting with the AR. Out of the two tumor suppressor genes *BRCA1* and *BRCA2*, germline mutations in the latter are shown to have a more substantial impact on PrCa risk (Nyberg et al. 2020). Nyberg et al. (2020) showed that *BRCA2* mutations are also linked to higher risk of aggressive PrCa. Genetic testing for *BRCA* variants is already implemented to some extent and likely will become more common in the future (National cancer institute 2021; Raghallaigh & Eeles 2021).

The vast heterogeneity in PrCa can also be seen in the expression levels of genes associated with PrCa risk. For instance, a microarray meta-analysis of four independent PrCa gene expression datasets by Rhodes et al. (2002) found a total of 500 genes upregulated as well as several downregulated in clinically localized PrCa compared to benign prostate tissue. Stuart et al. (2004) set to distinguish expression level differences between different cell types found in prostate tissue and identified several gene expression alterations to be cell type- and tumor-specific. Their findings supported over 300 of the genes reported by Rhodes et al. (2002) to be upregulated in prostate tumor tissue. Additionally, Chandran et al. (2007) found that over 400 genes are upregulated and over 350 genes downregulated in the metastatic form of PrCa, compared to primary tumor tissue. Knowledge of expression level changes can be implemented in the identification of tumor aggression. The role of these genes varies from cell-cell interaction control to transcription regulation. Wang et al. (2018) review some common genes and signaling pathways involved and dysregulated in PrCa. Taken together, these findings represent the fact that several biological processes are altered in PrCa.

## 1.3 Alternative splicing and cancer

### 1.3.1 Pre-mRNA splicing

Pre-mRNA (precursor messenger RNA) splicing (henceforth referred to as mRNA splicing) is a post-transcriptional mechanism essential for gene expression. At its core, mRNA splicing involves the removal of introns from pre-mRNA with two transesterification events catalyzed by a specialized machinery called the spliceosome. In addition to the spliceosome, *cis*-acting elements in the pre-mRNA sequence, as well as *trans*-acting proteins are needed for splicing to occur (Olender & Lee 2019). A spliceosome is a ribonucleoprotein (RNP) complex consisting of five small nuclear RNPs (snRNPs) and hundreds of associated proteins that are needed for accurate and stable splicing (Abramowicz & Gos 2018). The main snRNPs in the spliceosome, which themselves are composed of small nuclear RNAs (snRNAs), are U1, U2, U4/6 and U5 (Wahl et al. 2009). The accessory proteins have varying roles – some are RNA-binding proteins (RBPs; such as U2AF and SF1) and some are enzymes (such as helicases and kinases) (Cooper et al. 2009). RBPs are those which regulate when, where and how frequently splicing events occur. Major conformational rearrangements of the spliceosome complex are also crucial for splicing to occur (Wahl et al. 2009). Several RBPs are associated with structural modifications of the RNA sequence as well (Cooper et al. 2009). In addition, changes in RNA secondary structures also regulate splicing and gene expression (Olender & Lee 2019).

The first essential step in splicing is the recognition of splice sites (ss) which are located at exon-intron boundaries: one at the 3' end of the intron (3'ss) and another at the 5' end (5'ss). The most common splice site dinucleotides found at the exon-intron boundaries are GT at the 5'ss and AG at the 3'ss (Abramowicz & Gos 2018). In the first step of splicing, the 5'ss, also called the splice donor, is identified by the U1 snRNP, which then binds to the RNA sequence. The partially conserved 5'ss sequence spans positions -3 to +6, where the first three nucleotides reside in the exon and the six latter in the intron (Roca et al. 2005). After U1 binding, U2 snRNP base pairs with the branch point sequence (BPS) of the intron. Conformational rearrangements take place and rest of the snRNPs (U4/U6 and U5) are also recruited to the forming complex – the U6 snRNP, for instance, replaces U1. The complex experiences catalytic activation and the first transesterification event, where a 2'-hydroxyl group in the BPS attacks a phosphodiester bond in the 5'ss, takes place. At this point a lariat structure is formed, which consists of the intron to be cleaved and the 3' exon. After additional compositional and conformational changes to the complex, another transesterification reaction takes place, where a 3'-hydroxyl group attached to the 5'ss attacks the 3'ss, also known as the splice acceptor. After the two transesterification reactions, the intron is excised, and the two exons ligated together. All through the splicing process, RBPs bind and detach to help stabilize the RNA-RNA interactions and contribute to structural modifications. (Wahl et al. 2009).

Splice site selection is influenced by relative strength of the splice site and *cis*-acting elements, which include splicing enhancers and silencers. The strength of a splice site is defined by the *cis*-acting sequences' ability to bind *trans*-acting proteins (Olender & Lee 2019). Additionally, exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs) work to facilitate splicing, whereas exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs) have an opposite function.

### 1.3.2 Alternative splicing (AS)

A process called alternative splicing (AS) enables the production of multiple protein isoforms from a single gene by alternatively splicing either introns or exons. AS is part of normal development and differentiation. In fact, the great majority (>90 %) of human genes exhibit AS post-transcription, which greatly increases the repertoire of protein isoforms generated by genes. AS allows cells to respond to different environments and adapt to different developmental stages. (Olender & Lee 2019). Generally speaking, this means that AS outcomes are time- or spatial-dependent. According to Sugnet et al. (2004), AS events in humans and mice are highly conserved. This further demonstrates the importance of AS and how it provides an evolutionarily advantage.

Traditionally AS mechanisms are divided into five groups, which include exon exclusion (a.k.a. exon skipping), intron inclusion (a.k.a. intron retention), alternative 3' and 5' splice sites, and mutually exclusive exons, as reviewed by Wang & Aifantis (2020). Exon exclusion is thought to be the most frequent type of AS in vertebrates (Sugnet et al. 2004) as well as among different cancer types (Tsai et al. 2015). In exon exclusion, whole exons are omitted from the final transcript. Intron inclusion on the other hand means that an intron is retained in the final transcript, which tends to lead to a premature stop-codon and consequently nonsense-mediated decay (NMD). The use of alternative splice sites leads to only a segment of an exon being excluded or conversely only a segment of intron included. Mutually exclusive exons, as the name implies, are those where only one or the other is retained in the final transcript.

### 1.3.3 AS, mutagenesis, and cancer

Pre-mRNA processing requires a vast amount of different accessory and regulatory proteins and RNAs, as discussed above, which inevitably increases the probability of mutations occurring and consequently dysregulation of splicing (Cooper et al. 2009). Mutations in certain genomic regions affect mRNA splicing and are thus called splicing mutations. Splicing can be influenced through alterations of splice sites or splicing regulatory elements, or by creating new splice sites or activating existing, cryptic ones (Abramowicz & Gos 2018). Splicing mutations may also affect splicing machinery components (Wang & Aifantis 2020), yet these types of splicing mutations are more uncommon (Cooper et al. 2009). Abramowicz and Gos (2018) divide splicing mutations into different categories: (1) mutations can occur in traditional splice sites, where they lead to difficulties in RNA-protein interactions and consequently exon skipping; (2) a mutation in an intronic sequence can give rise to a new, cryptic splice site which, when activated, leads to part of an intron (cryptic exon) being included; (3) a cryptic splice site can arise in the exon by mutation, leading to part of an exon being excluded; (4) mutations in exonic sequences can disrupt the function of *cis*-acting elements; and finally (5) splicing can be affected by alterations of mRNA secondary structures. Fundamentally, splicing mutations exert their power through affecting the strength of a splice site – natural or cryptic. A cryptic splice site is used when it is stronger than the natural splice site. It is notable that a splicing mutation in a *cis*-element affects splicing of a single gene, in which the mutation occurs, whereas a splicing mutation in *trans*, such as splicing machinery components, can cause aberrant splicing in multiple genes (Faustino & Cooper 2003).

Out of all disease-causing mutations, 30—60 % are estimated to be splicing mutations, according to Jung et al. (2015). The great majority of splicing-associated factors and regulators are subject to disease-causing mutations as well as expression level dysregulation (Bonnal et al. 2020). Cancer cells are known to exploit AS to facilitate all hallmarks of cancer, and dysregulated splicing in cancer has gained more attention in cancer research in recent years (Wang & Aifantis 2020). Splicing dysregulation itself is regarded as a hallmark of cancer (Cooper et al. 2009) and contributes to the vast heterogeneity of cancer (Rajan et al. 2009).

Several different kinds of splicing mutations have been found to promote tumorigenesis, along with other, non-splicing related mutations. Cancer-related splicing mutations usually involve disrupting interactions between mRNA and RBPs or by activating cryptic splice sites (Jiang et al. 2000). Transcripts resulting from aberrant splicing may exert their tumorigenic properties via completely new mRNA isoforms, or by altering expression of already existing isoforms (Pajares et al. 2007; Tsai et al. 2015; Olender & Lee 2019). The former results from unnatural splicing patterns, whereas the latter is a result of alternative splice site selection (Faustino & Cooper 2003).

Splicing mutations may act as oncogenic drivers or passengers, and regularly lead to activation of proto-oncogenes or suppression of tumor suppressors. For the former, increased activation and gain-of-function are the cause of acquired oncogenic activity, whereas for the latter, frameshift mutations or generation of premature stop codons and resulting NMD are a common mechanism for inactivation (Olender & Lee 2019), as in the case of the famous tumor suppressor *TP53* (Jung et al. 2015). Consequences of aberrant splicing may include, for instance, epigenetic alterations, DNA damage, or gene expression changes (Olender & Lee 2019).

Differential expression of splicing-associated proteins has been reported in many tumor tissues (Wang & Aifantis 2020). The splicing factor SF2 (ASF, SRSF1), for example, is regularly overexpressed in cancer due to amplification and activation of the gene encoding for it, *SFRS1*. *SFRS1* is a proto-oncogene, as shown by Karni et al. (2007). They found that SF2 contributes to tumorigenesis by dysregulating expression of different splicing isoforms in many cancer-related genes.

Splicing-related therapy strategies have been intensively explored. Indeed, cancer-specific splice variants are a promising tool for diagnostics, identifying disease aggressiveness, and cancer therapy. Possible therapy targets include alterations in splice sites, as well as in *cis*-sequences and *trans*-acting proteins (Olender & Lee 2019). Wang & Aifantis (2020) in turn point out the utility of splicing-induced tumor neoantigens. These neoantigens, namely truncated protein isoforms that have evaded NMD, show potential for cancer vaccines. Additionally, spliceosome inhibitors (Wang & Aifantis 2020) and modulators (Paschalis et al. 2018) have been proposed as splicing-related therapeutic strategies. Yet another option is pre-mRNA-binding antisense oligonucleotides (ASOs), which work by preventing unfavorable splicing and promoting proper splicing (Cooper et al. 2009; Wang & Aifantis 2018). Antisense hybridization by other RNAs has also been developed to work in a similar manner as ASOs, and direct elimination of mRNA using RNA interference (RNAi) is also a potential therapy form (Cooper et al. 2009).

Cooper et al. (2009) suggest that targeting misfunctioning RNAs and RNPs involved in splicing regulation could be a more viable therapy approach than targeting the mis-spliced proteins. Additionally, Tsai et al. (2015) propose that splicing factors should work as therapy targets instead of AS events, which would be better suited for serving as biomarkers. To escape therapies, some oncogenes use differential splicing to generate alternative protein isoforms, unrecognizable to the treatment method being used (Wang & Aifantis 2020). Furthermore, splicing variants may impact how the cancer cell responds to other non-splicing related therapeutic approaches by making the cell more vulnerable to certain therapies or by increasing resistance (Bonnal et al. 2020).

### 1.3.4 AS and PrCa

As with other cancer types, PrCa has its own characteristic splicing landscape. AS is involved in all steps of PrCa, including development, progression, and even drug resistance. Splicing contributes to PrCa via many mechanisms, such as mutated splicing machinery components and splicing factors, altered expression of splicing factors, as well as through effects on cellular signaling pathways.

As discussed, persistent AR signaling is integral for treatment resistance in PrCa and the development of CRPC. AS has been identified as one factor contributing to development of resistance via alterations of the *AR* transcript. The *AR* gene is comprised of 8 exons and several AR splice variants (AR-Vs) have been reported (Paschalis et al. 2018). Aberrant splicing can uphold AR signaling by, for instance, production of AR isoforms that work independently of circulating androgens. One AR-V working by such mechanism is the splice variant AR-V7, which is the most clinically significant of all 20 or so AR-Vs identified (Antonarakis et al. 2016; Paschalis et al. 2018; Olender & Lee 2019). AR-V7 has been shown to increase risk of relapse and correlates with worse prognosis. AR-V7 is significantly upregulated in CRPC tissue, as well as bone metastatic tissue, a proven by Hörnberg et al. (2011). Consequently, it poses a potential biomarker for the disease. AR-V7 is a result of premature polyadenylation and alternative splice site selection near exon 3, which lead to cryptic exon 3 inclusion and a truncated protein product (Paschalis et al. 2018; Olender & Lee 2019). This truncated AR does not contain a ligand-binding domain needed for androgen binding, which eventually leads to androgen-independent activation (McCrea et al. 2011; Antonarakis et al. 2016; Olender & Lee 2019). Several regulatory mechanisms for AR-V7 expression have been proposed (Paschalis et al. 2018). Increased expression of AR-V7 is the result of increased *AR* expression and decreased AR signaling, of which both are achieved by ADT (Antonarakis et al. 2016). Hence, ADT indirectly leads to increased production of AR-V7, which then contributes to resistance. The functional role of splice variant AR-V7 is yet to be discovered (Paschalis et al. 2018).

Clinically significant AR splice variants are insensitive to antiandrogen therapies. This is due to absence of the aforementioned ligand-binding domain, which acts as the target of such therapies (Paschalis et al. 2018). Following from this, it ought to be beneficial to develop treatments targeting other domains of the AR, which have been proven to be sustained in AR-Vs as well (Antonarakis et al. 2016). Additionally, splicing factors have been identified that favor production of AR-V7, and these could potentially be used as therapeutic targets (McCrea et al. 2011).

Other genes with splice variants associated with PrCa include *FGFR2*, *Bcl-x*, *KLF6*, *CLK1* and *VEGF*, to mention a few (Rajan et al. 2009; Olender & Lee 2019). *KLF6* (Kruppel-like Factor 6) for instance has a splice variant called *KLF6-SV1* which has been associated with increased metastasis and poorer survival in men with PrCa (Narla et al. 2008). *KLF6* itself is a tumor suppressor, whereas splice variant *KLF6-SV1* possesses oncogenic properties. *KLF6-SV1* is a result of alternative 5'ss selection in exon 2 (Rajan et al. 2009). Narla et al. (2008) showed that the variant is overexpressed in PrCa and even more so in metastatic tissue. They tested the significance of *KLF6-SV1* in respect to PrCa progression and metastasis by inhibition with RNAi and found that tumor growth was suppressed. This suggests that *KFL6-SV1* could prove to be a potential therapeutic target.

### 1.3.5 Studying AS

The Human Gene Mutation Database estimates the proportion of splicing mutations out of all mutations in the genome to be approximately 9 %. Abramowicz and Gos (2018) point out that this estimate might be greatly underestimated since identifying splicing mutations, especially by only using DNA sequencing, is challenging. What makes studying aberrant splicing even more difficult is the tissue-specificity of splicing (Tsai et al. 2015).

Several programs and bioinformatic algorithms have been devised to predict the effect of a mutation or other genomic alteration on splicing (Di Giacomo et al. 2013; Abramowicz & Gos 2018). Due to the predictive nature of these *in silico* methods, as well as their tendency to often mispredict splicing variants' effects (Di Giacomo et al. 2013), *in vitro* functional studies are necessary. Predictive algorithms are important nevertheless for guiding towards further analysis of variants that are likely clinically significant (Spurdle et al. 2008).

At the time, the most efficient way to determine the effect of a mutation on splicing is to use reverse transcription PCR (RT-PCR) on RNA extracted from patient tissue or cell line and sequence the resulting cDNA. RT-PCR provides a relatively reliable and fast method for determining the splicing landscape of a tissue. It allows the detection of all transcript variants expressed in the tissue. A downside of direct RNA analysis by RT-PCR is the possibility of NMD, however. (Abramowicz & Gos 2018). Additionally, RNA is often not readily available, and so other methods have been devised (Spurdle et al. 2008).

Another method for examining splice variants is by microarrays designed specifically for detection of splicing changes. Microarrays rely on probes which bind specific fragments of DNA or RNA. This hybridization emits a signal which is then measured, and which reflects the relative expression of each fragment. Splicing-specific arrays differ from conventional microarrays in that they are designed to bind all exons (exon arrays) or exon junctions (exon junction array) (Rajan et al. 2009), or both (Fehlbaum et al. 2005). This allows direct measurement of exon expression and the full range of transcripts. These so-called splice-arrays can be designed to measure AS events of all kinds, as proven by Fehlbaum et al. (2005). Splice-arrays have drawbacks, such as signal variability and the impact of initial amounts of the splice variants in cells, especially if low, as pointed out by Rajan et al. (2009). These microarrays are developed based on existing knowledge of AS events derived from public databases. A major drawback therefore comes from the availability of known AS events, which is incomplete to say the least (Pajares et al. 2007).

Limitations of microarrays can be overcome by mRNA-sequencing (mRNA-Seq) (Rajan et al. 2009). At its core, mRNA-Seq involves a high-throughput sequencing-by-synthesis approach to generate short cDNA reads from RNA. These reads are then aligned with a reference genome sequence. By this method, both known and novel splicing events can be identified. The method is, however, sensitive to biases emerging from improper cDNA fragmentation (Bainbridge et al. 2006).

Yet another possibility is to use minigene splicing assays. These come in handy especially when RNA from a patient is not available. Aside from obviating the need of patient RNA, minigene splicing assays also bring about the benefit of avoiding problems brought by NMD and provide a clear picture of a possible causal effect (Di Giacomo et al. 2013). The assay relies on genomic DNA and involves amplification of the target region, cloning it to an expression vector, and transfecting the construct to a suitable cell line. RNA of transfected cells is then extracted, RT-PCR performed, and the resulting cDNA analyzed. Different splicing patterns are generally detected by resolving the RT-PCR-products by gel electrophoresis and sequencing the resulting fragments. (Bonnet et al. 2008; Di Giacomo et al. 2013).

Minigene splicing assays allow the comparison of splicing patterns of wild type and variant sequences in identical conditions, clarifying the effect of the mutation on splicing compared to normal. The target region usually contains the exon and surrounding intronic sequences, but larger constructs containing several exons can also be made. Minigene assays are not without drawbacks, however. As Spurdle et al. (2008) point out, several factors that are present in the gene's natural habitat, such as other genes and cell type-specific elements, which might affect splicing patterns, are absent. Additionally, the many steps in the assay generate technical challenges and vulnerabilities. These assays are often combined with bioinformatic methods to attain reliable results (Spurdle et al. 2008; Abramowicz & Gos 2018). The benefit of this combination approach was demonstrated by Di Giacomo et al. (2013) in their paper on splicing-affecting *BRCA2* exon 7 variants' prevalence.

