UNIVERSITY OF TRENTO - Italy

International PhD Program in Biomolecular Sciences

Department of Cellular, Computational and Integrative Biology – CIBIO
31 Cycle

"The effect of germline variants on the genesis of early somatic events in cancer explored via Cas9 genome editing"

Tutor/Advisor

Prof. Francesca Demichelis

Supervisor

Dr. Paola Gasperini

Ph.D. Thesis of

Blerta Stringa

CIBIO, University of Trento

Academic Year 2018-2019

I would like to thank in particular Prof. Francesca Demichelis.
Her support has been precious to me.
Thank you!
Faleminderit!

Declaration of authorship

I, Blerta Stringa, declare that this thesis titled “The effect of germline variants on the genesis of early somatic events in cancer explored via Cas9 genome editing” is my own work and the use of all material from other sources has been properly and fully acknowledged. I confirm that:

o This work was done while in candidature for a research degree at this University;

o Where I have consulted the published work of others, this is always clearly attributed;

o Where I have quoted from the work of others, the source is always given;

o I have properly and fully acknowledged all main sources of help;

o Where the work was done by myself jointly with others, I have made clear what exactly was done by others and what I have contributed myself.

Abstract

Although the understanding of genetic predisposition to prostate cancer (PCa) has been improved through genome-wide association studies (GWAS), little is known about the biological implication of germline variants residing in coding or non-coding regions in cancer development and progression. Our hypothesis is that inherited variants may predispose to specific early recurrent genomic events observed in PCa adenocarcinomas, possibly in the context of variable androgen receptor (AR) signaling that changes during a man's lifetime.

Recent *in silico* analysis by our group on potential association between germline variants and PCa specific somatic lesions identified a non-coding polymorphic regulatory element at the 7p14.3 locus associated with DNA repair and hormone regulated transcript levels and with an early recurrent prostate cancer specific somatic mutation in the Speckle-Type POZ protein (SPOP) gene (OR=5.54, P=1.22e-08) in human prostate tissue data. In order to functionally characterize the polymorphic 7p14.3 locus (rs1376350, single nucleotide polymorphism, G>A), we set up to establish isogenic cell lines harboring the minor allele by using the CRISPR/Cas9 system. In parallel, CRISPR/Cas9 system was used to knock out different portion of the region encompassing the 7p14.3 variant and to eliminate transcription factors (TFs) binding sites that were identified from previous *in silico* analysis (i.e. AR and CCAAT/Enhancer Binding Protein (C/EBP) beta (CEBPβ)). The transcriptomes of edited pools and edited single clones from macrodeletion (731 bp), microdeletion (50 bp) and alterations of TFs binding sites were analyzed and compared to the transcriptomes of isogenic cells heterozygous (A/G) and homozygous (A/A) for the minor allele A of the risk variant rs1376350 (with or without AR overexpression).

These data identified a set of genes scattered throughout the genome with the same pattern of deregulation suggesting the implication of the variant on the regulation of genes residing in different chromosomes. Additionally, ChIP-qPCR experiments for histone modification supported the identification of the 7p14.3 locus with enhancer activity. Furthermore, ChIP-qPCR of histone mark associated with transcriptional activation or repression in isogenic cells harboring the minor allele A upon AR overexpression showed that the activity of the locus is higher for the minor allele A compared to G, independently from AR activation.

Despite the limitations of our model and the current lack of validation in other cells, we confirmed that some of the differentially expressed genes that emerged from the comparative analysis of edited cells are deregulated in human normal and tumor prostate samples as well. This work is a proof of concept of germline predisposition to molecularly distinct cancer subclasses and has the potential to nominate new mechanisms of cancer development.

Future work aims to elucidate the mechanisms implicated in the deregulation of the transcriptome by combining the information obtained until now with potential new players that we expect to identify by Mass Spectrometry experiments. To clarify the link between the 7p14.3 variant and the somatic mutations in SPOP, we plan to express mutant SPOP in isogenic cells harboring the minor allele and to asses DNA damage response upon overexpression or silencing of TFs binding at and around the rs1376350 variant.

My work is an example of how the CRISPR/Cas9 system can be used to develop a technical framework with convergent approaches to functionally characterize polymorphic regulatory regions including but not limited to the establishment of isogenic cells upon single nucleotide editing.

Index

1 Introduction

1.1 Prostate

1.1.1 Prostate Cancer Etiopathogenesis

The prostate gland is a walnut-sized male reproductive organ located directly beneath the urinary bladder and the urethra. The function of the organ is to produce seminal fluid important for the nutrition and transport of sperm. The prostate anatomy includes three lobes: transient, peripheral and central zone, each made by muscles and glands (McNeal, 1988). The development of the prostate, that starts around the second and third trimester and completes at the time of birth, is under the control of androgen hormones produced by the fetal testes. Thirty days after birth the morphogenesis of the organ is completed, while the final growth and maturation occurs at puberty when androgen circulating levels rise markedly. Around age 50 the size of the prostate and the secreted seminal fluid decrease due to the reduction of androgen levels. Prostate activity is under the control of hormones such as those released by the hypophysis. Luteinizing hormone (LH) and follicle stimulating hormone (FSH) released by the pituitary gland control the number of Leydig cells and testosterone secretion (**Figure 1A**). 17-β hydroxysteroid dehydrogenase (HSD17β) enzyme synthetizes testosterone and its inhibition negatively impacts the normal development and function of the prostate gland (White, Xie and Ventura, 2013). Once synthetized, testosterone is secreted into the blood and carried to target cells in the male reproductive organs. Sex hormone binding globulin (SHBG) transports most of testosterone while the remaining free molecules can be converted to a more potent androgen, dihydrotestosterone (DHT) by 5alpha reductase that is expressed in prostate epithelial cells. On the other hand, adrenal glands, under a complex enzymatic activity, produce large amounts of Dehydroepiandrosterone (DHEA) and DHEA-sulfate (DHEA-S), which are transported in the blood to the prostate and to other peripheral tissues (**Figure 1B**). These inactive precursors are then transformed locally into the active androgens, testosterone and DHT. The enzymatic complexes DHEA sulfatase, 3α-HSD, 17α-HSD and 5α-reductase are all present in the prostatic cells. Once in the cytoplasm, unconverted testosterone and DHT recognize the binding domain of AR and activate a cascade of physiological events (**Figure 1C**) (Watson, Arora and Sawyers, 2015). Accumulation of DHT in the prostate increases the amount of active AR translocated into nucleus, which then triggers organ growth in both stroma and epithelia. This pathological condition known as Benign Prostatic Hyperplasia (BPH) (Gerald *et al.*, 2004) is detectable in one fourth of men by the fifth decade of life and in over 90% of men by the eighth decade. Development and progression of BPH is strongly associated with acute and chronic inflammation in adult prostate (Bushman, 2009). Although they anatomically arise from different districts of the prostate gland, BPH and prostate cancer (PCa) can co-exist in the same prostatic zone (White, Xie and Ventura, 2013). Genetic, hormonal, and inflammatory mechanisms have all been shown to be common pathophysiological driving mechanisms for the development of both BPH and PCa, potentially linking these conditions. However, on a cellular and molecular level, no study has shown that the development of BPH tissue has later converted into an oncological disease (Miah and Catto, 2019).