## 1.4 *ANO7*

### 1.4.1 *ANO7* and the TMEM16 protein family

*ANO7* (*Anoctamin 7*, *TMEM16G, NGEP [New Gene Expressed in Prostate], D-TMPP [Dresden-transmembrane protein of the prostate]*) is located in chromosome 2 band q37.3 (Kiessling et al. 2005). *ANO7* is part of the transmembrane protein 16 (TMEM16) protein family, also known as anoctamins (ANOs) (Guo et al. 2021). The TMEM16 family includes ten homologs (ANO1-10; TMEM16A-K) which all are composed of eight transmembrane domains (Das et al. 2008; Kunzelmann et al. 2019). Anoctamins are found in all mammals and have various functions in cells: some work as ion channels, some as lipid scramblases, while some possess both functions (Guo et al. 2021). For example, *ANO1* is a $Ca^{2+}$-activated $Cl^-$ channel (Caputo et al. 2008), whereas *ANO7* is thought to work as a $Ca^{2+}$-dependent lipid scramblase rather than an ion channel (Suzuki et al. 2013), though controversy remains (Guo et al. 2021). Guo et al. (2021) hypothesize, that *ANO7*, like its homolog *ANO6*, might actually have a dual-function, i.e. work as an ion channel as well as a lipid scramblase.

Kunzelmann et al. (2019) review the various roles of anoctamins in cells. Some anoctamins have been identified to be involved in cell proliferation, but roles of the different homologs differ and even contradict. Indeed, some have been identified to be involved in cell growth, like *ANO1*, whereas some in different forms of cell death (*ANO6*). Cell proliferation is, after all, partly regulated by intracellular $Ca^{2+}$ signals, which are regulated by anoctamins that work as $Ca^{2+}$-regulated ion channels. On the other hand, increased $Ca^{2+}$ levels can also contribute to cell death. Anoctamins have been proposed to be involved in cell proliferation by other mechanisms as well, for example by working as counter-ion channels. The many proposed mechanisms and roles of anoctamins in cells represent the depth of dubiety concerning the TMEM16 protein family.

The function of ANO7 is also debated. Marx et al. (2021) proposed that ANO7 has a role in dedifferentiation of prostate epithelial cells when expression is reduced, whereas Wahlström et al. (submitted manuscript) believe that ANO7 works as a tumor suppressor. ANO7 has also been found to affect cell morphology by forming aggregates, which was proven by preventing said aggregation by RNAi (Das et al. 2007). ANO7's role in vesicle formation, through its scrambling activity (Suzuki et al. 2013), as with its homolog ANO6, has been suggested as well (Kaikkonen et al. 2020). It is evident that the exact function of ANO7 still remains unknown and further research is needed.

It has been ascertained that *ANO7* is androgen-dependent. This is supported by the fact that among PrCa cell lines, *ANO7* transcripts are present only in androgen-dependent cells (LNCaP, VCaP, 22Rv1 and MDA PCa 2b cell lines), but not in PC-3 nor DU145 cell lines (androgen-independent cell lines) (Bera et al. 2004; Kiessling et al. 2005; Das et al. 2007). Kiessling et al. (2005) tested ANO7's androgen-dependence by addition of synthetic androgen to LNCaP cells' growth media. This led to increased expression. *ANO7* mRNA levels are, however, quite low in LNCaP cells, especially compared to patient tissue samples.

*ANO7* is reportedly expressed as two different length mRNA products, which are a result of two splice variants (Bera et al. 2004; Das et al. 2007). Other ANOs have also been reported to express alternatively spliced isoforms. *ANO1*, for example, has been shown to exist as at least four different splice variants (Caputo et al. 2008). An *ANO6* splice variant has also been reported and is associated with breast cancer (Dutertre et al. 2010). The predicted shorter variant of *ANO7*, *ANO7-S*, is encoded by the first four exons and is composed of 179 amino acids. However, its existence has not yet been proven. The longer variant, *ANO7-L*, is derived from all 25 exons, and is 933 amino acids long. (Das et al. 2007; Guo et al. 2021). *ANO7-S* is a consequence of internal polyadenylation, which is a result of a polyA-signal downstream of exon 4 (Bera et al. 2004). Because of this, the fourth and final exon in *ANO7-S* is longer than the corresponding exon in the full-length transcript. Other anoctamins also exhibit shorter transcripts (Hartzell et al. 2009).

Bera et al. (2004) found that the two ANO7 isoforms are localized in different regions in prostate tissue: ANO7-L in the plasma membrane and ANO7-S in the cytoplasm. Similar results regarding ANO7-L localization were obtained by Mohsenzadegan et al. (2013), and Das et al. (2007; 2008) in their subsequent papers. This implies that the absence of transmembrane domains in ANO7-S affects the localization of the protein product. Das et al. (2007) found ANO7-L to be localized especially at cell:cell contact regions in LNCaP cells and suggested that ANO7-L might therefore contribute to cell-cell interactions and possibly cell adhesion, which would explain the previously mentioned aggregate forming by ANO7.

### 1.4.2. *ANO7* and cancer

Many *TMEM16* genes have been documented to be overexpressed in cancer, making them feasible tumor biomarkers (Hartzell et al. 2009; Kunzelmann et al. 2019). Of the anoctamins, *ANO1* in particular has been largely studied in this regard. It has been coined a tumor marker in many cancers, such as pancreatic cancer, prostate cancer, breast cancer and lung cancer, to name a few (Kunzelmann et al. 2019). *ANO1* is significantly upregulated in cancer tissue and is considered a proto-oncogene due to its role in cell proliferation.

*ANO7* has also been associated with PrCa susceptibility (Dadaev et al. 2018; Kaikkonen et al. 2018). As its original name, *NGEP*, implies, *ANO7* is considered a prostate-specific gene. Indeed, many research groups (Bera et al. 2004; Das et al. 2007; Das et al. 2008; Mohsenzadegan et al. 2015) have found that *ANO7* is expressed solely in prostate tissue – benign, malignant and normal. Das et al. (2008) found *ANO7* to be expressed in 91 % of PrCa tissue samples studied, including metastatic tissue, whereas Mohsenzadegan et al. (2013) and Kiessling et al. (2005) detected *ANO7* transcripts in all prostate tissue samples studied. It is notable that *ANO7* transcripts have been found in some other tissues as well, such as the small intestine, colon, liver, and taste buds, as pointed out by Guo et al. (2021), but only in trace amounts (Kiessling et al. 2005; Kaikkonen et al. 2018). The almost exclusive expression of *ANO7* in prostate tissue has made it a promising immunotherapeutic target. Another contributing factor is the fact that *ANO7* is expressed on the cell surface (Bera et al. 2004; Das et al. 2007; Das et al. 2008; Mohsenzadegan et al. 2013).

Findings on *ANO7* expression levels in PrCa tissue have been contradicting. *ANO7* has been found to be under-expressed in PrCa tissue, and that expression levels have an inverse correlation with severity of the disease (Kiessling et al. 2005; Marx et al. 2021), Gleason score (Jhun et al. 2017), pathologic tumor stage and serum PSA levels (Mohsenzadegan et al. 2015). Significant downregulation of *ANO7* mRNA expression has also been found in metastatic prostate tumor samples compared to primary tumors (Chandran et al. 2007). Marx et al. (2021) argue that reduced *ANO7* expression levels could thus be used to reliably determine the aggressiveness of PrCa and predict poor patient prognosis. One research group has, however, found no correlation between *ANO7* expression levels and tumor grade (Das et al. 2008), and great variability in expression levels between individuals has also been documented (Kiessling et al. 2005).

Upregulation of *ANO7* mRNA expression in PrCa tissue has been reported. Kaikkonen et al. (2018) found that *ANO7* mRNA levels are increased in PrCa tissue, compared to other organs and other cancerous tissue. The group found increased *ANO7* expression to be correlated with poor prognosis, as opposed to what Marx et al. (2021) argued, and to what has previously been reported (see previous paragraph). They did not, however, study the difference in expression levels between tumor and normal prostate tissue. These conflicting findings regarding *ANO7* expression level changes in PrCa call for additional research.

Taken together, if *ANO7* were to become an immunotherapeutic target for PrCa, treatment should be adjusted for low-grade and high-grade PrCa patients. Mohsenzadegan et al. (2013) suggest that patients with more advanced PrCa could benefit more from combination therapy, where immunotherapeutic approaches are coupled with traditional therapy. *ANO7*-targeted therapy has been studied to some extent already, as briefly reviewed by Guo et al. (2021). Guo et al. (2021) highlight the versatile use of monoclonal antibodies (mAbs) that could be developed to target cell membrane localized ANO7. ANO7-targeted peptide vaccines have also been suggested. For instance, Cereda et al. (2009) managed to generate ANO7-specific T cells, capable of lysing ANO7 target cells. Even though ANO7 has great potential in terms of PrCa diagnostics and treatment, more research is clearly needed in order to determine the feasibility of ANO7 as a diagnostic marker and an immunotherapeutic target.

### 1.4.3 *ANO7* variants and PrCa

The variant rs77559646 in exon 4 of *ANO7* has been associated with risk of PrCa and predisposes to the aggressive form of PrCa (Kaikkonen et al. 2018). It is a G>A transition which is located five nucleotides downstream of exon 4. This region is part of the unique coding area for *ANO7-S*, as mentioned by Wahlström et al. (submitted manuscript). The variant allele has been shown to result in a missense mutation (Arg104His) (Dadaev et al. 2018). The area in question is also part of the 5' splice region of exon 4. The location of variant rs77559646 thus implies that the variant might have a dual effect, as pointed out by Wahlström et al. (submitted manuscript). Indeed, the variant has been shown to function as a splice site mutation as well. In their study, Wahlström et al. (submitted manuscript) found that the variant allele leads to dysregulated splicing and consequently to reduced ANO7 protein levels. In heterozygote individuals, ANO7 protein levels are still detectable in the apical membrane of prostate tissue due to one functioning wild type allele, whereas homozygous individuals exhibit close to none protein product apically.

Splicing dysregulation caused by variant rs77559646 is most likely a consequence of inadequate base pairing between splicing machinery and the RNA sequence. The U1 snRNP is required to bind the 5'ss for splicing to occur normally. In the reference sequence, one mismatch is already present at position +4, due to an A>C transition relative to the consensus sequence. This mismatch does not, however, disturb regular splicing since the splice site is strong enough for proper splicing regardless. An additional mismatch is introduced by the variant allele A in position +5. Two mismatches then are expected to cause problems with U1 snRNP base pairing.

Aberrant splicing due to variant rs77559646 leads to an increase in intronic RNA and exon 4 skipping (Wahlström et al. submitted manuscript). By comparing the reference and variant constructs, the variant has been shown to cause increased intron 3 expression and reduced splicing of exon 4. The exon definition theory, which states that splicing machinery assembly can occur across the exon, rather than intron (De Conti et al. 2013), could provide an explanation for why intron 3 splicing is altered as well, even though the mutation resides in intron 4. The 'exon definition' model applies when the exon in question is relatively short and the introns long, as in the case with exon 4, and introns 3 and 4. The idea is that communication is favored between splice sites with the shortest distance to one another. Additionally, expression of introns 4 and

5 were also elevated in carriers of variant rs77559646 (Wahlström et al. submitted manuscript).

Ultimately, the amount of ANO7-L protein decreases in the cell dramatically. This is detected at the apical membrane. Since homozygous carriers of the variant exhibit no protein product, contrary to non-carriers and heterozygous carriers, it is thought that the decrease in ANO7-L amount is associated with risk of aggressive PrCa. When examining heterozygotes, Wahlström et al. (submitted manuscript) showed, however, that the level of mRNA is not downregulated between reference and variant, which implies that NMD is not activated. In fact, Kaikkonen et al. (2018) found that mRNA levels of *ANO7* were significantly upregulated in prostate tumors of carriers and linked to poor overall survival. Wahlström et al. (submitted manuscript) suggest the possibility that mutant *ANO7* mRNA resides in the nucleus, rather than cytoplasm, where it would be subjected to NMD.

Another variant, rs78154103 (C>G), found in intron 7 has been linked to a cryptic splicing event and a decrease in the level of normally spliced *ANO7* mRNA, as shown by Wahlström et al. (unpublished). The variant is located 15 nt downstream of exon 7 and is reported to increase the usage of an alternative 5'ss (dinucleotide GC) in intron 7, located 170 nt downstream of exon 7. Aberrant splicing due to use of the cryptic splice donor leads to inclusion of a cryptic exon 7 sequence. The variant is very common in the general population (Ensembl genome browser 104 2021) and always cooccurs with variant rs77559646. This is not true the other way around, however. Unlike variant rs77559646, variant allele rs78154103 has not been associated with increased risk of PrCa.

Transcripts produced by the cryptic splicing event are detectable in homozygous wild types (C/C). However, according to RNA-Seq data, homozygous wild type patients display fewer splicing events from the cryptic donor site (dPSI [delta percent spliced in] = <0.1) than homozygous for the variant allele (G/G) (dPSI = >0.2). The difference between wild type and variant is detectable but subtle when tested experimentally, as previously shown by Wahlström et al. (unpublished) with a minigene splicing assay.

Wahlström et al. (unpublished) hypothesize that the cause of cryptic splicing due to variant rs78154103 is through changes in secondary structure of *ANO7*. Altered RNA secondary structures are known to affect mRNA splicing, mainly by preventing the association of RBPs (De Conti et al. 2013). The exon 7-exon 8 region is believed to form a stem-loop structure, which in turn could be disturbed by variant rs78154103. Alterations in a stem-loop structure can lead to changes in binding of a splicing regulatory protein to the 5'ss if it is located in the stem, as has previously been proven to be the case in intron 10 of Tau protein (Ray et al. 2011). Upon binding to the 5'ss, the protein stabilizes the stem-loop structure, which then prevents binding of U1 snRNP to the 5'ss. Following from this, the splicing machinery rather uses a cryptic splice donor that is accessible, as in this case would be the splice site further downstream of intron 7, ultimately leading to partial intron 7 retention.

In a previous study by Wahlström et al. (unpublished), the effect of variant rs78154103 on the usage of cryptic splice donor in intron 7 was tested using a minigene splicing assay. A construct which contained exon 7 and only a segment of intron 7 coupled with vector intron was designed and used as a minigene. Difference in the level of normally spliced *ANO7* mRNA between the reference and variant constructs was found to be small, yet statistically significant. Wahlström et al. (unpublished) also studied the variant's effect on splicing by using intron 7-specific primers and detected five different splice variants, which were more strongly expressed in the variant construct.

## 1.5 Aims of the study

As mentioned earlier, it is of great importance to identify genes, variants, and mechanisms of action associated with disease development, progression, and resistance. Since aberrant splicing contributes enormously to many disease types, including cancer, it is pivotal to study and understand its causes and consequences. Several genes with aberrant splice variants exhibiting oncogenic properties, and alterations in transcript isoform abundances have been implicated in many cancer types, such as prostate cancer. The search is ongoing and provides novel diagnostic and therapeutic opportunities.

Previously, the inclusion of exon 4 has been attempted to be restored by correcting the variant allele rs77559646. 22Rv1 cell line, a natural carrier of variant rs77559646, was subjected to CRISPR-Cas9 DNA base-editing, by which the variant allele A was converted back to G. In the process, a neighboring A allele in position +3 relative to the exon 4-intron 3 boundary was also converted to G. This study aims to determine whether this correction together with the unintended transition restores normal splicing of *ANO7*. The new splice donor sequence is expected to be strong enough for proper U1 binding and restore normal splicing, which would be seen as a decrease in both exon 4 skipping and intron 3 inclusion. This hypothesis is supported by the fact that in the consensus sequence, the allele in position +3 is a pseudo-uridine nucleotide (Ψ), which is capable of forming a base pair with either purine allele, G or A. To test this, RT-PCR will be performed for clones containing the variant and for clones with the corrected allele(s) to see for differences in intron 3 expression levels. Additionally, a minigene splicing assay will be conducted for a construct containing the base-edited sequence to see the overall change in exon 4 skipping relative to a reference and a variant construct.

As mentioned, the effect of variant rs78154103 on the usage of cryptic splice donor in intron 7 was previously tested using a minigene splicing assay, with a construct containing exon 7 and a synthetic intron 7, comprised of vector and endogenous intron 7. In this study, a longer construct will be designed and examined using a minigene splicing assay. Here, the minigene will contain the entire intron 7, along with surrounding exons. The use of a whole intron in the minigene rather than only a segment allows the RNA to form secondary structures in a similar fashion as endogenously, revealing secondary structure modification's possible contribution to aberrant splicing by variant rs78154103. An additional benefit from using the whole intron comes from avoiding effects caused by accessory vector intron. The minigene splicing assay will be performed using vector- and gene-specific primers.

# 2. Materials and methods

## 2.1 Intron 3 expression level determination

For the evaluation of intron 3 expression levels, CRISPR-Cas9 base-edited 22Rv1 cell line samples, kindly provided by PhD Christopher Löf, were used. Of the clones, five had undergone no change (heterozygous for variant rs77559646), three had undergone correction (+3 G/G and +5 G/G), and one had corrected rs77559646 as well as one of the bystander alleles (+3 A/G and +5 G/G). The clones had previously been stored in 1 ml volumes of TRIsure™ (Meridian Bioscience) at -80ºC. Total RNA was extracted according to TRIsure extraction protocol, with an additional chloroform extraction, and eluated into 40 µl of nuclease-free water. Total RNA was quantified using UV spectrophotometry and samples then stored at -80ºC until cDNA synthesis. Due to inadequate RNA concentrations, two samples were discarded, and seven clones chosen for further analysis. RNA quality of all remaining clones was measured using 2200 TapeStation System (Agilent Technologies).

cDNA conversion from RNA samples was performed using Maxima H Minus First Strand cDNA Synthesis Kit, with dsDNase (Thermo Fisher Scientific), according to protocol, including the additional step 4 of dsDNase inactivation using 10 mM DTT. 1 µg of total RNA was used for a 20 µl reaction. Reverse transcriptase minus (RT-) negative controls were also prepared for two of the randomly chosen cDNA syntheses to assess possible genomic DNA contamination. The primers used for first strand cDNA synthesis were oligo(dT) and random hexamer primers, 0.5 µl each per reaction, both of which were supplied by the kit. Since two primers were used, deviating from manufacturer's protocol, the incubation in step 4 of first strand synthesis was modified to be 10 min at 25ºC, followed by 30 min at 50ºC. All cDNA samples were stored at -80ºC.

PCR was performed using AmpliTaq Gold® 360 Master Mix (Applied Biosystems) and the following cycling conditions: initial denaturation at 95ºC for 10 min, 40 cycles of 95ºC for 30 s, 58ºC for 30 s, and 72ºC for 1 min 30 s, and final extension at 72ºC for 7 min. PCR reactions were made to 25 µl volumes according to manufacturer's protocol, with 2 µl of first strand cDNA synthesis product, corresponding to 100 ng total RNA, and 0.5 µl per 10 µM primer. Two sets of primers were used, in separate PCR reactions. The first pair of primers, ANO7_i3_F1 and ANO7_i3_R1 (see Appendix 1) are designed to amplify a 401 bp region in intron 3. The second pair of primers, ANO7_i3_F2 and ANO7_i3_R2 (Appendix 1) are designed to amplify a 1154 bp region in intron 3. The resulting PCR products were run on 1.2 % agarose gels and photographed using the ChemiDoc MP Imaging System (Bio-Rad). An additional, third PCR reaction was also made using actin amplifying primers ACTB-221-F and ACTB-221-R (Appendix 1), to determine the integrity of the cDNA. The cycling conditions were similar, with the exception of annealing at 52ºC. PCR products were similarly resolved on an agarose gel and photographed.

All fragments from primer pair 2 PCR were excised and purified. DNA extraction was performed using NucleoSpin Gel and PCR Clean-up Mini kit for gel extraction and PCR clean up (Macherey-Nagel). Purified DNA was eluated into 15 µl and concentrations measured using UV spectrophotometry. DNA was stored in -20ºC until Sanger sequencing (Eurofins Genomics).

Sequencing results led to additional cloning of the purified DNA using the StrataClone PCR Cloning Kit (Agilent). Manufacturer's protocol was followed, starting from ligation (step 4 onwards). 2 µl of the cloning reaction mixture was used for transformation to competent cells (supplied by the kit). Deviating from manufacturer's protocol, 10 µl was plated on the other plate, in a total volume of 110 µl with LB medium, and rest of the transformation mixture on the other, in a 100 µl total volume after centrifugation. Multiple single, white colonies were picked and returned to +37ºC overnight. Colony PCR was performed the next day using AmpliTaq Gold® 360 Master Mix (Applied Biosystems). PCR reactions were made to 30 µl total volumes, with 15 µl AmpliTaq master mix, 0.6 µl of 10 µM primers each (T3 and T7; Appendix 1), and 13.8 µl nuclease-free water. Small amounts of bacterial colony mass from the plates were used as template. The following cycling conditions were used: initial denaturation at 95°C for 10 min, 35 cycles of 95°C for 30 s, 52°C for 30 s, 72°C for 60 s, and final extension at 72°C for 7 min. PCR products were visualized on an agarose gel and purified using NucleoSpin Gel and PCR Clean-up Mini kit for gel extraction and PCR clean up (Macherey-Nagel). Concentrations were then measured, and the purified DNA stored in -20ºC until Sanger sequencing (Eurofins Genomics).

A miniprep was made from previously prepared bacterial colonies of one of the samples. The miniprep was prepared using NucleoSpin Plasmid EasyPure (Macherey-Nagel), according to manufacturer's protocol. Concentration of the miniprep was measured and the sample subjected to Sanger sequencing (Eurofins Genomics).

The obtained sequencing results were analyzed using various programs. Sequence quality was first confirmed using BioEdit (Hall 1999). The Basic Local Alignment Search Tool (BLAST) (NCBI) was used to confirm that all sequences originated from *ANO7* by performing a BLAST search against the whole human genome. Further BLAST analyses were performed for some sequences. Additionally, Pairwise Sequence Alignment tools (EMBOSS Water) as well as Multiple Sequence Alignment tools (Kalign, MAFFT and T-coffee) were utilized. All tools are provided by EMBL-EBI. To obtain sensible results, advanced option ‘gap open penalty’ was modified to be high and ‘gap extension penalty’ to be low. Moreover, the intron 3 reference sequence was subjected to RepeatMasker analysis. Results were visualized using SnapGene (Insightful Science; available at snapgene.com).

## 2.2 Plasmid construction

Genomic DNA from base-edited clones of the cell line 22Rv1 was used for plasmid construction. For minigene splicing assay of exon 4, a CRISPR-Cas9 base-edited clone in which a correction of rs77559646 (+3 G/G and +5 G/G) had occurred, was used as template. For splicing assay of intron 7, an unedited clone heterozygous for rs77559646, was used as template.

### 2.2.1 Exon 4 constructs

Primers for amplification of exon 4 and surrounding intronic sequences had previously been designed by PhD Gudrun Wahlström (Wahlström et al. submitted manuscript). The primers ANO7-ex4-F and ANO7-ex4-R (Appendix 1) had been designed to amplify an 806 bp region, consisting of 367 bp of intron 3, the entirety of exon 4 (143 bp) and 296 bp of intron 4, and flanked restriction sites for EcoRI and BamHI. PCR was performed using Phusion® Hot Start Flex DNA Polymerase (New England BioLabs Inc.) in 50 µl total volume and according to the protocol provided by manufacturer. 1 µl of genomic DNA of unknown concentration in DNAreleasy Advance (Nippon Genetics) was used for the reactions. The cycling conditions were the following: initial denaturation at 98ºC for 30 s, 35 cycles of 98ºC for 10 s, 60ºC/65ºC for 15 s, and 72ºC for 15 s, and final extension at 72ºC for 5 min. The resulting PCR fragments were resolved on a 1 % agarose gel. After this, PCR products were purified using NucleoSpin Gel and PCR Clean-up Mini kit for gel extraction and PCR clean up (Macherey-Nagel). Purified DNA was eluated into 30 µl and concentrations measured, after which DNA was stored in -20ºC until cloning.

The amplified region was cloned into the plasmid pSPL3 by first performing digestions for the extracted DNA and plasmid using 5 µl CutSmart™ -buffer (NewEngland

BioLabs) and restriction enzymes EcoRI and SacI (1 µl each). For DNA, the whole remaining volume was used, and for pSPL3 plasmid, 0.5 µl DNA, corresponding to 1.7 µg, was used. Reactions were filled to 50 µl with MQ water. The digestion reactions were incubated at 37ºC for 1 h. After incubation, digestion products were resolved on a 0.8 % agarose gel. Correct sized fragments were excised and purified using the above-mentioned kit, and eluated into 15 µl.

Ligation of insert to vector was done using 50 ng of vector and the Rapid DNA Ligation Kit (Thermo Fisher Scientific), according to manufacturer's protocol. Vector and insert volumes were calculated using the following formula:

$$\frac{vector\ bp}{insert\ bp} = \frac{vector\ ng}{insert\ ng}$$

The result was further multiplied by x3 to achieve a 3:1 insert:vector ratio. Ligation reactions were filled up to a final volume of 20 µl with Milli-Q water. A control reaction was also made, which did not contain an insert.

The plasmids were transformed into NEB 5-alpha Competent *E. coli* cells (NewEngland BioLabs), using 5 µl ligation product per 50 µl competent cells. The cells were plated on two agar plates containing ampicillin (100 µg / ml), so that the first plate contained 100 µl of the transformation mixture and the second contained rest of the mixture. After growth at +37ºC overnight, two colonies were picked from the 100 µl plate for further growth in 2 ml LB-medium containing ampicillin (100 µg / ml). These bacterial cultures were then grown overnight at 37°C and used to make minipreps the following day using NucleoSpin Plasmid EasyPure (Macherey-Nagel), according to manufacturer's protocol. Proper insertion of the amplified region was confirmed by linearization using 5 µl miniprep and restriction enzymes EcoRI and BamHI (1 µl each), and 5 µl FastDigest buffer (Thermo Fisher Scientific), to a total volume of 20 µl. Digestion reactions were incubated at 37ºC for 15 min, after which they were run on 1 % agarose gels to ensure correct insertion.

After confirmation of correct insertion, pure cultures were made from both clones. After growth overnight at 37ºC on agar plates containing ampicillin (100 µg / ml), further cultures were made with 5 ml LB-medium containing ampicillin (100 µg / ml). Bacterial cultures were returned to 37°C and incubated overnight by shaking at 225 rpm. The next day, glycerol stocks with 60 % glycerol (1/3 of total volume) were prepared from the bacterial cultures to final volumes of 1.8 ml. Both clones were done in duplicates and stored in -80ºC.

After all inserts were confirmed using BioEdit (Hall 1999) by aligning with the desired plasmid construct, midipreps were prepared to ensure sufficiently pure DNA for transfection. Some glycerol stock, from a sample proven correct by sequencing, was scraped off the top and suspended into 3 ml LB-medium containing ampicillin (100 µg / ml) and incubated for six hours at 37 °C by shaking at 225 rpm. Absorbance was measured using Jenway™ 7200 Visible Scanning Spectrophotometer (Jenway). Based on the absorbance values, proper amounts of the cultures were transferred to 100 ml cultures of LB-medium containing ampicillin (100 µg / ml) and left on overnight growth as above. The next day, absorbance was measured similarly. Midipreps were then prepared using NucleoBond Xtra Midi Plus kit (Macherey-Nagel) by following the provided protocol. DNA was eluated into 300 µl and concentrations measured. For the minigene splicing assay, dilutions with concentration of 1 µg / µl were made. To further ensure that the correct plasmid was obtained, a test digestion was done using 0.5 µl of midiprep, corresponding to 1 µg DNA, 2 µl FastDigest buffer, and restriction enzymes EcoRI and BamHI (1 µl each), to a total volume of 20 µl. Digestion reactions were incubated at +37ºC for 15 min after which they were visualized on a 1 % agarose gel. DNA was stored in -20ºC.

### 2.2.2 Intron 7 constructs

Intron 7 plasmid construction was performed for the most part as exon 4 plasmid construction (see chapter 2.2.1). For intron 7 plasmid constructs, primers were designed with the help of Primer3Plus (bioinformatics.nl) and Primer-BLAST (NCBI) tools. The primers, ANO7-in6-3-F and ANO7-in8-1-R (Appendix 1) were designed to amplify a 1523 bp region containing exon 7, intron 7 and exon 8, as well as 210 bp of upstream (intron 6) and 266 bp of downstream (intron 8) intronic sequences. The primers included sites for EcoRI and EagI. As with construction of exon 4 plasmid constructs, PCR was performed using Phusion® Hot Start Flex DNA Polymerase (New England BioLabs Inc.), following manufacturer's protocol. The cycling conditions differed from exon 4 plasmid construction and were the following: initial denaturation at 98ºC for 30 s, 35 cycles of 98ºC for 10 s, 65ºC for 15 s, and 72ºC for 30 s, and final extension at 72ºC for 5 min. Gel electrophoresis and DNA extraction were performed as in Chapter 2.2.1.

Cloning of the amplified region to the plasmid pSPL3 was done using restriction enzymes EcoRI and EagI, and CutSmart™ -buffer (New England BioLabs Inc.). The reactions were incubated at 37ºC for 1 h. For DNA, all remaining template was used, and for pSPL3 plasmid, 10 µl DNA was used (standard miniprep, concentration unknown). Reactions were filled to 50 µl with Milli-Q water. The following steps were done as in chapter 2.2.1, with the exception of picking six colonies for the 2 ml bacterial cultures since intron 7 constructs are heterozygous, and so to assure that both the variant and reference were picked. Additionally, restriction enzyme EagI was used instead of BamHI in test digestions of mini- and midipreps. Midipreps were made from two construct's glycerol stocks, proven correct by sequencing. Midipreps of intron 7 constructs were eluated into 400 µl, instead of 300 µl, and stored in -20ºC.

## 2.3 Minigene splicing assay

Transfection of plasmid constructs was carried out on mycoplasma-negative COS-7 cells. The cells were grown in DMEM (Lonza; exp. June 2020), supplemented with 10 % FBS (Biowest), 100 units/ml penicillin and 100 µg/ml streptomycin (Lonza), and 2 mM UltraGlutamineTM I Supplement (Lonza). The cells were cultured in T75-flasks and grown in 5 % $CO_2$ conditions at +37ºC. Cells were splitted twice a week and used within 15 passages since thawing.

COS-7 cells were plated onto 6-well plates with two cell amounts (1.5 x $10^5$ cells / well and 2.25 x $10^5$ cells / well; 5 ml medium / well) and transferred back to 5 % $CO_2$ conditions at +37ºC overnight. Plates with confluency best corresponding to that recommended by the transfection kit's protocol were selected for transfections the next day. The cells were transfected with 2.5 µg DNA and 7.5 µl Lipofectamine 3000 Reagent (Thermo Fisher Scientific; exp. March 2020), according to manufacturer's protocol. The previously prepared exon 4 construct (Ex4-5), as well as two exon 4 constructs provided by PhD Gudrun Wahlström containing the reference (Ex4-2) and the variant sequences (Ex4-4), in addition to intron 7 constructs (In7-1 and In7-2) were used as DNA to be transfected. All five constructs were done in duplicates. Additionally, a GFP-positive control was prepared. After incubation at 37ºC overnight, cells were inspected using the EVOS M5000 Imaging System (Thermo Fisher Scientific) to determine transfection efficiency by measuring GFP fluorescence. The transfected cells were then lysed in TRIsure (Meridian Bioscience) in a 1 ml volume and stored at -80°C until RNA extraction. RNA isolation was performed following manufacturer's instructions, with an additional chloroform extraction. RNA was eluated into 40 µl nuclease-free water, concentrations measured, and stored in -80ºC until cDNA conversion.

One reference intron 7 construct sample in assay 3 had an inadequate concentration, and so an additional RNA precipitation was performed for it and its duplicate. First, 0.1 volumes of 3 M NaAc was added and the samples vortexed. Then, 2.5 volumes of ice-cold 99 % EtOH was added, and the samples incubated in -20ºC for 7 h. After incubation, the precipitate was centrifuged down with a speed of 15,000 x g at +4ºC for 10 minutes. The supernatant was removed and 500 µl of cold 70 % EtOH added on the pellet. The samples were vortexed briefly and the centrifugation step repeated. The supernatant was again removed carefully, after which the pellet was let to air-dry for approximately 10 minutes at room temperature. The pellet was then dissolved into 15 µl of nuclease-free water, concentrations measured, and RNA stored in -80ºC.

For cDNA synthesis, 1 µg of total RNA was used. Reverse transcription was performed using an oligo(dT) primer and the Maxima H Minus First Strand cDNA Synthesis Kit, with dsDNase (Thermo Fisher Scientific) into 20 µl reaction volumes. PCR was performed using AmpliTaq Gold® 360 Master Mix (Applied Biosystems). Primers dUSD2 and dUSA4 (Appendix 1) were in 10 µM concentrations and used in 0.5 µl volumes. The total volume per PCR reaction was 25 µl and the cycling conditions were as follows: initial denaturation at 95ºC for 10 min, 40 cycles of 95ºC for 30 s, 60ºC for 30 s, and 72ºC for 30 s, and final extension at 72ºC for 7 min. The resulting PCR products were separated on a 1.2 % agarose gel. The minigene splicing assay protocol was performed a total of three times.

From one of the intron 7 reference constructs of the second splicing assay, fragments were excised and extracted using NucleoSpin Gel and PCR Clean-up Mini kit for gel extraction and PCR clean up (Macherey-Nagel). The purified DNA was eluated into 15 µl and concentrations measured. DNA was then stored in -20ºC until Sanger sequencing (Eurofins Genomics).

The sequencing results were first analyzed using BioEdit (Hall 1999) to determine sequence quality. After this, a BLAST search using the ‘align two sequences’ feature of BLAST against the minigene construct of intron 7 was done. The results were then visualized using SnapGene (Insightful Science; available at snapgene.com).

Intensities of bands in splicing assay gel photos were measured using ImageJ software (Schneider et al. 2012). From these, means and standard deviations between all splicing assays were calculated using Excel. Statistical analysis of differences between means was performed using the analysis of variance (ANOVA) test (SAS Enterprise 7.1 software). Graphs, in which means were represented as fractions, were constructed using Excel.

Splice site strengths of the different exon 4 constructs, as well as intron 7 constructs' canonical and cryptic splice sites were determined using the maximum entropy model (MAXENT). The algorithm scores splice sites up to a score of 14.

### 2.3.1 Intron 7-specific RT-PCR

cDNA of intron 7 constructs of each splicing assay were used as template. PCR was performed using AmpliTaq Gold® 360 Master Mix (Applied Biosystems) and primers ANO7-crex7-3-F and dUSA4-C (Appendix 1), provided by PhD Gudrun Wahlström. The primers amplify a 1779 bp region of the intron 7 plasmid construct containing 758 bp of intron 7, exon 8 (111 bp), 286 bp of intron 8, as well as 624 bp of vector. PCR reactions were made according to manufacturer's protocol, with 2 µl of first strand cDNA synthesis product, corresponding to 100 ng total RNA, and 0.5 µl per 10 µM primer in 25 µl total volumes. The cycling conditions were as follows: initial denaturation at 95ºC for 10 min, 35 cycles of 95ºC for 30 s, 61ºC for 30 s, and 72ºC for 20 s, and final extension at 72ºC for 7 min. PCR products were resolved on a 1.5 % agarose gel and the resulting fragments were excised from one lane only. DNA was extracted using the NucleoSpin Gel and PCR Clean-up Mini kit for gel extraction and PCR clean up (Macherey-Nagel). DNA was eluated into 15 µl and stored in -20ºC until Sanger sequencing (Eurofins Genomics).

Some samples had to be subjected to additional cloning using the StrataClone PCR Cloning Kit (Agilent). The whole procedure was performed as in Chapter 2.1. Analysis of sequencing results was performed as in Chapter 2.3.

# 3. Results

In this study the impact of an unintentional conversion, a byproduct of base-editing conducted to correct variant allele rs77559646, on *ANO7* splicing was examined and characterized. More specifically, the effect of the correction(s) was studied in terms of intron 3 expression level changes and the extent of exon 4 inclusion. First, the change in intron 3 retention was studied by RT-PCR conducted on clones that had undergone base-editing successfully and to those still containing the variant. Second, the change in exon skipping was examined by conducting a minigene splicing assay.

Another objective of this thesis was to determine the effect of variant rs78154103 on *ANO7* splicing. The splicing pattern of the surrounding region was studied by conducting a minigene splicing assay. Vector-specific as well as gene-specific primers were used.

## 3.1 Intron 3 expression level determination

The first part of this study was to evaluate the level of intron 3 expression between the different 22Rv1 cell line clones. RNA extraction and cDNA conversion were first performed to all seven clones. Specific regions in intron 3 were then amplified by RT-PCR using two different primer pairs, and the PCR products resolved on an agarose gel (Fig. 1). The first set of primers were designed to amplify a 401 bp region in intron 3, which five samples were observed to express, as seen in Figure 1a. An additional faint band, approximately 200 bp in size, was also observed in samples A10 and B7. The second set of primers were to amplify a 1154 bp region in intron 3. Figure 1b shows that five clones had expressed the correct sized fragment, yet two additional bands were also observed in each lane. Altogether only five out of seven clones exhibited intron 3 expression. RIN values revealed that the RNA integrity of the two clones (A4 and A5), which failed to express intron 3, was poor (RIN = 5.3 and RIN = 5.4). RIN-values of the other five clones were decent (RIN = >8). Overall, the goal was to seek major differences in intron 3 expression between the different base-edited clones. However, results revealed that the extent of endogenous intron 3 expression did not differ drastically between clones (Fig. 1).