PCa is the most common cancer in men worldwide and a major cause of cancer death (Abeshouse *et al.*, 2015) with a higher incidence in African Americans compared to Americans of Asian ancestry (Pernar *et al.*, 2018). In 2018, PCa was ranked third cause of cancer death in men after lung and colorectal cancer (Ferlay *et al.*, 2018). Historically, diagnosis, therapeutic decisions and prognosis of prostate cancer patients have been defined by prostate specific antigen (PSA) levels and by the Gleason score, the prostate cancer grading system. In the past few years, the community of urologists is debating about the widespread use of PSA testing for PCa diagnosis, which leads to an increased incidence of low risk tumors and potentially unnecessary biopsies (Andriole *et al.*, 2012; Schröder *et al.*, 2012). The study of genetic markers, *in situ* molecular analyses, DNA and RNA-sequencing capabilities are pursued to not only improve cancer diagnosis but also to avoid unnecessarily treatment of people with the less aggressive forms of the disease (Siegel, Miller and Jemal, 2017).

Although environmental factors such as smoking, alcohol consumption and fat diet are considered important risk factors, the genetic component remains the principal protagonist of PCa etiology. In fact, the risk of developing the disease for first-degree relatives of men with PCa is two-fold the risk for the general population. The risk increases four times more if the diagnosis of cancer for first-degree relatives occurs in men younger than 60 years (Bell *et al.*, 2015). PCa genetic

predisposition is evident by twin studies that report 50% higher risk in monozygotic than in dizygotic twins (Goldgar *et al.*, 1994; Farashi *et al.*, 2019). Family history in conjunction with ancestral differences in PCa risk highlights the contribution of genetics to its etiology.

Recent studies highlighted that PCa presents a higher degree of heterogeneity comparing to other types of tumors. This characteristic is not limited to inter-tumor heterogeneity, but expands to intra-tumor heterogeneity (Rubin and Demichelis, 2018). In line with the clinical heterogeneity of the disease, genetic, genomic and molecular studies are proposed by large consortia both in the setting of localized disease and of metastatic disease to better identify cancer subtypes to eventually guide selection of tailored targeted therapeutic interventions.

1.1.2 Molecular Taxonomy of localized PCa

Cancer is a complex genetic disease influenced by both inherited variants and somatic alterations. Somatic mutations occur in the genome of both normal and neoplastic cells through exposure to exogenous or endogenous mutagens or as a result of altered DNA replication. Somatic genomic mutations detected in cancer include 'driver' mutations, implicated in cancer development and progression, and 'passenger' mutation, which do not confer growth advantage but are present in the progenitor cell of the tumor clone. By investigating recurrent somatic alterations within and across cancer types, large consortia as The Cancer Genome Atlas (TCGA) and as the International Cancer Genome Consortium (ICGC), and independent collaborative groups detected widespread heterogeneity within and between tumors.

The first large scale NGS based PCa studies revealed alterations of genes involved in prostatic development, chromatin modification, androgen signaling, and cell-cycle regulation (Berger *et al.*, 2011; Baca and Garraway, 2012; Barbieri *et al.*, 2012; Baca *et al.*, 2013). In 2015, TCGA published on the molecular taxonomy of primary PCa (localized adenocarcinomas) based on multi/layer data from a large collection of more than 300 PCa patients reporting on somatic single nucleotide mutations, gene fusions, somatic copy number alteration (SCNA), gene expression and DNA methylation (Abeshouse *et al.*, 2015). To date the taxonomy of primary PCa is becoming even more detailed both for indolent and non-indolent disease (Fraser *et al.*, 2017; Wedge *et al.*, 2018) with intriguing information on the involvement of DNA damage repair pathway.

Overall, the molecular heterogeneity of prostate tumors poses challenges to genomic biomarker identification, whereas the knowledge on genetic alterations may help the prediction of the clinical course of disease and the individual's response to therapy (Rubin and Demichelis, 2018; Salami *et al.*, 2018).

Localized PCa is macroscopically defined by major subclasses. The largest one including ~50% of PCa patients is characterized by fusions involving members of the ETS family (E26 transformation-specific family), especially ERG, ETV1, ETV4 and FLI1. ETS fusions are mutually exclusive with *SPOP* and *FOXA1* point mutations, two subtypes with the highest AR transcriptional activity, and with IDH1 positive tumors. SCNA analysis from the TCGA landmark paper further described that homozygous deletions spanning the *PTEN* locus occur in 15% of PCas, while less frequent focal deletions span *TP53*, *CDKN1B*, *MAP3K1*, *FNACD2* and *SPOPL*. Moreover, DNA repair pathway alterations are reflected in 19% of patient with PCa, mainly involving lesions in BRCA2, ATM, CDK12 and FANCD2 (Abeshouse *et al.*, 2015). In contrast to the common well-defined structural genomic lesions in PCa, point mutations (missense, altering single amino acids in the protein and nonsense, resulting in truncations) and indels (small insertions or deletions potentially resulting in frameshifts deleterious to the gene product) are less common than in other tumor types (Barbieri *et al.*, 2013). The TCGA study confirmed SPOP as the most frequently point mutated gene in PCa (11%), in line with the original report (Barbieri *et al.*, 2012).