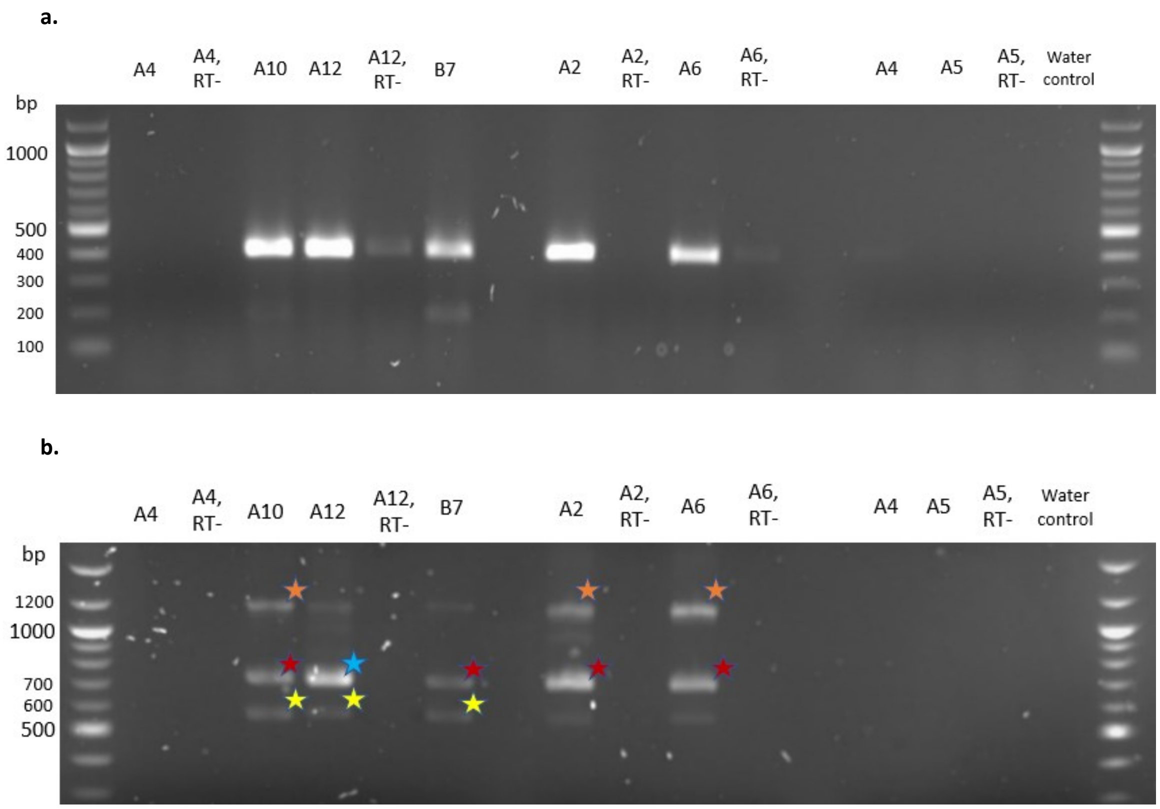


**Figure 1.** Expression of intron 3 in clones subjected to RT-PCR, visualized on a 1.2 % agarose gel. RT- = Reverse transcriptase minus negative control. 1 kb molecular weight marker was used as ladder. **a.** PCR products of primer pair 1 RT-PCR. Lane 5 (A12, RT-) most likely contains runoff from neighboring well. **b.** PCR products of primer pair 2 RT-PCR. *Blue star* denotes SEQ29. *Red stars* denote SEQ30. *Yellow stars* denote SEQ31. *Orange stars* denote SEQ32.

All denoted fragments from the gel shown in Figure 1b were extracted and purified. Corresponding fragments from different clones were pooled together (see Fig. 1b cutline). Reamplifications were performed to some of the fragments to obtain adequate concentrations. Furthermore, unsatisfactory sequencing results from the first rounds of sequencing led to cloning of the fragments, as well as miniprep construction of one of the samples. After these, a total of five samples (SEQ29-3, SEQ29-5, and SEQ30—32) were subjected to sequencing analysis. Reamplified fragments, from which initial sequencing was performed, are seen in Figure 2a and fragments obtained from colony PCR, and from which final sequencing was conducted, are seen in Figure 2b and 2c. SEQ32 was sequenced directly from the plasmid, due to unsuccessful colony PCR.

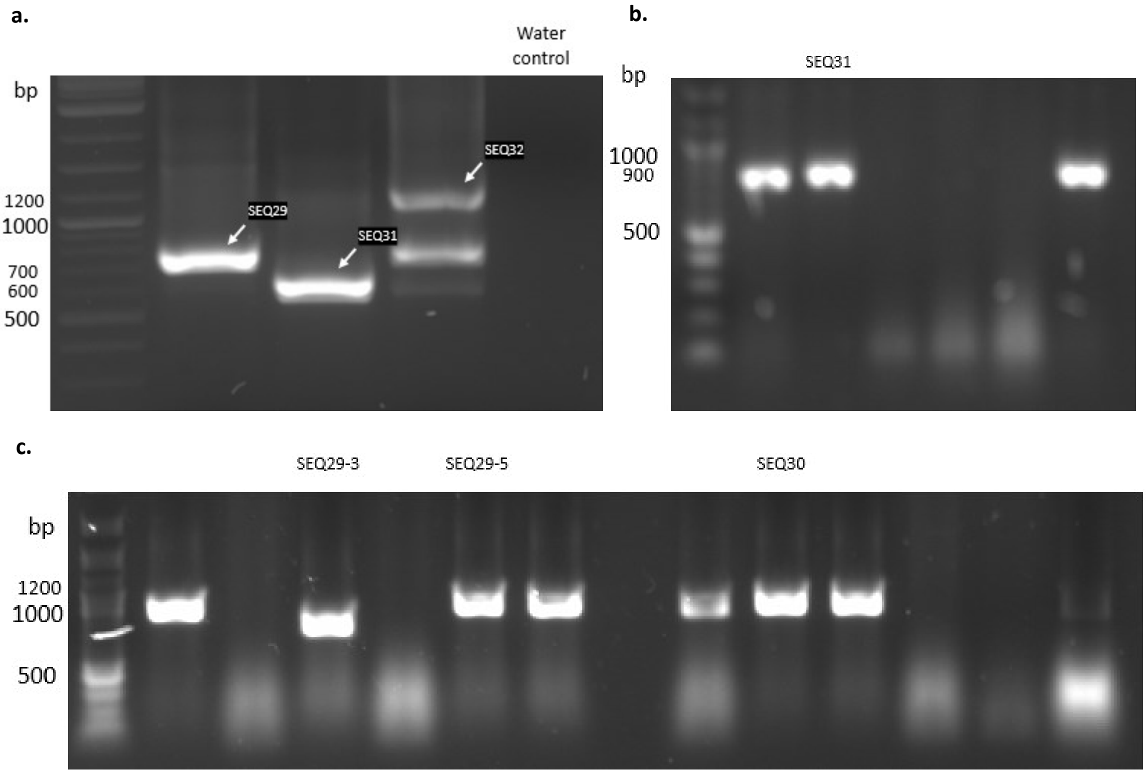


**Figure 2.** Reamplification and colony PCR products visualized on agarose gels. 1 kb molecular weight marker was used as ladder. **a.** Reamplified fragments SEQ29, SEQ31 and SEQ32 on a 1.2 % agarose gel. **b.** Colony PCR products of SEQ31 on a 1.5 % agarose gel. **c.** Colony PCR products of SEQ29 and SEQ30 on a 0.8 % agarose gel. Due to size differences, two SEQ29 fragments were excised (SEQ29-3 and SEQ29-5).

The anticipated sequence lengths and actual sequencing results are provided in Table 1. Sequencing revealed that each sample contained a region of intron 3, but four out of five also exhibited deletions. What is more, two samples, SEQ29-3 and SEQ31, were identical in sequence, despite size difference in the original gel (Fig. 1b). SEQ29-3 is probably a consequence of contamination, which had most likely occurred midst gel extraction of reamplified PCR products (Fig. 2a). Additionally, two samples (SEQ29-5 and SEQ30) were identical in length, as predicted (Fig. 1b). SEQ32 comprised the entire desired intron 3 sequence as proven by a BLAST search, apart from some individual variations (see next chapter). Taken together, multiple different transcripts were produced from the intron 3 sequence. Furthermore, all intron 3 fragments sequenced were in length what was anticipated originally, as evident from Table 1.

**Table 1.** Samples subjected to sequencing from intron 3 RT-PCR.

| Sequence | Origin | Primers | Fragment length in gel | Fragment length obtained by sequencing |
|---|---|---|---|---|
| SEQ29 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 700 bp (original) | See SEQ29-3, SEQ29-5 below |
| SEQ29-3 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 800 bp (after cloning)* | 577 bp |
| SEQ29-5 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 1000 bp (after cloning)* | 724 bp |
| SEQ30 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 700 bp (original) / 1000 bp (after cloning)* | 715 bp |
| SEQ31 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 600 bp (original) / 900 bp (after cloning)* | 577 bp |
| SEQ32 | Primer pair 2 RT-PCR | ANO7_i3_F2 and _R2 | 1200 bp | 1149 bp |
| | | | * = 192 bp cloning vector included | |

### 3.1.1. Defining sequence deletions

All sequencing results were BLAST searched against the human genome as well as intron 3 reference sequence. Search results revealed multiple matches in the genome. Therefore, repeat elements in intron 3 were deduced by RepeatMasker analysis. The repeat elements are depicted in Figure 3. Figure 3 also illustrates the locations of deleted regions of the four samples, compared to the reference sequence of intron 3 and the intron's repeat elements. The cut sites in each fragment were observed to be located in the *AluSx* as well as the *AluY* repeat sequences of intron 3. This was also detected when conducting a MAFFT analysis for all sequences, aligned against the intron 3 reference sequence (Appendix 2). To be more precise, the first cut site in each fragment was found in the former, and the second cut site in the latter repeat element. The deletion in samples SEQ29-3 and SEQ31 was 577 bp in length, whereas SEQ29-5 contained a 430 bp gap, and SEQ30 a 439 bp gap. SEQ29-3 and SEQ31 appeared to be identical, most likely due to contamination during extraction from gel, as mentioned above. It is notable that SEQ29-3/SEQ31 and SEQ30 contained a short identical sequence at both cut sites, making it hard to determine the exact location of the cut. No common splice sites were found at the cut sites.

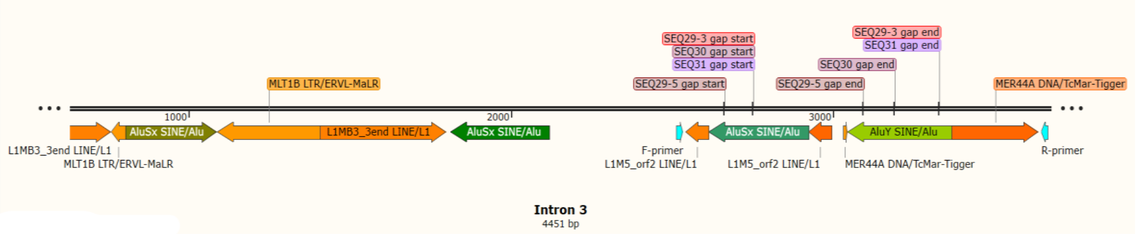


**Figure 3.** A schematic representation of the *ANO7* intron 3 sequence, including repeat elements and locations of sequence deletions. The repeat elements' locations, denoted by *green* and *orange arrows*, were determined using RepeatMasker. *Turquoise arrows* denote positions of primers used in RT-PCR.

BLAST searches and other analyses revealed that SEQ32 gave two matches, i.e. demonstrated high sequence similarity to two regions. Additionally, SEQ32 contained multiple mismatches (single nucleotide deviations) with the reference intron 3 sequence. Thus, *Alu* elements of SEQ32 were further analyzed (Appendix 3). Due to the number of mismatches in the first repeat sequence of SEQ32, the SEQ32 sequence was separately BLAST searched against the *AluSx* sequence of reference intron 3. This analysis revealed that the beginning of the sequence was a near-perfect match to *AluSx*, whereas multiple mismatches were observed towards the end, implying that the sequence originated from somewhere else. SEQ32 was then BLAST searched against the *AluY* sequence of reference intron 3. This search revealed two matches. Here, the first match seemed to be identical to that acquired with *AluSx*. The beginning of the sequence contained multiple mismatches and the end was a near-perfect match to *AluY*, opposite to what was observed with *AluSx*. The second repeat sequence in SEQ32, i.e. the second match when BLAST searched against *AluY*, was a near-perfect match to *AluY*, with only one mismatch and a small 3 nt deletion. Taken together, the first *Alu* repeat sequence in SEQ32 comprised of both *AluSx* and *AluY* sequences, and the second *Alu* in SEQ32 corresponded to *AluY* (Appendix 3).

### 3.2 Minigene splicing assay

To analyze how the unintended alteration near variant rs77559646 affects splicing of exon 4, a minigene splicing assay was conducted. Minigenes with the reference (Ex4-2), variant (Ex4-4), and base-edited (Ex4-5) sequences were generated by amplifying the target region containing exon 4 and flanking introns and cloning the sequence to a vector plasmid pSPL3. The minigenes were transfected into COS-7 cells, and the resulting RNA subjected to RT-PCR. PCR was conducted using vector-specific primers and the PCR products were resolved on an agarose gel (Fig. 4). Two distinct bands, sized 404 bp and 261 bp, could be observed in all constructs, as anticipated. The identities of the fragments have previously been reported by Wahlström et al. (submitted manuscript). They showed that the 404 bp product represents correct splicing of exon 4, whereas the 261 bp fragment represents complete exon 4 skipping. An unknown, weak fragment approximately 650 bp in size was also detected in the reference construct in this study, but which was not further analyzed.

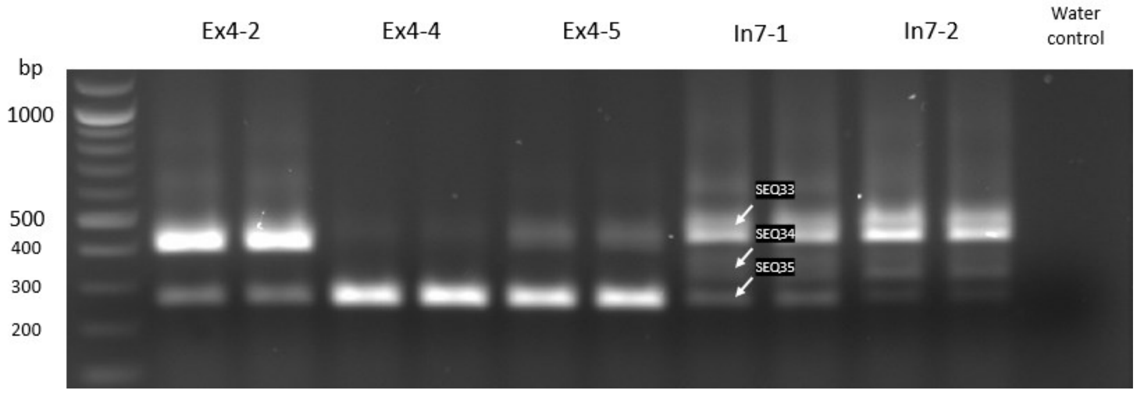


**Figure 4.** RT-PCR products of the second minigene splicing assay visualized on a 1.2 % agarose gel. All constructs are made in duplicates. 1 kb molecular weight marker was used as ladder. Ex4-2 = Reference construct. Ex4-4 = Variant construct. Ex4-5 = Base-edited construct. In7-1 = Variant construct. In7-2 = Reference construct. Excised and sequenced fragments from construct In7-1 are depicted.

Intensities of the two resulting bands were first analyzed across all splicing assays using ImageJ and Excel. Figure 5 illustrates the mean values of exon inclusion and exon skipping in the constructs, expressed as fractions. The results revealed that all three exon 4 constructs displayed exon skipping, but to varying extents. Variant allele A led to significantly higher levels of exon skipping compared to reference, as previously demonstrated by Wahlström et al. (submitted manuscript). Overall, complete exon skipping was clearly more prevalent than exon inclusion in Ex4-4 and Ex4-5. The base-edited Ex4-5 did exhibit improved exon 4 inclusion compared to Ex4-4 but did not reach the inclusion levels seen in Ex4-2. Statistical analyses performed using the ANOVA test supported these results (Appendix 4). Indeed, results revealed that the difference in exon 4 inclusion and skipping between the three constructs was statistically highly significant (p = <.0001). Additionally, MAXENT values support the observed pattern: the reference construct gave the highest score (MAXENT = 10.65), indicating that the canonical 5’ss is the strongest. The variant had the lowest score (MAXENT = 7.79), whereas the base-edited (MAXENT = 9.73) a slightly improved score. Taken together, conversion of the variant allele A back to G, accompanied with an unintentional conversion A>G nearby, resulted in a small, yet statistically highly significant change in exon inclusion.

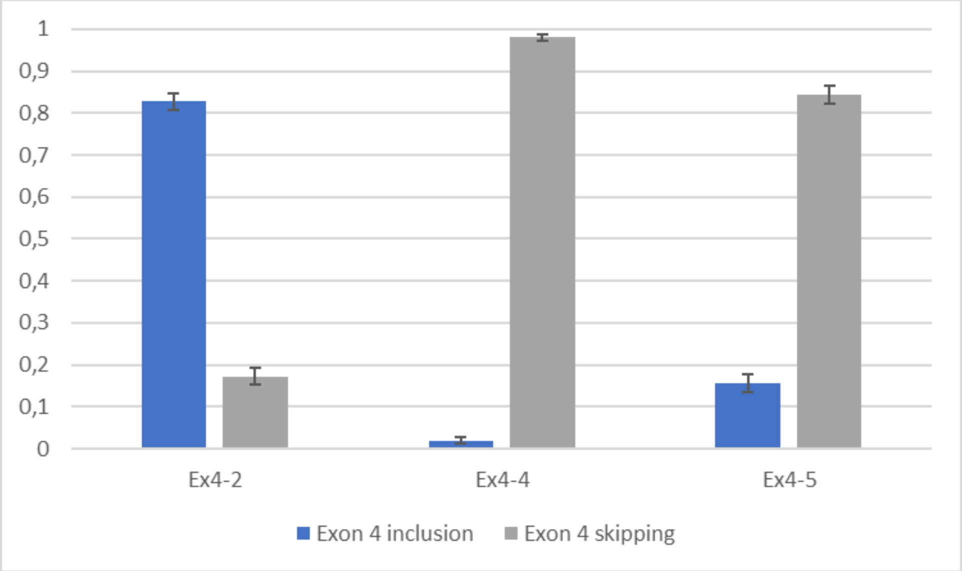


**Figure 5.** A bar graph representing the mean expression of exon 4 inclusion and skipping in exon 4 constructs of the minigene splicing assay. The Y-axis demonstrates the fraction (0—1) of expression. Ex4-2 = Reference construct. Ex4-4 = Variant construct. Ex4-5 = Base-edited construct. Standard deviations are included in the bars.

Figure 4 also depicts splicing assay results of intron 7 construct analysis. The variant allele rs78154103 carrier (In7-1) and reference (In7-2) constructs both resulted in three distinct products, sized 427 bp, 316 bp and 258 bp, as revealed by sequencing. The sequencing results are seen in Table 2 and Figure 6, with Figure 6a demonstrating the reference construct sequence. As seen from Figure 4, the largest fragment (SEQ33) is the most strongly expressed in both constructs. Sequencing revealed that SEQ33 represents exon inclusion, with both exons 7 and 8 retained in the final transcript (Fig. 6b) as a result of using canonical splice sites surrounding the exons. The middle fragment’s (SEQ34) expression seemed to differ only slightly between the two constructs. SEQ34 demonstrated aberrant splicing with only exon 7 retained, and exon 8 excluded (Fig. 6c). Figure 4 revealed that the smallest fragment (SEQ35), which exhibited complete exon 7 and 8 skipping (Fig. 6d), was least expressed in In7-2, but not in In7-1. To verify these observations, intensities of the bands were measured and analyzed as in exon 4 constructs.

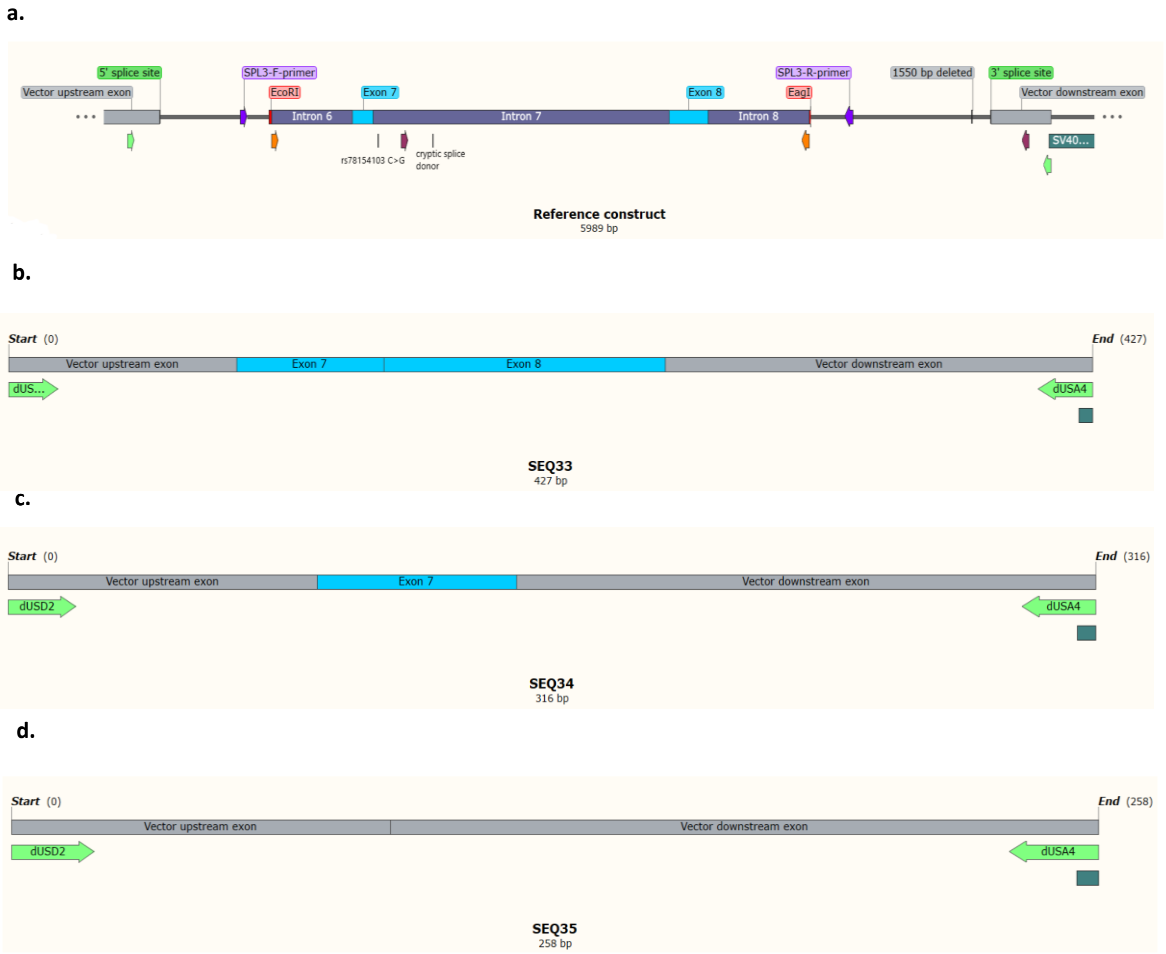


**Figure 6.** A schematic representation of intron 7 splicing assay sequencing results. *Green arrows* denote positions of dUSD2 and dUSA4 primers used in splicing assay RT-PCR. The far-right sequence contains the SV40_PA_terminator. **a.** The pSPL3-intron 7 -minigene construct, which against the sequencing results were aligned, showing all elements. *Orange arrows* denote primers used in amplification of the insert region. *Purple arrows* denote positions of vector-specific primers (SPL3-F and SPL3-R). *Plum arrows* denote primers ANO7-crex7-3-F and dUSA4-C used in the second intron 7 RT-PCR (see Chapter 3.2.1). **b.** SEQ33 sequence containing 90 bp vector upstream exon, exon 7 (58 bp), exon 8 (111 bp) and vector downstream exon (168 bp). **c.** SEQ34 containing 90 bp vector upstream exon, exon 7 (58 bp) and vector downstream exon (168 bp). **d.** SEQ35 containing 90 bp vector upstream exon and 168 bp vector downstream exon.

Figure 7 illustrates the variation in exon inclusion between the intron 7 constructs. As mentioned, the correctly spliced transcript, exhibiting exon 7 and 8 inclusion, was the most prevalent in both constructs. What is more, the fragment is more strongly expressed in In7-2 compared to variant In7-1 (Fig. 7). Statistical analysis (Appendix 4) revealed that the difference was highly significant (p = <.0001). Exon 8 skipping, however, did not differ much between the constructs (p = 0.0631). Conversely, exon 7 and 8 skipping exhibited statistically highly significant (p = <0.0001) differences between the constructs. All in all, the different constructs showed a clear difference in inclusion of exons 7 and 8.

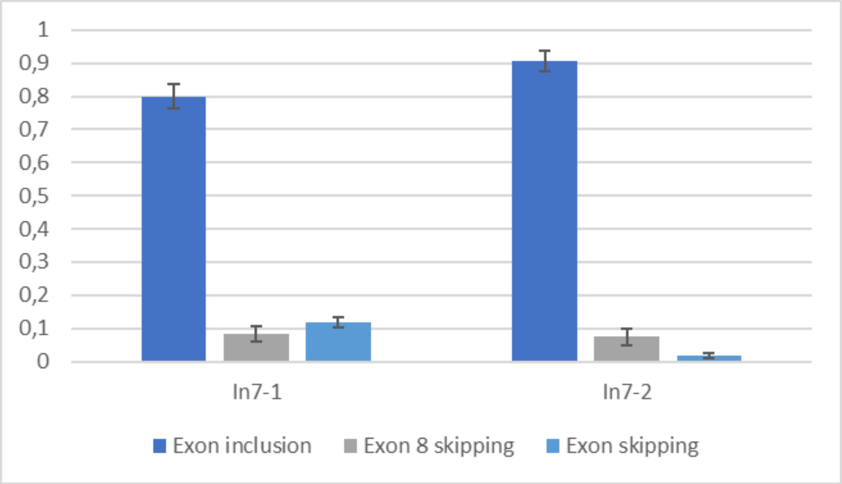


**Figure 7.** A bar graph representing the mean expression of exon 7 and 8 inclusion, exon 8 skipping, and exon 7 and 8 skipping in intron 7 constructs of the minigene splicing assay. The Y-axis demonstrates the fraction (0—1) of expression. In7-1 = Variant construct. In7-2 = Reference construct. Standard deviations are included in the bars.

MAXENT scores of the canonical 5'ss of intron 7 and the cryptic splice site located 170 nt downstream exon 7 indicated that the cryptic splice site is not stronger than the natural one. The scores were 7.35 and -0.49, respectively.

### 3.2.1 Intron 7 -specific

To demonstrate how the cryptic splice site in intron 7 is utilized, an intron 7-specific primer along with a vector-specific primer were used to analyze the intron 7 constructs. Figure 8 illustrates the obtained RT-PCR products on an agarose gel. Both constructs, In7-1 and In7-2 produced three fragments, which were approximately 450 bp, 350 bp and 310 bp in size (Fig. 8a). Additionally, a fourth very faint band, approximately 400 bp, could be seen. All fragments were observed to be expressed in similar quantities across all samples. The three distinct fragments (SEQ36—38) were excised and purified from one of the reference In7-2 samples (Fig. 8a), and SEQ37 and SEQ38 subjected to sequencing. The largest band (SEQ36) was reamplified (Fig. 8b). Here, the fourth band, barely visible in Figure 8a, was clearly visible. The band was not extracted, however. Direct sequencing of fragments SEQ36 and SEQ37 did not yield decent results, and so cloning was performed. Products from colony PCR (Fig. 8c) were then extracted and subjected to sequencing.

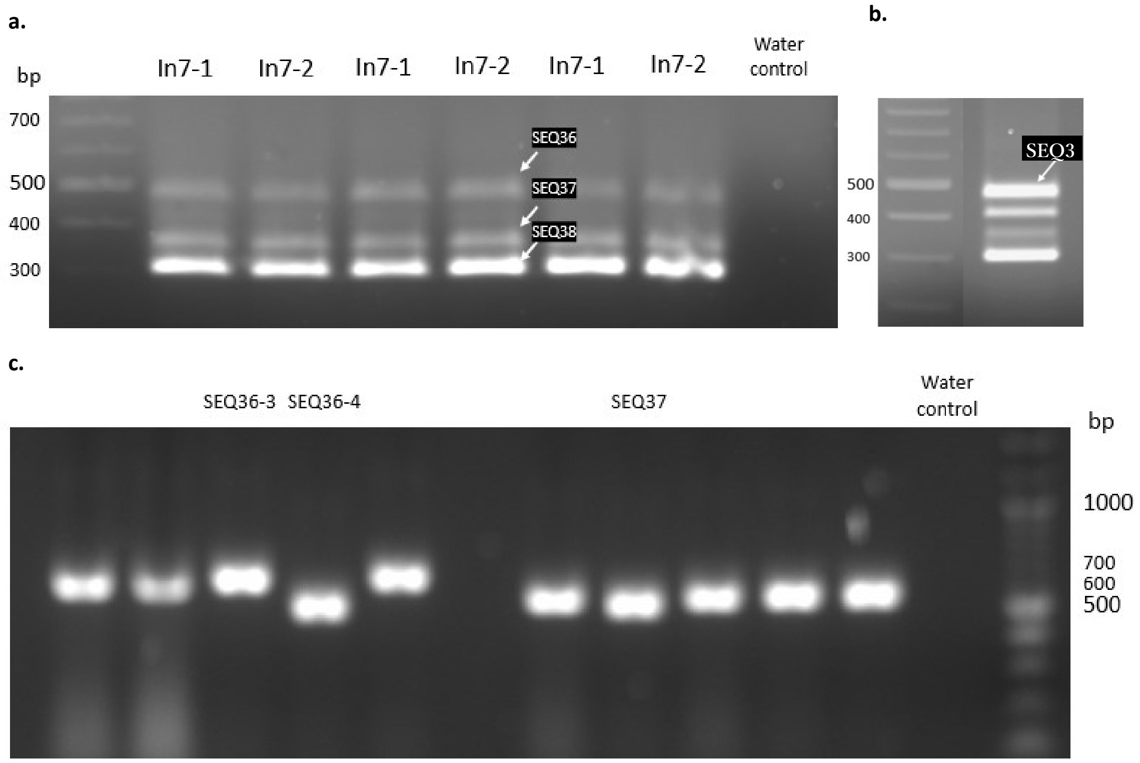


**Figure 8.** PCR products of intron 7-specific splicing assay RT-PCR visualized on a 1.5 % agarose gel. 1 kb molecular weight marker was used as ladder. **a.** The original PCR products of all intron 7 constructs from every splicing assay (1st, 2nd, 3rd assay, respectively). **b.** Products of the reamplification of SEQ36. **c.** Colony PCR products of SEQ36 and SEQ37. Due to size differences, two SEQ36 fragments were excised (SEQ36-3 and SEQ36-4).

Table 2 and Figure 9 illustrate the sequencing results. The intron 7 minigene construct (Fig. 6a) was used as reference to which all sequences were aligned against. All fragments were proven to include 89 bp of intron 7, exon 8 (111 bp), and 108 bp of vector downstream exon. 89 bp of intron 7 is a result of using the cryptic splice site. SEQ36-3 also included 116 bp of vector cryptic exon. Here the splicing had occurred in the vector – a region normally spliced out. The smallest fragment (SEQ38) corresponded to SEQ36-3 with the exception of absent vector cryptic exon. Therefore, SEQ38 is result of using canonical splice sites in exon 8 and vector, in addition to the cryptic splice site in intron 7. Surprisingly, SEQ36-4 and SEQ37 were also identical to the aforementioned despite size difference in gel (Fig. 8). SEQ36-4 and SEQ37 were quite possibly a result of contamination. The results thus imply that the largest (SEQ36-3) and smallest (SEQ38) fragments were successfully sequenced, whereas the middle fragment (SEQ37) could not be obtained.

**Table 2.** Samples subjected to sequencing from splicing assay.

| Sequence | Origin | Primers | Fragment length in gel | Fragment length obtained by sequencing |
|---|---|---|---|---|
| SEQ33 | Splicing assay | dUSD2 and dUSA4 | 430 bp | 427 bp |
| SEQ34 | Splicing assay | dUSD2 and dUSA4 | 330 bp | 316 bp |
| SEQ35 | Splicing assay | dUSD2 and dUSA4 | 260 bp | 258 bp |
| SEQ36 | Splicing assay | ANO7-crex7-3-F and dUSA4-C | 450 bp (original) | See SEQ36-3, SEQ36-4 below |
| SEQ36-3 | Splicing assay | ANO7-crex7-3-F and dUSA4-C | 700 bp (after cloning)* | 424 bp |
| SEQ36-4 | Splicing assay | ANO7-crex7-3-F and dUSA4-C | 600 bp (after cloning)* | 306 bp |
| SEQ37 | Splicing assay | ANO7-crex7-3-F and dUSA4-C | 350 bp (original) / 600 bp (after cloning)* | 306 bp |
| SEQ38 | Splicing assay | ANO7-crex7-3-F and dUSA4-C | 310 bp | 306 bp |
| | | | * = 192 bp cloning vector included | |

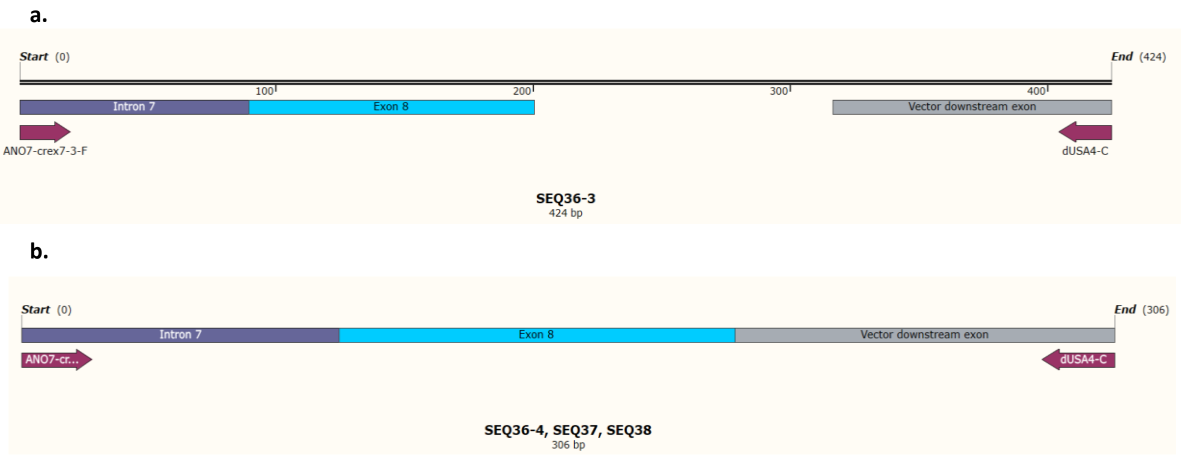


**Figure 9.** A schematic illustration of the second intron 7 RT-PCR's sequencing results. The sequences were aligned against the intron 7 minigene construct seen in Figure 6a. Primers ANO7-crex7-3-F and dUSA4-C used in RT-PCR are denoted by *plum arrows*. **a.** SEQ36-3 containing 89 bp of intron 7, exon 8 (111 bp), 116 bp vector cryptic exon, and 108 bp vector downstream exon. **b.** SEQ36-4, SEQ37 and SEQ38 containing 89 bp of intron 7, exon 8 (111 bp), and 108 bp vector downstream exon.

# 4. Discussion

This thesis work focused on revealing the impact of two mutations located in the genomic region of *ANO7*, a prostate cancer susceptibility gene, on pre-mRNA splicing. More specifically, two aims were sought after: one focusing on a variant (rs77559646) and an unintentional alteration in intron 4 of *ANO7*, and the other on a variant (rs78154103) in intron 7.

The variant rs77559646 (G>A) has previously been associated with the aggressive form of PrCa (Kaikkonen et al. 2018). The variant allele A has been documented to lead to aberrant splicing, which can be seen as exon 4 skipping and increased intron 3 retention (Wahlström et al. submitted manuscript). To correct the variant allele back to G, 22Rv1 cell line clones had been subjected to CRISPR-Cas9 base-editing. However, the treatment also led to conversion of another A allele to G, 2 bp upstream of rs77559646. Both positions, +3 and +5 relative to the exon-intron boundary, are located in the splice donor site of intron 4. This study aimed to determine whether a) the level of intron 3 expression has decreased in clones subjected to base-editing, and b) how the new mutation in position +3 behaves in a minigene splicing assay.

Another variant, rs78154103 located in intron 7, has been linked to a cryptic splicing event as evident from RNA-Seq data (Wahlström et al. submitted manuscript). The variant has been shown to activate a cryptic splice donor site in intron 7, which consequently leads to partial intron retention. The reason for why a cryptic donor site is favored, remains to be elucidated. Changes in secondary structure due to the variant is a proposed explanation. In this study, splicing patterns of the region in question were analyzed using a minigene splicing assay in which a construct containing the entire intron 7 along with surrounding exons was used, contrary to what had previously been done (Wahlström et al. unpublished). In their study, Wahlström et al. (unpublished) constructed a minigene containing only part of endogenous intron 7. By including the entire intron 7 here, secondary structures similar to those formed naturally in the cell are allowed to form, which provides clues as to whether secondary structures affect splicing of the region.

## 4.1 rs77559646

Using traditional RT-PCR to determine the level of intron retention in the cell between successfully base-edited clones and clones carrying the variant allele rs77559646, did not yield anticipated results. It was thought that if intron retention was a consequence of splicing defects caused by the variant, correction of the mutation would lead to decreased intron expression. Indeed, it was expected that intron 3 expression would normalize after correction of the variant allele, regardless of the unintentional conversion in position +3. The new mutation +3G is thought to allow proper binding of U1 snRNP and consequently promote proper splicing, since according to a consensus sequence, +3G and +3A are both favored (Roca et al. 2008). This was tested using two different primer pairs that amplify a specific region in intron 3. RT-PCR products showed no notable differences in intron 3 expression between the clones using either primer pair after visualization on an agarose gel. Some minor differences were detected between clones, but there was no pattern in relation to whether the clone had a base-edited sequence or not. These alterations were most likely due to varying RNA qualities. Additionally, two clones produced no PCR products. The lack of expression was found to be a result of poor RNA quality. It is possible that the reason for why this was the case might be that something happened to the samples during the extraction process, such as overheating of samples. It is also possible that the clones were already different during cell culture. In the future, it would be wise to take extra care in making sure all clones are subjected to identical treatment conditions.