1.1.3 SPOP mutated subclass in PCa

As a human based data study from our laboratory identified the *SPOP* mutated subclass as associated with inherited variants, we here summarize the main observations emerged in the last years regarding the SPOP mutation phenotype. *SPOP* encodes the substrate-recognition component of an E3-ubiquitin ligase (Zhang *et al.*, 2006), and all *SPOP* mutations observed in PCa occur exclusively in the structurally defined substrate binding subunit (MATH domain). Importantly, SPOP is responsible for the ubiquitination of several proteins, including the key AR coactivator Steroid Receptor Coactivator (SRC)-3 (Geng *et al.*, 2013), in line with the observed enhanced AR signaling in *SPOP* mutant PCa (Abeshouse *et al.*, 2015). Silencing of SPOP wt or overexpression of loss of function SPOP mutant (F133V MATH variant domain) has been associated with enhanced invasiveness of PCa cells. Moreover, recent work showed that SPOP mutation subclass is associated with upregulation of three metabolic pathways; tricarboxylic acid (TCA) cycle, fatty acid metabolism and glycerophospholipid metabolism (Yan *et al.*, 2017). While mutually exclusive with ETS fusions, SPOP mutant PCas frequently harbor deletion of the chromodomain-helicase-DNA-binding protein (CHD1), as confirmed across multiple cohorts and, unexpectedly, lack TP53 lesions (Barbieri *et al.*, 2012; Blattner *et al.*, 2014). Boysen et al. (Boysen *et al.*, 2018) found that metastatic castrate resistant prostate cancer (mCRPC) patients (resistance mechanisms reviewed in (Lorenzin and Demichelis, 2019)) with both *SPOP* mutation and *CHD1* deletion respond better to abiraterone treatment. WGS has revealed that genomic rearrangements in patients with *SPOP* mutation are also distinctive. Indeed, the SPOP mutation subclass has a significantly increased number of total genomic rearrangements compared with other classes (Baca *et al.*, 2013; Boysen *et al.*, 2015). In addition, SPOP mutation cancers harbor characteristic intrachromosomal rearrangements associated with significant copy number alterations, rather than balanced interchromosomal rearrangements observed in non-SPOP mutation PCa patient. Reporter assays for Homologous Recombination (HR) and Error-prone Non-homologous end Joining (NHEJ) showed that SPOP wt increased HR competence while SPOP-F133V and SPOP knockdown increased NHEJ (Boysen *et al.*, 2015) and further showed that *SPOP* mutations increase sensitivity to PARP inhibitor treatment. *In vivo* studies in human SPOP mutation mouse model demonstrated a deregulation of both PI3K/mTOR and AR pathways (Blattner *et al.*, 2017). Taken together, these results implicate SPOP as a novel participant in double-strand DNA brake (DSB) repair, suggest that *SPOP* mutation drives prostate tumorigenesis in part through genomic instability, and indicate that mutant SPOP tumors may respond to DNA-damaging therapeutics differently than other PCas.

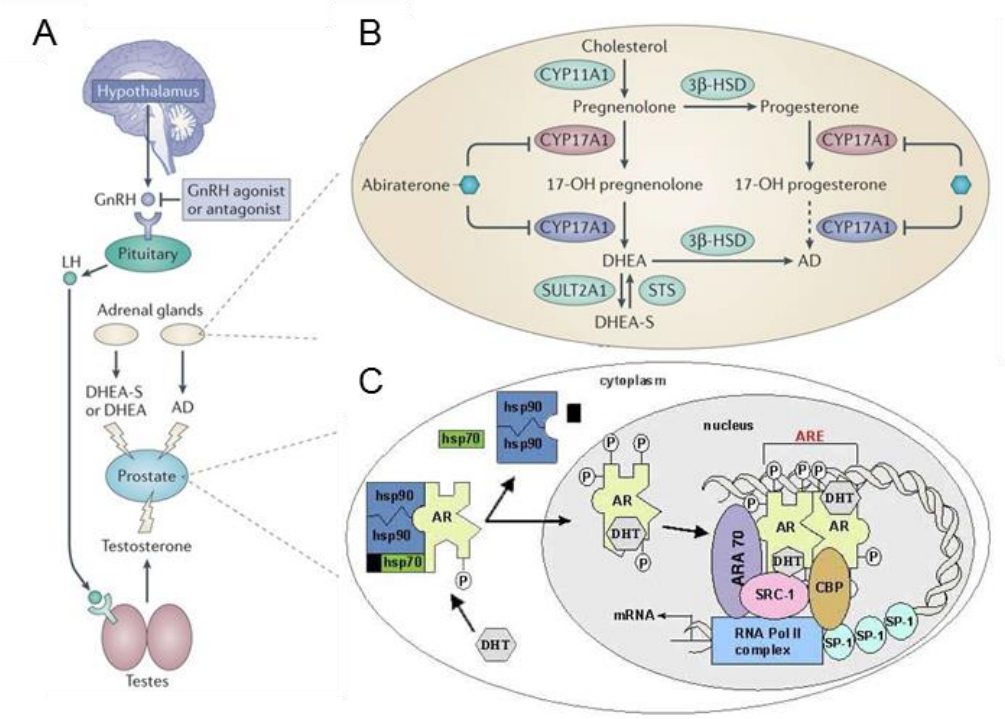


Figure 1. Hypothalamic–pituitary–adrenal axis, Adrenal androgen de novo steroidogenesis and Androgen Receptor activation in the prostate (modified image from (Sadar, 2003) and (Watson, Arora and Sawyers, 2015)). A) GnRH, released by hypothalamus, travels in hypothalamic-hypophysis portal blood system. Once arrived in anterior lobe of pituitary gland stimulate the release of FSH, LH and ACTH. LH and FSH hormones activity is located in testis cells which control production of testosterone and estrogens. Under ACTH control, the adrenal glands secrete DHEA-S, predominantly, DHEA and AD into the circulation. B) In the adrenal glands precursors of testosterone are produced under two pathways lead by CYP17A1. The enzyme has 17α

hydroxylation (red) and 17, 20-lyase (blue) activities; both are inhibited by abiraterone. Conversion of adrenal androgens to DHT occurs in prostate. C) Androgens such as DHT diffuse through the plasma membrane and bind to the AR. Upon ligand binding, the AR undergoes conformational changes involving an NH2-/carboxyl-terminal interaction and receptor stabilization. The AR translocate to the nucleus where dimerization and DNA binding to regulatory androgen response elements (ARE) occurs. Hormones gonadotropin-releasing hormone (GnRH); follicular stimulant hormone (FSH); luteinizing hormone (LH); Adrenocorticotropic hormone (ACTH); androgens dehydroepiandrosterone (-sulfate) (DHEA-S); androstenedione (AD); (cytochrome P450 family 17 subfamily A polypeptide 1 (CYP17A1); DHT (dihydrotestosterone); CBP (CREB-binding protein); ARE (androgen response element); hsp (heat shock protein); SRC-1 (steroid receptor coactivator 1).

1.2 Germline variants and prostate cancer risk

1.2.1 Role of germline variants in cancer

The genomes of two individuals differ for ~20,000–25,000 single nucleotide polymorphisms (SNPs) within coding regions that affect the functions of hundreds of genes. Germline studies related to cancer susceptibility unraveled key information years before the advent of whole-exome sequencing and whole-genome sequencing (WGS), thanks to array based technologies. On one hand, studies of inherited family history with rare germline mutations suggested high penetrance polymorphic loci. On the other hand, direct testing and sequencing of large number of cases (affected individuals) and controls (no affected individuals) has addressed the identification of thousands of germline variants associated with pediatric, childhood and adult cancer. Genetic information will eventually help clinical decision, which is increasingly under the influence of new knowledge of cancer-related pathogenic germline variants.

By integrating the genetic background of patients with tumor cells NGS analysis, cancer genomic studies could query what are the links between somatic alterations that arise during tumor development within a genetic component. A large integrative analysis of germline and somatic mutations in over 4,000 cancer cases from TCGA showed that the frequency of rare germline truncations within genes associated with cancer susceptibility demonstrated high and variable frequencies in some tumor types as 4% in acute myeloid leukaemia (AML) and 19% in ovarian cancer (Lu *et al.*, 2015). Huang and colleagues recently presented a catalog of germline variants from 10,389 individuals including 33 cancers types; a total of 853 pathogenic, or likely pathogenic, variants were discovered in 8% of adult cancer cases (Huang *et al.*, 2018). Ideker and co-workers through a TCGA pan-cancer analysis showed that germline variants are not only associated to the type of gene mutations, but also on tissue site of tumor development (Carter *et al.*, 2017). These computational analyses laid the foundation of one of the most intriguingly hypothesis discussed nowadays. Might germline variants act as potential co-oncogenes? What has emerged from recent studies is that the key biological processes required for the initiation of carcinogenesis can be altered by the effect of germline variants, but these events alone may not be sufficient to initiate malignant transformation. Compensatory pathways in normal tissues hide the potential malignity of these germline variants, but once this equilibrium is dysregulated an oncogenic or onco-suppressor process begins. What is plausible is that somatic mutations disable these compensatory pathways or activate complementary oncogenic processes. However, there is currently a substantial knowledge gap between germline variants cancer associations and the full understanding of how these risk variants contribute to human diseases (Li *et al.*, 2013; Chen *et al.*, 2014; Kanchi *et al.*, 2014).

1.2.2 Non-coding variants in cancer

TCGA and ICGC consortia sequenced the protein-coding portion (2%) of the genomes of approximately 11,000 individuals across 33 types of cancer (https://portal.gdc.cancer.gov) and of 20,000 individuals across 22 types of cancer (https://dcc.icgc.org), respectively. When noncoding regions were investigated by both consortia as part of the Pan Cancer Analysis of Whole Genomes (PCAWG) project, DNA alterations in noncoding regions involved in gene regulation showed that alterations in non-coding regions of DNA occur at a similar frequency to those in the protein-coding regions (Weinhold *et al.*, 2014). Non-protein-coding regions studies in the context of cancer susceptibility are highly relevant, in light of the Encyclopedia of DNA Elements (ENCODE) project analysis that demonstrated that 80% of the non-coding genome is -in fact- related to non-coding functional RNA, pseudogenes, introns, *cis* and *trans*-

regulatory elements, telomeres and transposons; all these elements can be involved directly or indirectly in transcriptional gene regulatory processes (Consortium *et al.*, 2012). The transcription mechanism involves several components of the genome starting from genes that regulate the transcription of other genes, enhancers, promoters, operators, insulators and silencers. SNPs present in transcription factor binding motifs may influence the binding affinity of a transcription factor (TF) in that locus. In addition, events such exposure to DNA damage or hormone levels changes can cause a second alteration in the locus or an alteration of the TF gene expression. Together, two events may alter transcription regulation that may affect the phenotype. Instead of residing in coding regions and altering protein function, SNPs can reside in each of these regulatory regions and control directly or indirectly gene expression (Amin Al Olama *et al.*, 2015; Chen *et al.*, 2015). However, the alteration of normal transcription activity due to germline and somatic mutation may or may not translated in a poor prognosis. Indeed, it was shown that bladder cancer patients harboring a SNP in the TERT promoter present better survival compared to patient carrying the ancestral allele (Rachakonda *et al.*, 2013).

To understand the potential effect of a SNP within a non-coding region is necessary to determine the features of the functional element that embeds the variant. ENCODE has provided important information to the research community by delineating all functional elements in the human genome. Histone and methylation marks, DNAase sites, and transcription factors–binding sites in protein coding and non-coding regions are important annotations which may provide more insight into Genome Wide Association Studies (GWAS) of SNPs as risk variants. High-throughput assays that capture biochemical changes such as ChIP-qPCR, ATAC-seq and Hi-seq at chromatin levels provided important information regarding the upstream and downstream effects of quantitative trait loci (eQTLs) on the regulatory machinery of the cells, including alteration of chromatin structure (Delaneau *et al.*, 2019). Furthermore, studies of three dimensional structure (3D) of genome translated into chromosome conformation capture (3C) technology have demonstrated that regulatory sequences can control transcription by looping to and physically contacting target coding genes that are located far way. The 3C technology probes one-versus-one contacts in the 3D space of the genome has evolved in probe one-versus-all (4C), many-versus-many (5C) and all-versus-all (HiC) contacts (de Laat and Dekker, 2012; Hughes *et al.*, 2014). Moreover, eQTL analysis based on the hypothesis that risk loci contain variants located within regulatory elements enables the investigation of the effect of SNPs on gene-expression levels, which may in turn affect phenotype (C. and T., 2013; Lappalainen *et al.*, 2013). Importantly, single risk variants might not be able to disrupt the normal equilibrium of a cell, but the combination of more than one SNP may influence cancer predisposition (Jansen *et al.*, 2017). For example, although common SNPs only confer weak to modest risk to PCa, some studies indicate that if a man carries multiple risk alleles, the risk increases (Eeles *et al.*, 2013).

All this information coming from advance technologies, including bioinformatics and statistical tools, have been crucial for the investigation of biological mechanism of interaction of germline variants in functional regions. Characterization of putative risk variants present in non-coding regions are difficult to achieve due to the fact that the functionality of the region might be still not annotated, or because risk variants may regulate transcription through a 3D physical contact by involving other players and further increasing the difficulty of experimental validations. Alternatively, the functional variants, discovered by GWAS, are in linkage disequilibrium (LD) with exonic variants that directly influence gene products. The *in silico* prediction of possible implication of germline variants in pathological condition needs to be followed by *in vitro* and/or *in vivo* validation.

1.2.3 Role of germline variants in PCa

A significant effect of heritability was observed for PCa and the estimated hereditary component is slightly higher in the youngers than in the olders (Lichtenstein *et al.*, 2000; Hjelmborg *et al.*, 2014). GWAS have been crucial to decipher PCa genetic component by identifying more than 160 common polymorphic loci associated with susceptibility to the disease. More recent work also found enrichment for germline variants in DNA Damage Repair (DDR) genes in metastatic PCa compared to localized PCa, including BRCA1/2 and ATM (Pritchard *et al.*, 2016; Mandelker *et al.*, 2017). Further, men carrying with the risk allele of rs25673 (19p13.13) were fourfold more likely to harbor PTEN mutations (Carter *et al.*, 2017).