### 4.1.1 Intron 3 sequence deletions

Fragments from the second primer pair's RT-PCR were subjected to Sanger sequencing to determine the origin of the incorrect sized fragments. Surprisingly, sequencing results revealed deletions in three fragments, in addition to a correct sequence. The number of different deletions suggests that some unknown mechanism is the cause. If only two events were observed, the whole region and one with a deletion, it might have been possible that germline mutations are to blame. However, in this case that explanation seems unlikely due to the occurrence of more than two different deletions. A more likely explanation is the presence of somatic deletions. It is also worth considering whether these deletions have occurred in the DNA, or post-transcriptionally. One possible way to examine this is by sequencing chromosomal DNA from said cell line, to elicit whether the deletions have occurred on the chromosomal level.

It is notable that the cut sites did not display common splice site dinucleotides, and so the deletions are not a cause of splicing due to presence of cryptic splice sites. However, the samples exhibiting deletions contained short, identical sequences at both cut sites. It is uncertain whether these duplications are a cause of the deletions. As the sites exhibit microhomology, occurrence of recombination is a possible explanation for such deletions. The observation that two samples exhibited identical deletions, and some deletions began at the same site supports this notion. Then again, it is possible that these are purely coincidental. The locations of the breakpoints, however, provide another angle for determining the cause of the deletions.

Intron 3 contains a high number of *Alu* repeat elements, as revealed by RepeatMasker analysis. The cut sites were all situated in *Alu* elements of intron 3, implying that *Alu* elements mediated the deletions. The mechanism in question is called *Alu*/*Alu*-mediated rearrangement (AAMR) (Song et al. 2018), also termed *Alu*-mediated recombination. This theory is further supported by the observation that primer pair 1 PCR products exhibited no variants, possibly a consequence of the absence of repeat elements in the region the primers amplify. It is also notable, that the region is much shorter.

*Alu* elements are transposable elements, more specifically non-LTR (long terminal repeats) retrotransposons, and part of the SINE (short interspersed elements) family. They are found throughout the genome and are considered the most abundant repeat elements (Pastor et al. 2009). Additionally, *Alu*s can be further subdivided to *AluJ*, *AluS* and *AluY* elements (in descending order of evolutionary age), which share high sequence similarity (Song et al. 2018). *Alu* elements are known to affect mRNA-splicing through disrupting splicing regulatory elements by insertion to introns, or by gain of function (Pastor et al. 2009). They also regularly contribute to diseases (Boone et al. 2014).

Fundamentally, the mechanism of action in *Alu*-mediated recombination is debated. Some suggested mechanisms include nonallelic homologous recombination (NAHR), microhomology-mediated end joining (MMEJ) and microhomology-mediated break-induced replication (MMBIR). Homologous recombination, however, is an unlikely explanation since it requires a longer stretch of homology (~300—500 bp), whereas *Alu* elements themselves are only approximately 300 bp in length (Boone et al. 2014; Song et al. 2018). The two other aforementioned mechanisms, that are based on homeology (partial homology) and/or are microhomology-mediated are far more plausible, as proven by Boone et al. (2014). Microhomologies at the breakpoints can be as small as 2 bp according to data by Boone et al. (2014) and Song et al. (2018).

The cut sites were located in *AluS*x and *AluY* sequences of intron 3, both of which are relatively young *Alu* elements. According to Song et al. (2018), younger *Alu*s mediate AAMR events more frequently, for they share higher similarity between each other than with older *Alu*s. Additionally, the two sequences were directly oriented, and in antisense orientation relative to the intron 3 sequence. The mutual orientation of *Alu*s participating in the AAMR event determines the type of rearrangement taking place (Majer & Sikora 2021). If the *Alu*s are in opposite orientation, the result is an inversion, where the region between the *Alu*s is inverted. Conversely, when the *Alu*s located in the same chromosome are both in the same orientation, as in this case, the result is a deletion or duplication. Boone et al. (2014), for instance, detected that breakpoint junctions of sequenced CNVs in the *SPAST* gene were mostly located in *Alu*s that were in direct orientation. An *Alu*-mediated deletion results in a chimeric hybrid-*Alu* comprising of half-and-half of both *Alu*s involved in the AAMR event, and the hybrid junction contains the microhomology region (Boone et al. 2014; Song et al. 2018). This was indeed observed in the results of this study.

Deletions caused by *Alu*s are recognized in the literature, but all focus on deletions of larger size than observed here. Searching '*Alu*-mediated deletions' mentioned in the title in PubMed yielded merely 13 articles in total. Deletions in these studies were most often a few kb in size, in contrast to the few hundred bp seen here. Song et al. (2018), however, report deletions ranging from 836 bp to as long as >4 Mb. In most studies the deletions were coined disease-causing. One such study performed by Nyström-Lahti et al. (1995), recognized a 3.5 kb genomic deletion, which corresponded to a 165 bp deletion in cDNA, and which was predisposing to hereditary colon cancer. This deleted segment corresponded to a whole exon. The mutation responsible for this deletion is presumably caused by *Alu*-mediated recombination. It is possible that smaller deletions occurring within an intron, such as the ones in this thesis study, have gone undetected and thus no literature on the subject is found. Nevertheless, *Alu*-mediated recombination seems like the most likely explanation for the results obtained here.

The sample containing the entire amplified region (SEQ32) revealed multiple mismatches with the intron 3 reference sequence. These mismatches could be a result of PCR mutations or preexisting SNPs, but these explanations are improbable due to the vast number of mismatches in SEQ32. Additionally, most mismatches seemed to concentrate on one *Alu*. Therefore, the *Alu* elements of the sample were looked at in more detail. The first *Alu* element (*AluSx*) of the sample did not perfectly align with the intron 3 reference sequence, yet the second *Alu* element (*AluY*) was a near-perfect match with *AluY* of intron 3, as it should. Surprisingly, the latter part of the first *Alu* aligned perfectly with the second *Alu* (*AluY*). The reason for this is unknown, and to the best of my knowledge, no such phenomena have been documented before. Involvement of AAMR events cannot be ruled out. In the scope of this thesis, it can only be stated that the fragment did not entirely correspond to the reference sequence.

Even though this study did not reveal notable differences in expression levels of intron 3, it is still possible that subtle changes are present. Possible changes could be detected using quantitative RT-PCR (RT-qPCR), which simultaneously measures the concentration of the cDNA sequence being amplified. RT-qPCR is ideal for splice variant detection, for it measures the ratios of different variants directly. Another benefit of RT-qPCR is that it suits well for analysis of mRNA found in low-abundance (Bustin 2000). This comes in handy when dealing with *ANO7*, which cell lines express only moderately. Bustin (2000) lists additional benefits of using RT-qPCR in detecting splice variants. Even so, based on the results obtained from the minigene splicing assay, which revealed that base-editing had not restored proper splicing, it was not reasonable to further analyze intron 3 expression of the clones using RT-qPCR in this study. Additionally, RT-qPCR effectiveness could be disturbed due to the deletions.

### 4.1.2 Minigene splicing assay

The mutation that resulted from base-editing was also studied by conducting a minigene splicing assay. A minigene construct carrying exon 4 and surrounding sequences was generated for the assay, and the final RT-PCR products visualized on an agarose gel. By comparing the products of the base-edited construct to those of a reference and variant, it was evident that the extent of exon skipping was altered. It was expected that exon 4 inclusion would improve, as predicted by a MAXENT analysis. That indeed was the case, yet only moderately. The idea behind the hypothesis was that U1 snRNP would bind more efficiently to the modified sequence than to the variant sequence.

The U1 snRNP binds to the splice donor site when it contains a specific sequence. Correction of variant rs77559646 restores a G (+5G) needed for the binding domain, whilst the conversion 2 nt upstream results in another G (+3G). The latter nucleotide in question can be an A or a G for proper binding to occur, for the U1 snRNP consensus sequence exhibits a pseudo-uridine nucleotide (Ψ) in position +3. A Ψ can base pair with either A or G (Roca et al. 2008). +3A and +3G are both reported to be highly conserved (Madsen et al. 2006; Roca et al. 2008). Thus, it was expected that correct splicing would be restored regardless of the unintentional conversion.

Despite a high level of conservation, many +3A>G mutations have been identified as disease-causing (Madsen et al. 2006; Roca et al. 2008). For instance, Tzetis et al. (2001) showed that a splice donor site mutation 621+3A>G in the *CFTR* gene reduces expression of normal *CFTR* mRNA. Consequently, the result is a more severe phenotype. *CFTR* (cystic fibrosis transmembrane conductance regulator) is the gene responsible for cystic fibrosis, a severe inherited disease mostly affecting the lungs, but also other organs. Drawn from this, it is evident that +3G does not always work as efficiently as +3A. Whether this is a direct consequence of disturbed base pairing with U1 snRNP, was not elucidated in this study. Conversely, search for +3G>A mutations in the literature was not as successful.

Indeed, it was observed that improvement in inclusion of exon 4 occurred, but only slightly. Perhaps in this case a G instead of A is not as suitable for some reason. It could be possible that the answer lies in the preexisting mismatch +4C – perhaps an interaction between the two positions disrupts splicing. Indeed, it has long been thought that the positions in the 5'ss are independent, but some modern analysis methods desert this idea of independence. Roca et al. (2008) show that there is indeed dependence between the different positions of the 5'ss and that these pairwise associations are important for, for example, U1 base pairing. Therefore, if this interaction between positions is disrupted, correct splicing is disturbed.

Roca et al. (2008) demonstrate different combinations of modifications in the 5'ss sequence that are unfavorable. For instance, they showed that the combination +3G and +4C causes aberrant splicing, a combination seen in the base-edited sequence studied here. The +3G in their study was successfully activated when the +4C was corrected back to consensus +4A. Roca et al. (2005; 2008) argue that correction in position +4 is enough to restore correct splicing, for it presumably contributes to the stability of the wobble G-Ψ base pair at +3. +3A provides stable enough base pairing with U1 even in the presence of non-consensus alleles in other position, such as +4C, but +3G requires consensus alleles in other positions (Madsen et al. 2006; Roca et al. 2008). Drawn from this, the mutated +3G produced by base-editing would probably not have a disadvantageous effect on splicing if the allele in position +4 were consensus, i.e. +4A.

It is also worth mentioning that U6 snRNP, also required for correct splicing, base pairs with the splice donor site, similarly as U1. U6 binds to the 5'ss after dissociation of U1. Binding of U6 is more limited, for the sequence motif responsible for base pairing is even more highly conserved than that of U1 (Roca et al. 2008). Involvement of U6 in splicing dysregulation here seems an unlikely explanation, however, since, even though the sequence that underwent base-editing introduces two mismatches (+4C and +5G) to the highly conserved U6 sequence, the reference and variant provide an even more incompatible match, with three mismatches (+3A, +4C and +5G(reference)/+5A(variant)) introduced by both. Conclusions of the involvement of U6 in this case, nor U1 for that matter, cannot be made without experimental testing.

Taken together, it is evident that the obtained results did not fully support the hypothesis that correct splicing would be restored despite introduction of the new mutation. Correction of variant allele rs77559646 does restore correct splicing, but this recovery seems to be impeded by the new splicing mutation in position +3, which most likely exerts its adverse power through interaction with the preexisting +4C. In conclusion, correction of variant allele rs77559646 (G>A) by base-editing is not reasonable due to its only minor impact on restoring correct splicing. In addition, the intron 3 region clearly exhibits a lot of noise, which does not make it a suitable candidate region for analysis of expression levels. It is essential for the region used in such studies to stay intact, which was proven not to be the case. Deletion of segments leads to a wrong kind of expression level reduction.

## 4.2 rs78154103

The intron 7 variant allele rs78154103 was similarly studied using a minigene splicing assay. A construct carrying the variant allele was compared to a reference construct. Results revealed that both constructs expressed the same splice variants when using vector-specific primers, although the extent of exon inclusion varied. The transcripts revealed inclusion of exons 7 and 8, skipping of both, and surprisingly skipping of only exon 8. Exon skipping was more frequent in the variant construct. No cryptic exon 7 was detected, at least to an extent that would be visible when visualized on an agarose gel, contrariwise to what was expected. This suggests that use of the cryptic splice site is more prevalent in patient samples, where its utilization is ~20 %, as revealed by dPSI scores. If cryptic splice site use were to be as frequent as in patient samples, a detectable PCR product would have been obtained.

Skipping of only exon 8 in the middle fragment was not expected. It is possible that the reason for such a transcript is that the splicing complex was unable to form properly. The cause of this is unknown. Additionally, the middle fragment in the gel might actually comprise two separate fragments. This would have been revealed by running the gel longer.

Cryptic exon 7 generated by use of the cryptic splice donor was, however, detected when using an intron 7-specific primer in addition to a vector-specific primer. Using a primer that binds to intron 7 results in amplification of only transcripts containing the intron. Additionally, exon 8 was included in the final transcripts obtained from this RT-PCR. These results were confirmed by sequencing. All in all, three, as well as possibly a fourth very faint one, fragments were produced, of which two were successfully sequenced. The middle fragment was not obtained, possibly due to contamination by the lower, more intense band. Therefore, more care should have been taken while excising bands. Also, selection of the desired sized fragment after colony PCR would have been more reliable if the gel were run for a longer period of time and/or more thought would have been put to which band to excise.

The fragments were subjected to sequencing to determine their composition. As mentioned, sequence of the middle fragment was not obtained, but the largest and smallest fragments were successfully sequenced. The smallest fragment exhibited cryptic exon 7 and exon 8 inclusion, as expected. The largest fragment, on the other hand, revealed an unexpected vector cryptic exon, in addition to cryptic exon 7 and exon 8, indicating that cryptic splice sites in the vector were utilized. An explanation for why such splice sites have been selected was not found in the literature. One may speculate that splicing in the region occurs slow, and thus the vector intron sequence was not spliced out in time.

In their study, Wahlström et al. (unpublished) detected five fragments after using the same intron 7-specific primer for their minigene construct. Sequencing revealed the same cryptic exon 7 as obtained here. Results between these two studies deviated, however, because exon 8 was included in the construct used in this study. Exon 8 is surrounded by active, normally working splice sites. This explains why the results in these studies differ in terms of transcript number and composition.

Surprisingly, cryptic splice site use was as prevalent in the reference construct as in the variant, as observed from the gel where RT-PCR products were visualized. Wahlström et al. (unpublished), however, detected intensity differences between reference and variant: expression of all fragments was lower in the reference constructs, indicating that cryptic splice site usage is more infrequent. This was the expectation for results in this study. It is plausible that, as with intron 3 expression measurements, RT-qPCR could have revealed possible differences in expression. However, it is evident that differences would nevertheless not be as considerable as in patient RNA-Seq data reported by Wahlström et al. (unpublished).

The 5’ss of reference intron 7 is already relatively weak for a splice site, as revealed by MAXENT analysis. The cryptic splice site downstream, in turn, gives a negative score, evidently indicating that it is weaker. However, it is notable that the cryptic splice site contains a noncanonical dinucleotide GC, which presumably perverts the MAXENT score. Be that as it may, the results of this study showed that the cryptic splice site is favored in some instances and cryptic splice variants are produced. Therefore, prediction programs such as MAXENT do not tell the whole story. This contradiction suggests that the reason for cryptic splice selection in intron 7 is not caused by differences in strength of the different splice sites, but rather by some other mechanism.

A suggested explanation is that changes in RNA secondary structure lead to alternative splice donor site use in intron 7 (Wahlström et al. unpublished). Modified RNA secondary structures are known to regularly affect splicing (De Conti et al. 2013; Abramowicz & Gos 2018; Olender & Lee 2019). To find out whether the region in question altogether forms specific secondary structures, such as a stem-loop as suggested for intron 7, RNA folding algorithms exist that are designed to predict secondary structures formed by the provided sequences (e.g. mFold and RNAfold). Secondary structures can also be studied by *in vitro* functional experiments, such as by using antisense RNA (Donahue et al. 2006). Since modified secondary structures can contribute to aberrant splicing and disease, targeting these structures is yet another possible therapeutic approach.

Indeed, changes in RNA secondary structure could be the reason for cryptic splice site selection in intron 7. Such a mechanism has been reported with the microtubule binding protein Tau. Exon 10 of *MAPT*, the gene encoding Tau, is regularly alternatively spliced to form different isoforms of the Tau protein. These isoforms are maintained in a constant ratio by splicing regulation, but changes in splicing due to mutations disturb this ratio and contribute to tauopathy (Donahue et al. 2006). Tauopathy is a term referring to neurodegenerative disorders where Tau is involved. One region in particular, the exon 10-intron 10 boundary, is prone to mutations that affect exon 10 splicing. *MAPT* pre-mRNA is shown to form a stem-loop structure in this region. Hence, mutations in this area affect the stem-loop structure, leading to changes in binding of a specific splicing regulatory protein to the 5’ss located in the stem (Ray et al. 2011). Upon binding to the 5’ss, the protein stabilizes the stem-loop structure, which in turn prevents binding of U1 snRNP. This ultimately leads to exon 10 exclusion. Mutations in the area usually lead to increased exon inclusion, however, by disrupting the stem-loop structure’s stable formation (Donahue et al. 2006). Jiang et al. (2000)

have shown that even one SNV is enough to disrupt regular splicing of exon 10 of the *MAPT* gene by altering the secondary structure.

Wahlström et al. (unpublished) suspect that intron 7 also forms a similar stem-loop structure, which is affected by variant rs78154103. When the variant is present, the stem-loop is predicted to be in a more stable conformation. This in turn would prevent binding of the U1 snRNP to the natural 5’ss, and rather uses a cryptic splice donor further downstream of intron 7. This theory requires experimental verification.

In conclusion, the results of this study support the hypothesis that variant allele rs78154103 (C>G) leads to cryptic splice site selection in intron 7. However, the difference in splicing between variant and reference was not as remarkable as anticipated. Another aim was to determine the splice variants produced by cryptic splice site usage, which was accomplished adequately. Moreover, this study demonstrates the utility of using a longer construct, containing an entire intron, instead of a shorter one with only part of an intron when determining splicing patterns of such a region.

## Acknowledgments

First and foremost I would like to thank my supervisors Gudrun Wahlström and Christina Nokkala for all the help and guidance I received during this thesis project. No question of mine went unanswered, no matter the nature. I would also like to thank Johanna Schleutker and her research group for making this study possible, both financially and in terms of assistance. Special thanks to Christoffer Löf for providing me the material studied here and Jukka Karhu for always helping me whenever needed.

My family deserves a big thanks for being supportive throughout my studies and always believing in me. I would not have gone this far without them. An ambivalent thanks to my dalmatian puppy, as well, who continually challenged me during this journey – with good intentions, I'm sure.

# References

Abramowicz A & Gos M (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics* 59, 253—268.

Antonarakis E S, Armstrong A J, Dehm S M, Luo J (2016). Androgen receptor variant-driven prostate cancer: clinical implications and therapeutic targeting. *Prostate Cancer Prostatic Dis.* 19, 231—241.

Bainbridge M N, Warren R L, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis E R, Sadar M D, Siddiqui A S, Marra M A, Jones S J M (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7.

Bera T K, Das S, Maeda H, Beers R, Wolfgang C D, Kumar V, Hahn Y, Lee B, Pastan I (2004). NGEP, a gene encoding a membrane protein detected only in prostate cancer and normal prostate. *PNAS* 101, 3059—3064.

Bonnal S C, López-Oreja I, Valcárcel J (2020). Roles and mechanisms of alternative splicing in cancer — implications for care. *Clinical oncology* 17, 457—474.