Notably, almost all reported PCa risk variants reside in non-protein coding regions of the genome. Thus, finding the actual targets of these risk regions and determining how they drive the development of complex traits present a major challenge. 8q24 region has been largely study due to the presence of alleles that predispose to many cancers including prostate, breast and colon cancer (Ahmadiyeh *et al.*, 2010; Tong *et al.*, 2018). The 8q24 identified variants are located

in a "gene desert," a few hundred kilobases telomeric to the Myc gene. Histone methylation and acetylation marks and 3C assays suggested that these SNPs are within regions that act as enhancers for MYC in a tissue-specific manner (Sotelo *et al.*, 2010).
The difficulty on studying variants residing in non coding regions of the human genome make it still unclear what is the big picture of the biological implication of the inherited component of PCa. Indeed, only few studies so far have deeply investigated the mechanisms involved in the progression of PCa in the presence of risk variants.

In 2018, two interesting papers demonstrated that a SNP on chromosome 19, rs11672691, is associated not only with predisposition to PCa, but also with the aggressiveness of the disease and suggested the mechanism of action behind its role (Gao *et al.*, 2018; Hua *et al.*, 2018). The germline variant resides within an intron of a noncoding RNA, PCAT19. This lnRNA is more expressed in PCa tumors than in normal prostate and is adjacent to CEACAM21, located 100 kb away. To understand the mechanism of action of rs11672691 in CEACAM21 and or/ PCAT19 regulation, the authors performed a series of experiments in 22Rv1 prostate cells, that are heterozygous for the minor allele (A/G).

Gao and his group found that the region is enriched for enhancer-associated histone modifications and determined that the risk allele (A>G) creates a higher-affinity binding site for the transcription factor HOXA2 (Gao *et al.*, 2018). This claim was supported by ChIP-qPCR in 22Rv1cells, heterozygous for the variant. They used CRISPR/Cas9 system to obtain homozygous 22RV1 cells for the minor allele (G/G) as described in Ran et al. (Ran *et al.*, 2013) In those cells they observed a higher binding of HOXA2 and an increased expression of PCAT19 and CEACAM21. On the other hand, in VCaP cells, homozygous for the ancestral allele (A/A), the opposite effect was observed. Moreover, they observed that the expression of HOXA2 in human tumors is higher compared to normal prostate tissues and that high levels associate with increased risk of recurrence and shorter survival. Their work implicated HOXA2 in PCa progression and specifically as a direct mediator of increased gene expression at the PCAT19/CEACAM21 locus associated with the A-to-G 'mutation' at rs11672691.

Hua and colleagues investigated the same locus and drew complementary conclusions (Hua *et al.*, 2018). In their motif analysis, they identified a potential binding site for the tumor suppressor NKX3.1 on the rs11672691 region. Using the heterozygous 22Rv1 cell line, they showed that the A-to-G mutation reduces the affinity for this factor; moreover, they showed that an additional PCa-associated SNP located 36 bp away similarly affects binding of the transcription factor YY1. The rs11672691 SNP resides within the promoter region of the shorter isoform of PCAT19 and the A-to-G mutation results in the downregulation of the short form in favor of the longer form, PCAT19-long, the promoter of which is located 40 kb away. PCAT19-long isoform, together with another protein HNRNPAB, activate other genes involved in the progression of tumorigenesis.

The two studies suggested a model in which two transcription factors, a tumor suppressor (NKX3.1) and HOXA2, compete for binding to a control element that can switch between promoter and enhancer function to govern the isoform of lncRNA (PCAT19), which in turn contributes to prostate cancer progression. These studies provide an excellent example of functional characterization of risk SNPs in the context of PCa.

1.3 Hormones and PCa

1.3.1 Androgens and the Androgen receptor

Androgens are steroid hormones, which in men are formed primarily in the testes (95%) and the adrenal gland (5%). Formation of androgens in endocrine glands occurs by two biosynthetic pathways; pregnenolone and progesterone (**Figure 1**). Examples of androgens in adult males are testosterone, DHT, androstenedione, dehydroepiandrosterone, and androstenediol. In the prostate, DHT is the most potent androgen and is required for growth, development, and secretion of seminal fluid (Watson, Arora and Sawyers, 2015).

AR, together with the estrogen receptor (ER), the glucocorticoid receptor (GR), the progesterone receptor (PR) and the mineralocorticoid receptor (MR) is a member of the steroid hormone group of nuclear receptors. In prostate, AR is differently localized as described below; the nuclei of all luminal epithelial cells, most stromal cells and occasional basal epithelial cells. AR signaling plays a key role in the development and normal function of the prostate and its alteration leads to PCa development and progression. Despite a good response to androgen deprivation (AD) therapy, most patients

develop CRPC with an alteration of AR levels, hypersensitivity or AR variants. The *AR* gene, located on the X chromosome (q11-12), is more than 90 kb long and codes for a protein that has four basic domains: an unstructured regulatory region, the N-terminal domain; the DNA-binding domain (DBD); the hinge region (H) and Ligand-binding domain (LBD). Prior to ligand binding of DHT and testosterone, the AR is held inactive through association with heat shock proteins (HSP90, HSP70, and p23). DHT displaces HSPs from AR and the receptor rapidly translocate to the nucleus. With respect to the testosterone molecular structure, the DHT molecule lacks a single double bond on ring A; this peculiarity was shown to increase the affinity for AR of two-folds and to decrease the rate of dissociation fivefold relative to testosterone (Sadar, 2003). Once translocated into nucleus, AR binds to promoter and enhancer of its target genes and potentially to other unknown functional regions of genome. Despite the important implication of AR in genomic regulatory process, several studies have shown a cytoplasmic involvement in which AR nuclear translocation or AR-DNA binding are not required. A cytoplasmic activity of AR through mitogen-activated protein kinase (MAPK) cascades converges to extracellular signal-regulated kinase (ERK) activation (Bluemn *et al.*, 2017).

AR can bind enhancer and promoters simultaneously and regulate the expression of their target genes but how this complex works remains still unclear. Experimental procedure like electrophoretic mobility shift assay (EMSA), DNAseI footprinting, chromatin immunoprecipitation (ChIP) assay and functional analysis have been directed towards resolving the manner in which ligand-activated AR interacts with AREs. Like all nuclear receptor, DNA binding domain (DBD) is involved in contacting the protein domain with the consensus motif. All classical steroid receptors recognize identical response elements, which consist of two hexamer half-sites (5'-AGAACA-3') arranged as inverted repeat separated by 3 nucleotides. AR forms a symmetric, head-to-head dimer that is nearly identical with the dimer seen in the ER DBD–DNA and GR DBD–DNA structures (Shaffer *et al.*, 2004). This common pattern of structure conformation and binding site is differing from selective AREs, which consist of two hexamer half-sites repeating in the same strand. Zinc finger structure helps to distinguish binding affinity for AR and GR. One zinc finger is involved in direct DNA binding, which recognizes the specific hormone response element half-site 5′-AGAACA-3′. The other zinc finger is involved in a "head-to-head" dimerization of AR through the D-box (Tan *et al.*, 2015). Monomeric AR binding and different organization of two hexamers, as direct, inverted or everted repeats separated by different number of nucleotides are also proposed as binding patterns of this nuclear receptor. Although *in silico* prediction may suggest strong affinity for AR in the AREs sites, *in vitro* validation not always confirmed these results. An example is reported for TMPRSS2, which enhancer is situated 13,5 kb upstream of the gene and is supposed to bind AR with a strong affinity. Instead, transient transfection experiments observed a low affinity for AR, suggesting that other TFs are cooperating, making traditional way of AREs identification more difficult (Wang, Carroll and Brown, 2005).

The androgen activity is a central axis in primary PCa and its targeting remains the main strategy of treatment used by now. AR is implicated in ETS fusion genes regulation, in tumors with *SPOP* or *FOXA1* mutations associated with higher AR transcriptional activity (Abeshouse *et al.*, 2015).

1.3.2 Cofactors, cooperators and collaborators of AR

Nuclear receptors can activate gene transcription through interaction with basal transcription machinery or through the interaction with a more complex system made by coactivators such as p160 family (SRC-1, 2 and 3), CBP and p300, and cofactors represented by histone acetyltransferase and (HAT) and histone deacetylation (HDAC). Most coactivators host a HAT domain and promote histone acetylation and gene transcription. Instead, HDAC is correlated to gene repression in presence of AR and corresponding hormones (Shang, Myers and Brown, 2002). It has been shown that HAT-containing coactivator complex recruit AR to both enhancer and promoter region, followed by recruitment of RNA polymerase II and histone acetylases. Conversely, AR repression complex involves the recruitment of corepressor only in promoter regions (Takayama and Inoue, 2013).

SRC-1, -2, -3 are ubiquitous cofactors of the p160 group with similar structure. One of their roles is facilitating the recruitment of AR in AREs. Recent papers investigated the link between AR activity and coactivator proteins with the aim to elucidate the mechanism leading to castration resistance and metastatic events in PCa. SRC-2 is implicated in the regulation of PCa cell survival, metabolism and promotion of metastatic growth under AR deprivation conditions. The role of SRC-3 in PCa progression and resistance has been also investigated. In fact, *in vivo* studies in PTEN knock down mice shown that deletion of SRC-3 reverts tumor aggressiveness and its levels correlate with PCa grade (Gnanapragasam *et al.*, 2001). As for the other SRC members, SRC-3 is subject of several post-translation modifications such as

ubiquitination by SPOP protein part of Ubiquitin E3 ligase family (Culig, 2016). In contrast to SRC-2 and 3, SRC-1 seems to be involved in the regulation of AR activity by potentiating the effect of low concentrations of androgenic hormones similar to those measured in patients under androgen withdrawal therapy.

In addition to all studies focused on the activity of AR proximal to regulatory elements of target genes, many tried to elucidate the mechanism by which AR mediates transcription from far apart. Before the era of genome-wide mapping TF binding, AR binding activity was confined on Kallikrein-3 (*KLK3*) gene locus which is upregulated in cells overexpressing AR (Young *et al.*, 1991; Henttu, Liao and Vihko, 1992). Advanced technologies display a new model of distal cis-regulatory elements for AR-binding sites which turns out in the control of AR not only in protein coding regions but also in non-coding regions of the genome. The spatial communication of distal enhancers and target promoters is sustained by transcripts from active enhancer regions, termed enhancer RNAs (eRNAs), leading to transcriptional activation elements (Hsieh *et al.*, 2014).

Forkhead box 1 (FOXA1) and GATA-family member are two DNA-binding collaborating factors that establish and maintain AR-mediated chromatin loops. FOXA1, described as a pioneering factor, induces an open chromatin state through interactions with histones H3 and H4 (Cirillo *et al.*, 1998, 2002) and helps AR interaction with a broad spectrum of enhancer sites. Thanks to FOXA1, this interaction is expanded in distal regulatory elements at specific gene loci (Ernst *et al.*, 2011). FOXA1 expression regulates the transcription of some AR targets including CCND3, MYC and CDK6 (Wang *et al.*, 2007) while its inhibition does not affect the expression of other AR target genes such as KLK3, TMPRSS2 and PDE9A (Sahu *et al.*, 2011). GATA family member play a role similar to FOXA1, inducing open chromatin formation, maintenance and recruitment of AR to target genes (Cirillo *et al.*, 2002). Beyond coactivators, cofactors and collaborators, recent evidence suggests that localized chromatin interactions such as those involving AR are largely defined by global chromatin organization governed by CCCTC-binding factor (CTCF).

The specific combinations of cofactors, collaborators and coactivators recruited to AREs likely provide a mechanism for tissue specific and ligand specific gene expression. Through these actions, it is apparent that AR promotes PCa survival and cell proliferation (Lin *et al.*, 2018). If the direct implication of AR in tumor growth and sustainment has been already demonstrated, the mechanisms contributing to the androgen independence lesions remain still unclear. Genetic alterations that support androgen-independent prostate cancer growth, including activation of MAPK, phosphatidylinositol 3 kinase /AKT and protein kinase C pathways, which converge on activation of AR (Edwards and Bartlett, 2005; Shand and Gelmann, 2019), need to be investigated in a contest of cellular and molecular changes associated with AR.

1.3.3 CEBP family and AR

The CCAAT/enhancer-binding proteins (C/EBPs) are members of the basic leucine zipper (bZIP) class of TFs that contain a C-terminal basic DNA-binding domain; a leucine zipper domain involved in homo- or hetero-dimerization; a N-terminal region containing transcription activation (TAD) and regulatory domains (RD) that interact with basal transcription apparatus and transcription co-activators. These domains are strongly conserved regions (CRs) in all members of the CEBP family. The CEBP element has a divergent dyad repeat sequence RTTGCGYAAY, in which R and Y represent A/G and C/T, respectively (Osada *et al.*, 1996). Six members (α, β, γ, δ, ε and ς) of the CEBP family and CEBPs are involved in fundamental cellular processes, including proliferation, apoptosis, differentiation, inflammation, senescence and energy metabolism (RAMJI and FOKA, 2002), but experiments with C/EBPβ and α knockout suggest that these are the most essential members of the CEBP family (Bégay, Smink and Leutz, 2004). From all the protein members, CEBPα has the unique role of inhibiting cell proliferation and it also acts as a tumor repressor. The CEBPβ gene encodes an intron-less transcript which is translated into three isoforms, namely the long LAP* and LAP and the short LIP isoform (Descombes and Schibler, 1991). Both LAPs are transcriptional activators, whereas the LIP isoform lacks the transactivation domain and part of the regulatory domain (Descombes and Schibler, 1991). Beyond some important function related to myelopoiesis, macrophage functions, mammalian gland formation, liver regeneration (Luedde *et al.*, 2004) and bone metabolism, CEBPβ alteration activity is linked to tumorigenesis. CEBPβ activity seems to be connected to different tumor types such as leukemia, ovarian, colorectal and prostate cancer (Adamo *et al.*, 2018; Liu *et al.*, 2018), indicating a huge implication of this TF with some of the most important pathways that lead to an incontrollable cell proliferation and differentiation. By investigating CEBPβ implication in physiological and