Bonnet C, Krieger S, Vezain M, Rousselin A, Tournier I, Martins A, Berthet P, Chevrier A, Dugast C, Layet V, Rossi A, Lidereau R, Frébourg T, Hardouin A, Tosi M (2008). Screening *BRCA1* and *BRCA2* unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. *Journal of Medical Genetics* 45, 438—446.

Boone P M, Yuan B, Campbell I M, Scull J C, Withers M A, Baggett B C, Beck C R, Shaw C J, Stankiewicz P, Moretti P, Goodwin W E, Hein N, Fink J K, Seong M-W, Seo S H, Park S S, Karbassi I D, Batish S D, Ordóñez-Ugalde A, Quintáns B, Sobrido M-J, Stemmler S, Lupski J R (2014). The *Alu*-Rich Genomic Architecture of *SPAST* Predisposes to Diverse and Functionally Distinct Disease-Associated CNV Alleles. *The American Journal of Human Genetics* 95, 143—161.

Bratt O (2002). HEREDITARY PROSTATE CANCER: CLINICAL ASPECTS. *The Journal of Urology* 168, 906—913.

Bustin S A (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology* 25, 169—193.

Caputo A, Caci E, Ferrera L, Pedemonte N, Barsanti C, Sondo E, Pfeffer U, Ravazzolo R, Zegarra-Moran O, Galietta L J V (2008). TMEM16A, A Membrane Protein Associated with Calcium-Dependent Chloride Channel Activity. *Science* 322, 590—594.

Cereda V, Poole D J, Palena C, Das S, Bera T K, Remonbo C, Gulley J L, Arlen P M, Yokokawa J, Pastan I, Schlom J, Tsang K Y (2009). New gene expressed in prostate: a potential target for T cell-mediated prostate cancer immunotherapy. *Cancer Immunol Immunother* 59, 63—71.

Chandran U R, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, Michalopoulos G, Becich M, Monzon F A (2007). Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* 7, 64.

Cooper T A, Wan L, Dreyfuss G (2009). RNA and Disease. *Cell* 136, 777—793.

Dadaev T & Saunders E J, et al. (2018). Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nature communications* 9, 2256.

Das S, Hahn Y, Nagata S, Willingham M C, Bera T K, Lee B, Pastan I (2007). NGEP, a Prostate-Specific Plasma Membrane Protein that Promotes the Association of LNCaP Cells. *Cancer Res* 64, 1594—1601.

Das S, Hahn Y, Walker D A, Nagata S, Willingham M C, Peehl D M, Bera T K, Lee B, Pastan I (2008). Topology of NGEP, a Prostate-Specific Cell:Cell Junction Protein Widely Expressed in Many Cancers of Different Grade Level. *Cancer Res* 68, 6306—6312.

De Conti L, Baralle M, Buratti E (2013). Exon and intron definition in pre-mRNA splicing. *WIREs RNA* 4, 49—60.

Di Giacomo D, Gaildrat P, Abuli A, Abdat J, Frébourg T, Tosi M, Martins A (2013). Functional Analysis of a Large set of *BRCA2* exon 7 Variants Highlights the Predictive Value of Hexamer Scores in Detecting Alterations of Exonic Splicing Regulatory Elements. *Wiley Human Mutation* 34, 1547—1557.

Donahue C P, Muratore C, Wu J Y, Kosik K S, Wolfe M S (2006). Stabilization of the Tau Exon 10 Stem Loop Alters Pre-mRNA Splicing. *J Biol Chem* 281, 23302—23306.

Duterte M, Lacroix-Triki M, Driouch K, de la Grange P, Gratadou L, Beck S, Millevoi S, Tazi J, Lidereau R, Vagner S, Auboeuf D (2010). Exon-Based Clustering of Murine Breast Tumor Transcriptomes Reveals Alternative Exons Whose Expression Is Associated with Metastasis. *Cancer Res* 70, 896—905.

Faustino N A & Cooper T A (2003). Pre-mRNA splicing and human disease. *Genes and Development* 17, 419—437.

Fehlbaum P, Guihal C, Bracco L, Cochet O (2005). A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Research* 33, e47.

Guo J, Wang D, Dong Y, Gao X, Tong H, Liu W, Zhang L, Sun M (2021). ANO7: Insights into topology, function, and potential applications as a biomarker and immunotherapy target. *Tissue and Cell* 72, 101546.

Hall, T A (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95—98.

Hartzell H C, Yu K, Xiao Q, Chien L-T, Qu Z (2009). Anoctamin/TMEM16 family members are $Ca^{2+}$-activated $Cl^{-}$ channels. *The Journal of Physiology* 587, 2127—2139.

Heinlein C A & Chang C (2004). Androgen Receptor in Prostate Cancer. *Endocrine Reviews* 25, 276—308.

Hjelmborg J B, Scheike T, Holst K, Skytthe A, Penney K L, Graff R E, Pukkala E, Christensen K, Adami H-O, Holm N V, Nuttall E, Hanse S, Hartman M, Czene K, Harris J R, Kaprio J, Mucci L A (2014). The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol Biomarkers Prev* 23, 2303—2310.

Hörnberg E, Ylitalo E B, Crnalic S, Antti H, Stattin P, Widmark A, Bergh A, Wikström P (2011). Expression of Androgen Receptor Splice Variants in Prostate Cancer Bone Metastases is Associated with Castration-Resistance and Short Survival. *PLoS ONE* 6, e19059.

Jiang Z, Cote J, Kwon J M, Goate A M, Wu J Y (2000). Aberrant Splicing of tau Pre-mRNA Caused by Intronic Mutations Associated with the Inherited Dementia Frontotemporal Dementia with Parkinsonism Linked to Chromosome 17. *Molecular and Cellular Biology* 20, 4036—4048.

Jhun M A, Geybels M S, Wright J L, Kolb S, April C, Bibikova M, Ostrander E A, Fan J-B, Feng Z, Stanford J L (2017). Gene expression signature of Gleason score is associated with prostate cancer outcomes in a radical prostatectomy cohort. *Oncotarget* 8, 43035—43047.

Jung H, Lee D, Lee J, Park D, Kim Y J, Park W-Y, Hong D, Park P J, Lee E (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature Genetics* 47, 1242—1251.

Kaikkonen E, Rantapero T, Zhang Q, Taimen P, Laitinen V, Kallajoki M, Jambulingam D, Ettala O, Knaapila J, Boström P J, Wahlström G, Sipeky C, Pursiheimo J-P, Tammela T, Kellokumpu-Lehtinen P-L, PRACTICAL Consortium, Fey V, Maehle L, Wiklund F, Wei G-H, Schleutker J (2018). ANO7 is associated with aggressive prostate cancer. *International Journal of Cancer* 143, 2479—2487.

Kaikkonen E, Takala A, Pursiheimo J-P, Wahlström G, Schleutker J (2020). The interactome of the prostate-specific protein Anoctamin 7. *Cancer Biomarkers* 28, 91—100.

Karni R, de Stanchina E, Lowe S W, Sinha R, Mu D, Krainer A R (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural & Molecular Biology* 14, 185—193.

Kiessling A, Weigle B, Fuessel S, Ebner R, Meye A, Rieger M A, Schmitz M, Temme A, Bachmann M, Wirth M P, Rieber E P (2005). *D-TMPP*: A Novel Androgen-Regulated Gene Preferentially Expressed in Prostate and Prostate Cancer That Is the First Characterized Member of an Eukaryotic Gene Family. *The Prostate* 64, 387—400.

Kunzelmann K, Ousingsawat J, Benedetto R, Cabrita I, Schreiber R (2019). Contribution of Anoctamins to Cell Survival and Cell Death. *Cancers* 11.

Käypä hoito -suositus (2021). Eturauhassyöpä. <https://www.kaypahoito.fi/hoi11060> [accessed 10.11.2021]

Madsen P P, Kibæk M, Roca X, Sachidanandam R, Krainer A R, Christensen E, Steiner R D, Gibson K M, Corydon T J, Knudsen I, Wanders R J A, Ruiter J P N, Gregersen N, Andresen B S (2006). Short/branched-chain acyl-CoA dehydrogenase deficiency due to an IVS3+3A>G mutation that causes exon skipping. *Journal of Human Genetics* 118, 680—690.

Majer F & Sikora J (2021). Easy and fast PCR-based protocol allows characterization of breakpoints resulting from *Alu/Alu*-mediated genomic rearrangements. *Molecular Genetics & Genomic Medicine* e1830.

Marx A, Koopmann L, Höflmayer D, Büscheck F, Hube-Magg C, Steurer S, Eichenauer T, Clauditz T S, Wilczak W, Simon R, Sauter G, Izbicki J R, Huland H, Heinzer H, Graefen M, Haese A, Schlomm T, Bernreuther C, Lebok P, Bonk S (2021). Reduced anoctamin 7 (ANO7) expression is a strong and independent predictor of poor prognosis in prostate cancer. *Cancer Biol Med* 18, 245—255.

McCrea E, Sissung T M, Price D K, Chau C H, Figg W D (2016). Androgen receptor variation affects prostate cancer progression and drug resistance. *Pharmacological Research* 114, 152—162.

Mohsenzadegan M, Madjd Z, Asgari M, Abolhasani M, Shekarabi M, Taeb J, Shariftabrizi (2013). Reduced expression of NGEP is associated with high-grade prostate cancers: a tissue microarray analysis. *Cancer Immunol Immunother* 62, 1609—1618.

Narla G, DiFeo A, Fernandez Y, Dhanasekaran S, Huang F, Sangodkar J, Hod E, Leake D, Friedman S L, Hall S J, Chinnaiyan A M, Gerald W L, Rubin M A, Martignetti J A (2008). KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis. *Journal of Clinical Investigation* 118, 2711—2721.

National Cancer Institute (2021). Genetics of Prostate Cancer (PDQ®)–Health Professional Version. <https://www.cancer.gov/types/prostate/hp/prostate-genetics-pdq> [accessed 08.05.2021]

Nyberg T, Frost D, Barrowdale D, Evans D G, Bancroft E, Adlard J, Ahmed M, Barwell J, Brady A F, Brewer C, Cook J, Davidson R, Donaldson A, Eason J, Gregory H, Henderson A, Izatt L, Kennedy M J, Miller C, Morrison P J, Murray A, Ong K-R, Porteous M, Pottinger C, Rogers M T, Side L, Snape K, Walker L, Tischkowitz M, Eeles R, Easton D F, Antoniou A C (2020). Prostate Cancer Risks for Male *BRCA1* and *BRCA2* Mutation Carriers: A Prospective Cohort Study. *European Urology* 77, 24—35.

Nyström-Lahti M, Kristo P, Nicolaides N C, Chang S-Y, Aaltonen L A, Moisio A-L, Järvinen H J, Mecklin J-P, Kinzler K W, Vogelstein B, de la Chapelle A, Peltomäki P (1995). Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nature Medicine* 1, 1203—1206.

Olender J & Lee N H (2019). Role of alternative splicing in prostate cancer aggressiveness and drug resistance in African Americans. *Adv Exp Med Biol* 1164, 119—139.

Pajares M J, Ezponda T, Catena R, Calvo A, Pio R, Montuenga L M (2007). Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol* 8, 349—357.

Paschalis A, Sharp A, Welti J C, Neeb A, Raj G V, Luo J, Plymate S R, de Bono J S (2018). Alternative splicing in prostate cancer. *Nature Reviews Clinical Oncology* 15, 663—675.

Pastor T, Talotti G, Lewandowska M A, Pagani F (2009). An *Alu*-derived intronic splicing enhancer facilitates intronic processing and modulates aberrant splicing in ATM. *Nucleic Acids Research* 37, 7258—7267.

Niu Y, Altuwaijri S, Lai K-P, Wu C-T, Ricke W A, Messing E M, Yao J, Yeh S, Chang C (2008). Androgen receptor is a tumor suppressor and proliferator in prostate cancer. *PNAS* 105, 12182—12187.

Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, Seppä K (2020). Cancer in Finland 2018. *Cancer Society of Finland Publication* 94.

Raghallaigh H N & Eeles R (2021). Genetic predisposition to prostate cancer: an update. *Familial Cancer.*

Rajan P, Elliott D J, Robson C N, Leung H Y (2009). Alternative splicing and biological heterogeneity in prostate cancer. *Nature Reviews Urology* 6, 454—460.

Ray P, Kar A, Fushimi K, Havlioglu N, Chen X, Wu J Y (2011). PSF Suppresses Tau Exon 10 Inclusion by Interacting with a Stem-Loop Structure Downstream of Exon 10. *Journal of Molecular Neuroscience* 45, 453—466.

Rhodes D R, Barrette T R, Rubin M A, Ghosh D, Chinnaiyan A M (2002). Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer. *Cancer Res* 62, 4427—4433.

Roca X, Sachidanancam R, Krainer A R (2005). Determinants of the inherent strength of human 5’ splice sites. *RNA* 11, 683—698.

Roca X, Olson A J, Rao A R, Enerly E, Kristensen V N, Børresen-Dale A-L, Andresen B S, Krainer A R, Sachidanandam R (2008). Features of 5-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome research* 18, 77—87.

Rodríquez C, Calle E E, Miracle-McMahill H L, Tatham L M, Wingo P A, Thun M J, Heath C W (1997). Family History and Risk of Fatal Prostate Cancer. *Epidemiology*, 8, 653—657.

Saarelma O, Duodecim (2021). Eturauhassyöpä. <https://www.terveyskirjasto.fi/dlk00210> [accessed 01.10.2021]

Schneider, C A, Rasband, W S, Eliceiri, K W (2012). "NIH Image to ImageJ: 25 years of image analysis". *Nature Methods* 9, 671—675.

Shen M M & Abate-Shen C (2010). Molecular genetics of prostate cancer: new prospects for old challenges. *Genes and Development* 24, 1967—2000.

Song X, Beck C R, Du R, Campbell I M, Coban-Akdemir Z, Gu S, Breman A M, Stankiewicz P, Ira G, Shaw C A, Lupski J R (2018). Predicting human genes susceptible to genomic instability associated with *Alu/Alu*-mediated rearrangements. *Genome Research* 28, 1228—1242.

Spurdle A B, Couch F J, Hogervorst F B L, Radice P, Sinilkova O M (2008). Prediction and Assessment of Splicing Alterations: Implications for Clinical Testing. *Wiley Human Mutation* 29, 1304—1313.

Stuart R O, Wachsman W, Berry C C, Wang-Rodriguez J, Wasserman L, Klacansky I, Masys D, Arden K, Goodison S, McClelland M, Wang Y, Sawyers A, Kalcheva I, Tarin D, Mercola D (2004). *In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer. *PNAS* 101, 615—620.

Sugnet C W, Kent W J, Ares M Jr., Haussler D (2004). Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice. *Pacific Symposium on Biocomputing* 9, 66—77.

Suzuki J, Fujii T, Imao T, Ishihara K, Kuba H, Nagata S (2013). Calcium-dependent Phospholipid Scramblase Activity of TMEM16 Protein Family Members. *The Journal of Biological Chemistry* 288, 13305—13316.

The Human Gene Mutation Database (2021). <http://www.hgmd.cf.ac.uk/ac/index.php> [accessed 31.03.2021]

Tsai Y S, Dominguez D, Gomez S M, Wang Z (2015). Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget* 6, 6825—6839.

Tzetis M, Efthymiadou A, Doudounakis S, Kanavakis E (2001). Qualitative and quantitative analysis of mRNA associated with four putative splicing mutations (621+3A→G, 2751+2T→A, 296+1G→C, 1717–9T→C-D565G) and one nonsense mutation (E822X) in the *CFTR* gene. *Journal of Human Genetics* 109, 592—601.

Wang E & Aifantis I (2020). RNA Splicing and Cancer. *Trends in Cancer* 6, 631—644.

Wang G, Zhao D, Spring D J, DePinho R A (2018). Genetics and biology of prostate cancer. *Genes and Development* 32, 1105—1140.

Wahl M C, Will C L, Lűhrmann R (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* 136, 701—718.

Zeeger M P A, Jellema A, Ostrer H (2003). Empiric Risk of Prostate Carcinoma for Relatives of Patients with Prostate Carcinoma. *American Cancer Society* 97, 1894—1903.

# Appendices

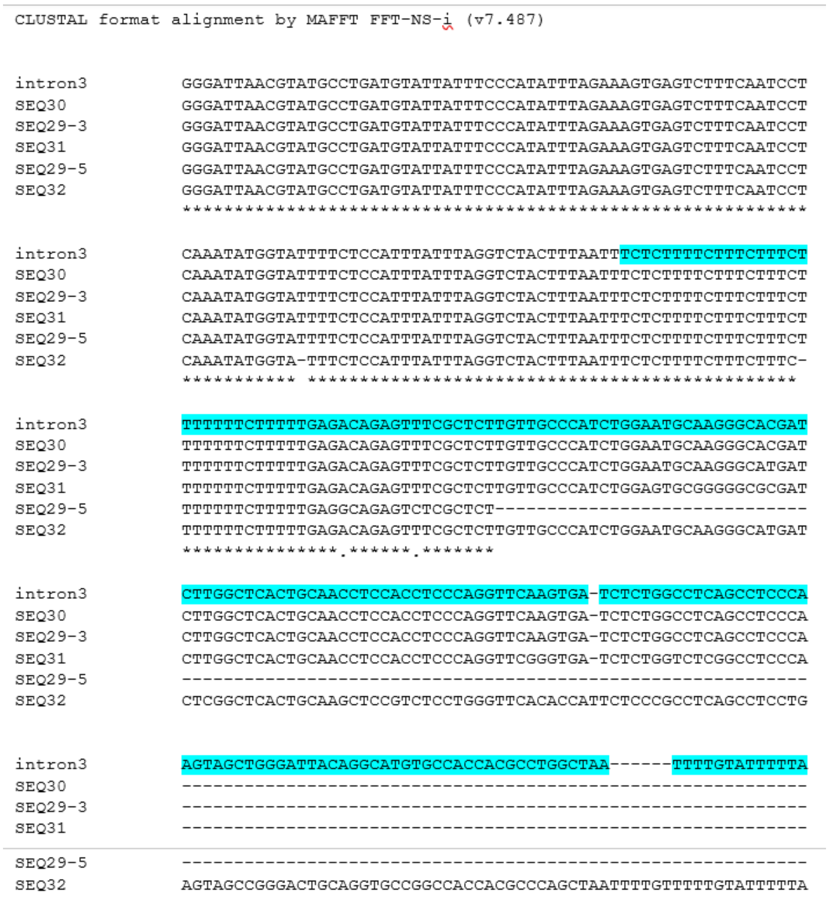
## Appendix 1 – Primer table

| Primer | Sequence |
|---|---|
| ANO7_i3_F1 | 5’-GTTGTCCTTACTCCTCGGGTG-3’ |
| ANO7_i3_R1 | 5’-TGGCATGAGACAGCAATGGA-3’ |
| ANO7_i3_F2 | 5’-GGGATTAACGTATGCCTGATGT-3’ |
| ANO7_i3_R2 | 5’-ACACCCGAGGAGTAAGGACA-3’ |
| ACTB-221-F | 5’-GCCTCGCCTTTGCCGA-3' |
| ACTB-221-R | 5’-CTCGTCGCCCACATAGGAAT-3' |
| T3 | 5’-ATTAACCCTCACTAAAGGGA-3’ |
| T7 | 5’-TAATACGACTCACTATAGGG-3’ |
| ANO7-ex4-F | 5’-GGGGAATTCTTCTCTATGGCAGGACGGGA-3’ |
| ANO7-ex4-R | 5’-GGGGGATCCGGAGACGGGACAGGTGAATG-3’ |
| ANO7-in6-3-F | 5’-TCCTGGGGATGTGAAAAGGC-3’ |
| ANO7-in8-1-R | 5’-ACCCAGAGAGACAGAGCCAT-3’ |
| dUSD2 | 5’-CCTGCACCTGAGGAGTGAAT-3’ |
| dUSA4 | 5’-GCTCACAAATACCACTGAGAT-3’ |
| ANO7-crex7-3-F | 5’-CAGAAAGGCCTGCTTGAGAC-3’ |
| dUSA4-C | 5’-CCTGCACCTGAGGAGTGAAT-3’ |