pathological conditions some of the mechanisms shared by this TF and other proteins involved in gene regulation have emerged. The interplay between HATc/HADACs domains with TAD of CEBPβ indicates that CEBPβ can act not only as a repressor but also an activator of gene expression (Mink *et al.*, 1996; Guo *et al.*, 2001; Kovács *et al.*, 2003; Di-Poï *et al.*, 2005; Ki *et al.*, 2005). The implication of CEBPβ in gene regulation are supported by further studies of interaction with the SWI/SNF chromatin remodeling complex, which shuffles nucleosomes along the DNA and thereby facilitates the binding of the gene transcription machinery (Steinberg *et al.*, 2012). CEBPβ is a target gene of mTOR and low mTOR levels leads to increased expression of two LAPs isoforms and deceased LIP expression with a consequence of metabolic alteration in healthy mice (Zidek *et al.*, 2015).

As a TF involved in metabolism, C/EBPβ regulates also the expression of steroidogenic genes and its activity is modulated in response to DHT, estrogen and progesterone. Describing implication of CEBPβ in different mechanism of cell equilibrium maintenance nominate CEBPβ as an important protagonist in cancer development and progression as well, including PCa. C/EBPβ activity affects several facets of PCa disease progression, starting from first stage of inflammation (Wang, Bergh and Damber, 2007), to the regulation of metastatic genes and PCa cell survival (Kim and Field, 2008; Kim, Minton and Agrawal, 2009). It has been suggested that CEBPβ can act as a co-repressor of AR in PCa while androgen dependent cells present low levels of CEBPβ due to the inhibitory activity of AR on the promoter of CEBPβ gene. On the other hand loss of C/EBPβ leads to a reduction in the number of senescent cells following AD while ectopic expression of C/EBPβ induces the expression of senescent markers. These results indicate that C/EBPβ plays a central role in cellular senescence induced by AD, and that impeding the senescent response via inhibition of C/EBPβ expression keeps PC cells susceptible to chemotherapy (Barakat *et al.*, 2015).

In the data set of Grasso et al. 2012 (Grasso *et al.*, 2012), CEBPβ expression levels were significantly elevated in 35 CRPC samples compared with 59 localized PCa samples. The controversy information coming from Barakat and Grasso data set of possible implication of AR and CEBPβ in CRPC can be explained by the fact that there is substantial divergence in AR gene targets when comparing castrate-resistant to androgen sensitive cells. A considerable implication of CEBPβ in some of the most critic phases of tumor development and progression juxtapose to AR suggests the requirement for further investigation of cooperation between these two TFs in PCa development and progression.

1.4 Genome editing to study the effect of variants on cell biology

Most of the susceptibility loci identified through GWAS studies fall within non-coding regions of the DNA and are particularly enriched in regions characterized by high chromatin accessibility and TF occupancy (Ernst *et al.*, 2011; Consortium *et al.*, 2012). Luciferase assay, ChIP-qPCR, Starr-seq are some of the methods that can be applied *in vitro* on cells harboring the allele of interest to understand the allele dependent functionality. When cell lines with the relevant genotype(s) of interest are not available, experimental characterization of the functional impact becomes more complicated.

Genome editing is a group of technologies that enable DNA alteration. Through these technologies genetic material can be added, removed, or modified at particular locations in the genome. These include the functional knockout or knock-in of endogenous genes or alleles, and the targeted induction or correction of specific mutations or other polymorphisms. In addition to allow precise modelling of genetic contributions to normal cellular function, genome editing has obvious potential as therapeutic tool for the treatment of a wide range of diseases. Fundamental for the development of genome editing approaches is the availability of targeted endonucleases to produce DSBs at specific locations in the genome in order to stimulate the activation of the endogenous repair machinery either for NHEJ-mediated knockouts or homologous recombination (HR). To date, four main classes of targeted nucleases have been used: meganucleases, zinc finger nucleases (ZFNs), Transcription Activator-Like Effector (TALE) nucleases (TALENs) and the more recently discovered Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/ CRISPR associated (Cas) nucleases (Byrne, Mali and Church, 2014; Adli, 2018) (**Figure 2A**). CRISPR-Cas9 is a system that has generated a lot of excitement in the scientific community because it is faster, cheaper, more accurate, and more efficient than other existing genome editing methods. The mechanism has evolved into bacteria which developed this system to block infection by foreign genetic material, such as bacteriophages or plasmids. The activity of Cas9 endonuclease is coordinated by sgRNA(s) originated from the fusion between the crRNA and the tracrRNA through a four base pair loop. This resulted in a two-component system that can target virtually any genomic locus and generate DSBs. DSBs are subsequently

repaired via either NHEJ repair pathway or the precise homology direct repair (HDR) pathway. NHEJ can be used to generate gene knockouts and HDR can be used for precise editing of DNA sequences. As NHEJ is much more frequent than HDR across eukaryotic cells, especially in non-dividing cells, precise gene editing is challenging and its application in gene therapy is limited (Elliott *et al.*, 1998; Cong *et al.*, 2013; Lin *et al.*, 2014). One of the limitations is the required presence of a homologous template containing the desired change for HDR to be simultaneously delivered to the DSB in the target cells (Sternberg and Doudna, 2015). It has been shown that the efficiency of recombination decreases as the insert size increases which is considered a big issue if big sequences should be edited (Li *et al.*, 2014).