## Appendix 2 – MAFFT analysis of all intron 3 sequences against intron 3 reference sequence

GAP OPEN PENALTY 3.0
GAP EXTENSION PENALTY 0.01

🟦 = AluSx
🟩 = AluY

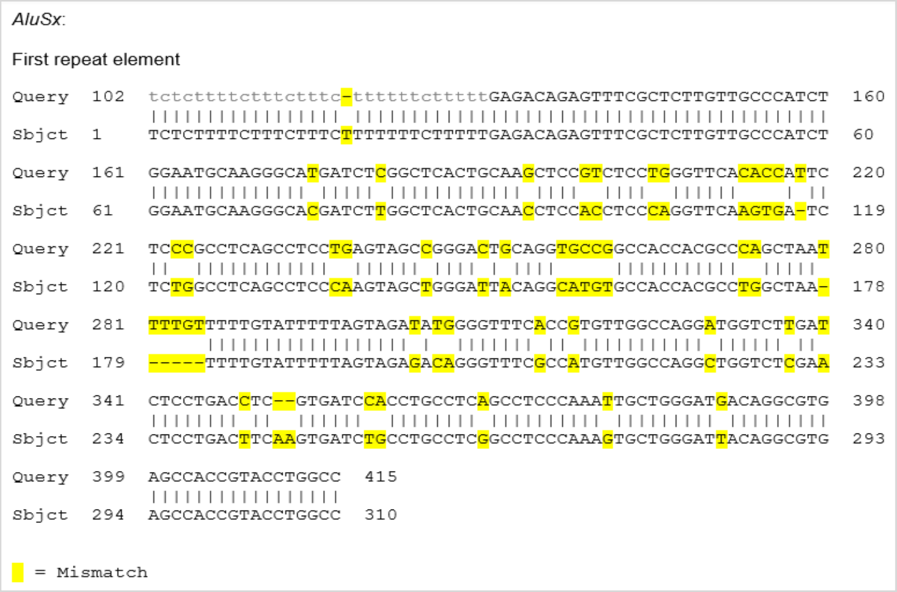


```
CLUSTAL format alignment by MAFFT FFT-NS-i (v7.487)


intron3         GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
SEQ30           GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
SEQ29-3         GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
SEQ31           GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
SEQ29-5         GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
SEQ32           GGGATTAACGTATGCCTGATGTATTATTTCCCATATTTAGAAAGTGAGTCTTTCAATCCT
                ************************************************************

intron3         CAAATATGGTATTTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTCT
SEQ30           CAAATATGGTATTTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTCT
SEQ29-3         CAAATATGGTATTTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTCT
SEQ31           CAAATATGGTATTTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTCT
SEQ29-5         CAAATATGGTATTTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTCT
SEQ32           CAAATATGGTA-TTTCTCCATTTATTTAGGTCTACTTTAATTTCTCTTTTCTTTCTTTC-
                *********** *********************************************** 

intron3         TTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCTGGAATGCAAGGGCACGAT
SEQ30           TTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCTGGAATGCAAGGGCACGAT
SEQ29-3         TTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCTGGAATGCAAGGGCATGAT
SEQ31           TTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCTGGAGTGCGGGGGCGCGAT
SEQ29-5         TTTTTTCTTTTTGAGGCAGAGTCTCGCTCT------------------------------
SEQ32           TTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCTGGAATGCAAGGGCATGAT
                ***************.******.*******                              

intron3         CTTGGCTCACTGCAACCTCCACCTCCCAGGTTCAAGTGA-TCTCTGGCCTCAGCCTCCCA
SEQ30           CTTGGCTCACTGCAACCTCCACCTCCCAGGTTCAAGTGA-TCTCTGGCCTCAGCCTCCCA
SEQ29-3         CTTGGCTCACTGCAACCTCCACCTCCCAGGTTCAAGTGA-TCTCTGGCCTCAGCCTCCCA
SEQ31           CTTGGCTCACTGCAACCTCCACCTCCCAGGTTCGGGTGA-TCTCTGGTCTCGGCCTCCCA
SEQ29-5         ------------------------------------------------------------
SEQ32           CTCGGCTCACTGCAAGCTCCGTCTCCTGGGTTCACACCATTCTCCCGCCTCAGCCTCCTG
                                                                            

intron3         AGTAGCTGGGATTACAGGCATGTGCCACCACGCCTGGCTAA------TTTTGTATTTTTA
SEQ30           ------------------------------------------------------------
SEQ29-3         ------------------------------------------------------------
SEQ31           ------------------------------------------------------------
SEQ29-5         ------------------------------------------------------------
SEQ32           AGTAGCCGGGACTGCAGGTGCCGGCCACCACGCCCAGCTAATTTTGTTTTTGTATTTTTA
```

```
intron3          GTAGAGACAGGGTTTCGCCATGTTGGCCAGGCTGGTCTCGAACTCCTGACTTCAAGTGAT
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          ------------------------------------------------------------
SEQ32            GTAGATATGGGGTTTCACCGTGTTGGCCAGGATGGTCTTGATCTCCTGACCTC--GTGAT


intron3          CTGCCTGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGTACCTGGCCT
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          ------------------------------------------------------------
SEQ32            CCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGGCGTGAGCCACCGTACCTGGCCT


intron3          TTAATTTTTTTTTTTTTTTTTTTTGCTGCTTTCGTGAGATCTTGCACATCTTTTAGTAGA
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          ------------------------------------------------------------
SEQ32            TTAA-----TTTTTTTTTTTTTTTGCTGCTTTCGTGAGATCTTGCACATC-TTTAGTAGA


intron3          TGTATTCCAAATAGCATTACTTGGAAAATTTCACTGCCTATGTGTCATAGGTATATAGAA
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          ------------------------------------------------------------
SEQ32            TGTATTCCAAATAGCATTACTTGGAAAATTTCACTGCCTATGTGTCATAGGTATATAGAA


intron3          ATTATTTATTTCCTTTTTTTTTTTTTTTTGAGATGGAGTCTCGCTCTGTTGCCCTGTTGC
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          -----------------------------------------------GTTGCCCTGTTGC
SEQ32            ATTATTTATTTCC---TTTTTTTTTTTTTGAGATGGAGTCTCGCTCTGTTGCCCTGTTGC


intron3          CCAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCCGTCTCCTGGGTTCAC
SEQ30            ------------------------------------------------------------
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          CCAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCTGTCTCCTGGGTTCAC
SEQ32            CCAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCCGTCTCCTGGGTTCAC


intron3          ACCATTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCC
SEQ30            -------------------------AGTAGCTGGGACTACAGGTGCCGGCCACCACGCCC
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          ACCATTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCC
SEQ32            ACCATTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCC


intron3          AGCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTAGCCAGGATGG
SEQ30            AGCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTAGCCAGGATGG
SEQ29-3          ------------------------------------------------------------
SEQ31            ------------------------------------------------------------
SEQ29-5          AGCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTAGCCAGGATGG
SEQ32            AGCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTGGCCAGGATGG


intron3          TCTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAG
SEQ30            TCTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAG
SEQ29-3          -------------------------------------------AATTGCTGGGATGACAG
SEQ31            -------------------------------------------AATTGCTGGGATGACAG
SEQ29-5          TCTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAG
SEQ32            TCTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAG
                                                            *****************

intron3          GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGCGAGACAGAATGCCCA
SEQ30            GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGGGAGACAGAATGCCCA
SEQ29-3          GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGGGAGACAGAATGCCCA
SEQ31            GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGGGAGACAGAATGCCCA
SEQ29-5          GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGGGAGACAGAATGCCCA
SEQ32            GCGTGAGCCACCAAACCCAGCCACAATAAGATATTGTAAAAGAGGGAGACAGAATGCCCA
                 ******************************************** ***************
```

```
intron3         CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
SEQ30           CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
SEQ29-3         CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
SEQ31           CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
SEQ29-5         CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
SEQ32           CATTCACATAACTTTTACTGCAGCATTTTGTTATAATTATTCTGTTTTATTATTAGTTAT
                ************************************************************

intron3         TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
SEQ30           TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
SEQ29-3         TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
SEQ31           TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
SEQ29-5         TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
SEQ32           TGTTCATCTCTTACTTTACCTAATTCATAAATTAACCTTCACCATAGGAATGTATATCAT
                ************************************************************

intron3         AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTTACTCT
SEQ30           AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTTACTCT
SEQ29-3         AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTTACTCT
SEQ31           AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTTACTCT
SEQ29-5         AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTCACTCT
SEQ32           AGGTATGTACGTATTATAGGTATGTATAGGGAAAGACAGTATATATAGGTTTGTTACTCT
                ******************************************************.*****

intron3         GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCTGCAGTTAAGGGGACTCCG
SEQ30           GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCCGCAGTTAAGGGGACTCCG
SEQ29-3         GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCCGCAGTTAAGGGGACTCCG
SEQ31           GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCCGCAGTTAAGGGGACTCCG
SEQ29-5         GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCCGCAGTTAAGGGGACTCCG
SEQ32           GCACAGTTTCAGGCATCCACCGGGTTCTTGGATAGATCCCCCGCAGTTAAGGGGACTCCG
                *****************************************.******************

intron3         TTGTCCTTACTCCTCGGGTGT
SEQ30           TTGTCCTTACTCCTCGGGTGT
SEQ29-3         TTGTCCTTACTCCTCGGGTGT
SEQ31           TTGTCCTTACTCCTCGGGTGT
SEQ29-5         TTGTCCTTACTCCTCGGGTGT
SEQ32           TTGTCCTTACTCCTCGGGTGT
                *********************
```

# Appendix 3 – BLAST searches of SEQ32’s first and second repeat elements against *AluSx* and *AluY* repeat sequences

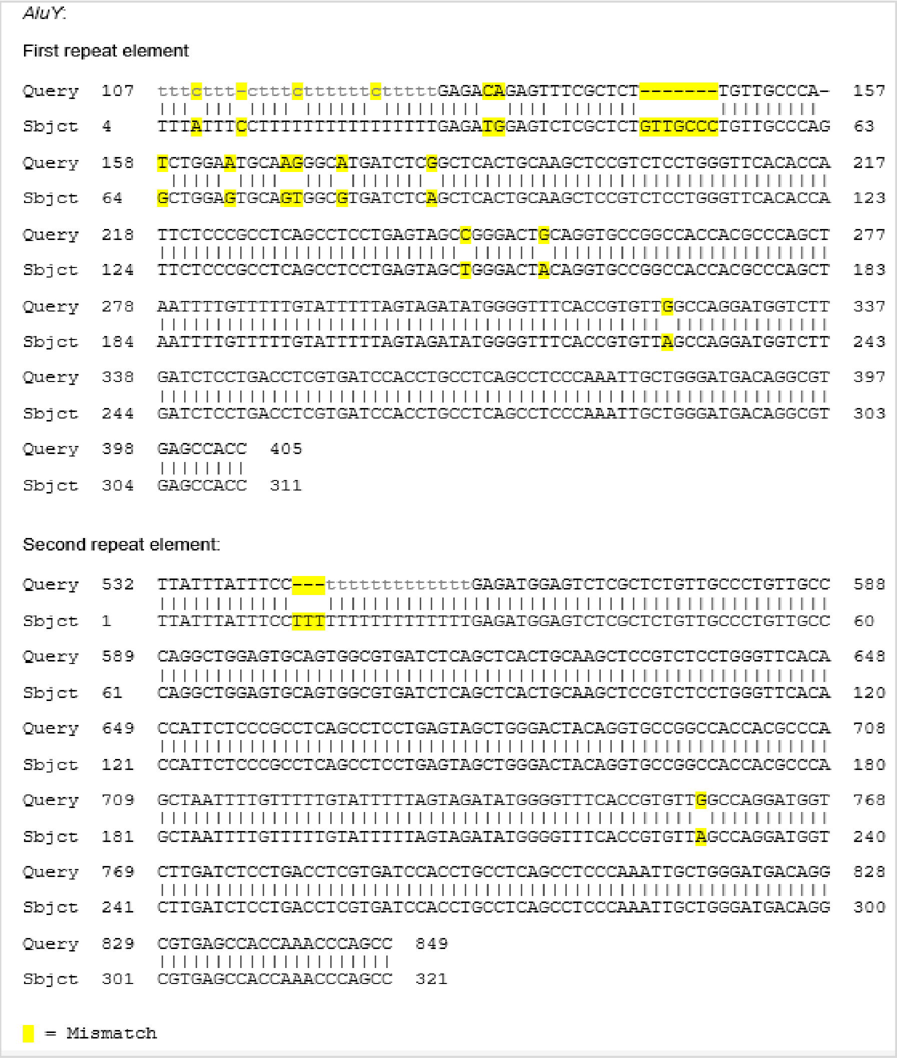


*AluSx*:

First repeat element

```
Query  102  tctcttttctttctttc-ttttttctttttGAGACAGAGTTTCGCTCTTGTTGCCCATCT  160
            ||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TCTCTTTTCTTTCTTTCTTTTTTTCTTTTTGAGACAGAGTTTCGCTCTTGTTGCCCATCT  60

Query  161  GGAATGCAAGGGCATGATCTCGGCTCACTGCAAGCTCCGTCTCCTGGGTTCACACCATTC  220
            |||||||||||||| ||||| |||||||||||| ||||  ||||  ||||||    | ||
Sbjct  61   GGAATGCAAGGGCACGATCTTGGCTCACTGCAACCTCCACCTCCCAGGTTCAAGTGA-TC  119

Query  221  TCCCGCCTCAGCCTCCTGAGTAGCCGGGACTGCAGGTGCCGGCCACCACGCCCAGCTAAT  280
            ||  ||||||||||||  |||||| |||| | ||||     |||||||||||  |||||
Sbjct  120  TCTGGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGCATGTGCCACCACGCCTGGCTAA-  178

Query  281  TTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTGGCCAGGATGGTCTTGAT  340
                 |||||||||||||||||| |  ||||||| || ||||||||||| |||||| ||
Sbjct  179  -----TTTTGTATTTTTAGTAGAGACAGGGTTTCGCCATGTTGGCCAGGCTGGTCTCGAA  233

Query  341  CTCCTGACCTC--GTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGGCGTG  398
            |||||||| ||  ||||||  |||||||| |||||||||| ||||||||| |||||||||
Sbjct  234  CTCCTGACTTCAAGTGATCTGCCTGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTG  293

Query  399  AGCCACCGTACCTGGCC  415
            |||||||||||||||||
Sbjct  294  AGCCACCGTACCTGGCC  310
```

🟨 = Mismatch



*AluY*:

First repeat element

```
Query  107  tttcttt-ctttcttttttcttttttGAGACAGAGTTTCGCTCT-------TGTTGCCCA-  157
            ||| ||| |||| |||||| |||||||||  |||| |||||||       |||||||||
Sbjct  4    TTTATTTCCTTTTTTTTTTTTTTTTTGAGATGGAGTCTCGCTCTGTTGCCCTGTTGCCCAG  63

Query  158  TCTGGAATGCAAGGGCATGATCTCGGCTCACTGCAAGCTCCGTCTCCTGGGTTCACACCA  217
             ||||| ||||  ||| ||||||| |||||||||||||||||||||||||||||||||||
Sbjct  64   GCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCCGTCTCCTGGGTTCACACCA  123

Query  218  TTCTCCCGCCTCAGCCTCCTGAGTAGCCGGGACTGCAGGTGCCGGCCACCACGCCCAGCT  277
            ||||||||||||||||||||||||||| |||||| |||||||||||||||||||||||||
Sbjct  124  TTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCCAGCT  183

Query  278  AATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTGGCCAGGATGGTCTT  337
            ||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  184  AATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTAGCCAGGATGGTCTT  243

Query  338  GATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGGCGT  397
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  244  GATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGGCGT  303

Query  398  GAGCCACC  405
            ||||||||
Sbjct  304  GAGCCACC  311
```

Second repeat element:

```
Query  532  TTATTTATTTCC---tttttttttttttGAGATGGAGTCTCGCTCTGTTGCCCTGTTGCC  588
            ||||||||||||   |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TTATTTATTTCCTTTTTTTTTTTTTTTTGAGATGGAGTCTCGCTCTGTTGCCCTGTTGCC  60

Query  589  CAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCCGTCTCCTGGGTTCACA  648
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCAAGCTCCGTCTCCTGGGTTCACA  120

Query  649  CCATTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCCA  708
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CCATTCTCCCGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGTGCCGGCCACCACGCCCA  180

Query  709  GCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTGGCCAGGATGGT  768
            |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||
Sbjct  181  GCTAATTTTGTTTTTGTATTTTTAGTAGATATGGGGTTTCACCGTGTTAGCCAGGATGGT  240

Query  769  CTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGG  828
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CTTGATCTCCTGACCTCGTGATCCACCTGCCTCAGCCTCCCAAATTGCTGGGATGACAGG  300

Query  829  CGTGAGCCACCAAACCCAGCC  849
            |||||||||||||||||||||
Sbjct  301  CGTGAGCCACCAAACCCAGCC  321
```

🟨 = Mismatch

## Appendix 4 – Statistical analyses by SAS Enterprise

| Assay | Response variable | Construct | Mean (fraction) | 95% CI, lower | 95% CI, upper | F value | P value |
|---|---|---|---|---|---|---|---|
| Exon 4 | | | | | | | |
| | Exon inclusion | | | | | F(2,13) = 2491.66 | <.0001 |
| | | Ex4-2 | 0.8291 | 0.7938 | 0.8593 | | |
| | | Ex4-4 | 0.01984 | 0.01467 | 0.02680 | | |
| | | Ex4-5 | 0.1556 | 0.1275 | 0.1275 | | |
| Intron 7 | | | | | | | |
| | Exon 7 and 8 inclusion | | | | | F(1,8) = 158.45 | <.0001 |
| | | In7-1 | 0.8025 | 0.7298 | 0.8594 | | |
| | | In7-2 | 0.9096 | 0.8688 | 0.9387 | | |
| | Exon 8 skipping | | | | | F(1,8) = 4.65 | 0.0631 |
| | | In7-1 | 0.08024 | 0.05041 | 0.1254 | | |
| | | In7-2 | 0.07138 | 0.04464 | 0.1123 | | |
| | Total exon skipping | | | | | F(1,8) = 265.71 | <.0001 |
| | | In7-1 | 0.1166 | 0.09285 | 0.1455 | | |
| | | In7-2 | 0.01831 | 0.01304 | 0.02564 | | |
| | | | | | | | |
| | | | | CI = Confidence interval | | | |