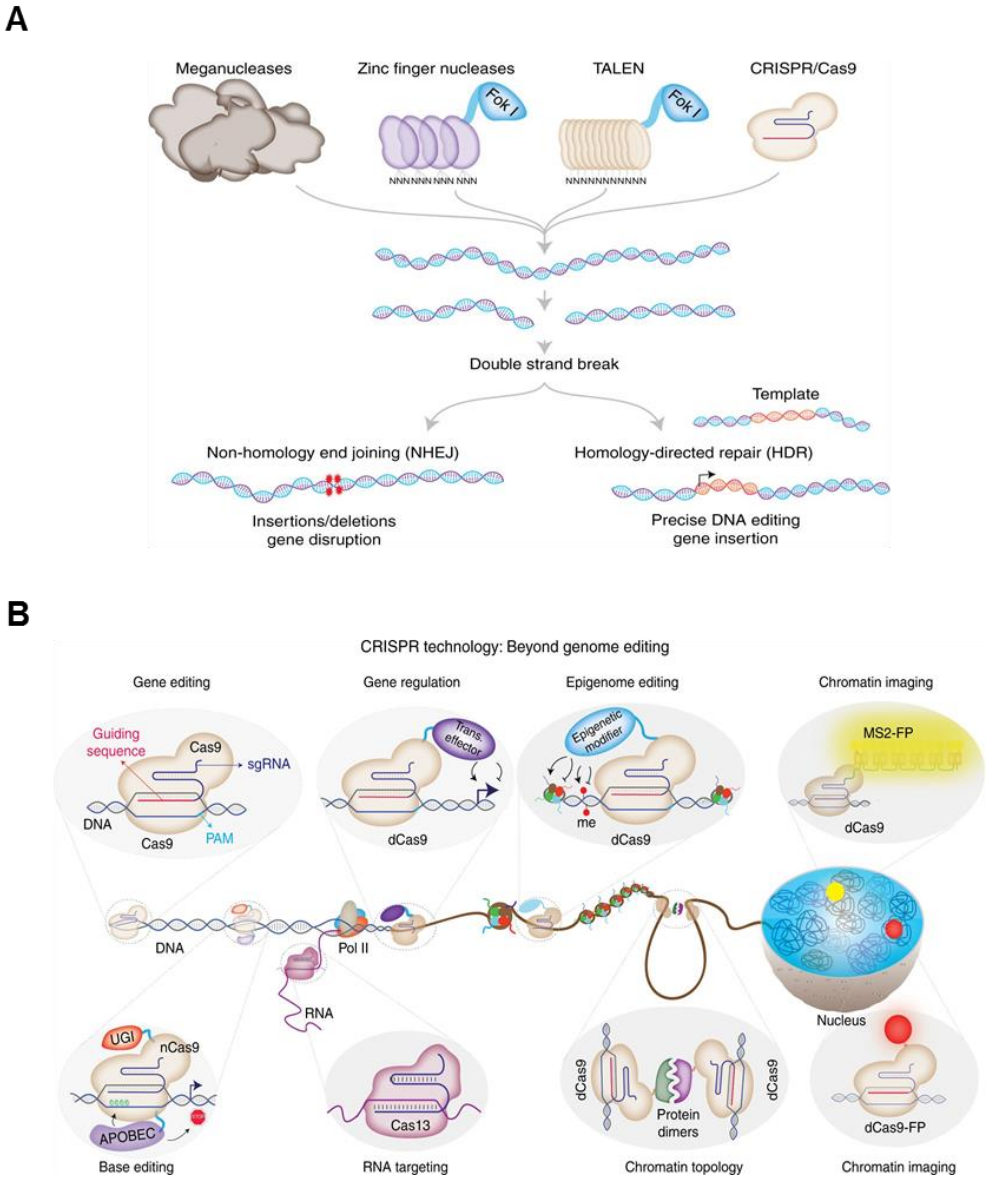


Figure 2. Representation of principle working of major genome-editing technologies. Engineered restriction enzymes Meganucleases recognize long stretches of DNA sequences (14-44 bp). The DNA binding domain of each zinc finger nuclease recognizes three bp whereas each TALE recognizes an individual base. CRISPR/Cas9 targeting specificity is mediated by RNA–DNA base pairing and the PAM sequence. Activity of each of these proteins result in DNA double-strand breaks, which are repaired either by NHEJ or HDR. While NHEJ results in random indels and gene disruption at the target site, HDR insert a specific DNA template (single stranded or double stranded) at the target site for precise gene editing. B) New era of CRISRP/Cas9 technology beyond the classic genome editing through the cleavage of the DNA by Cas9 enzyme. Altered catalytically activity of Cas9 allows a different method of editing, epigenome editing, chromatin imaging, and chromatin topology manipulations. Furthermore, the catalytically impaired nickase Cas9 enzyme has been used as a platform for base editing without double strand breaks. In addition, novel RNA-targeting CRISPR/Cas9 system has been added to the list of Cas9 targets (Adli, 2018).

Since the advent of the powerful of CRISPR/Cas9 technique in genome editing field, several attempts were made to overcome the limitations connected to the HDR efficiency and NHEJ frequency. For example, silencing of NHEJ key molecules KU70, KU80 or DNA ligase IV (Chu *et al.*, 2015) in human and mice cell line increases the HDR events by 4-5 fold. Cell cycle synchronization with chemical drugs and timed delivery of Cas9 ribonucleoprotein complexes (RNPs) maximize the efficiency of HDR up to 33%, decreases off-target events and increases cell viability due to low toxicity of the Cas9 delivery method (Lin *et al.*, 2014).

Cas9 application is not limited only in removal and substitution of genomic regions but also in transcriptional regulation, epigenetic modification and fluorescence tracking (Eid, Alshareef and Mahfouz, 2018) (**Figure 2B**). Indeed, CRISPR/Cas9 systems are now harnessed for high-throughput screening of the noncoding genome to uncover functional regulatory elements and to define their precise functions with superior speed (Sharon *et al.*, 2018). Due to high success of DNA modification as a result of different indels (insertion/deletion) induced by cooperation of Cas9 endonuclease and the guide RNA, larger scale screens have been performed using Cas9 and pooled sgRNA libraries to saturate a chosen noncoding region with indel mutations (Shukla and Huangfu, 2018). Moreover, for the application on targeting chromatin modification, the Cas9 enzyme has been modified to generate a nuclease-deficient version (dCas9) that does not induce DNA cleavage. Furthermore, the dCas9 has been implemented in activation or repression of target gene expression (CRISPRa and CRISPRi respectively).

The application of genome editing in the study of germline variants in noncoding regions has foundations in the Cas9 endonuclease activity. Basically, the technique, applied to obtain isogenic cells with the variant, is based on the DSB induced by the enzyme and the substitution of single nucleotide *via* HDR in presence of a donor DNA. Insertion of exogenous DNA up to ~100 bp can be performed with a single-stranded, synthetic DNA oligo or double strand donor DNA. Design of a single-stranded donor template (ssDNA) plays a big role in knock-in efficiency; important parameters include homology arm (HA) length, homologous arms symmetry and chemical modifications. Homology arm lengths of 30, 40 and 50 bp have optimal knock-in efficiency, while lengths of 60 and 70 bp have lower efficiency. Phosphorothioate-modified DNA templates with symmetric HA gives optimal knock-in. The efficiency of knock-in increase even more if chemical modifications are present in both 5' and 3' of oligo (Renaud et al., 2016; Richardson et al., 2016). As described above most of the limits of the approach consist on the very low frequency of HDR events comparing to NHEJ. Even though new variant of Cas9 and new strategies of increasing HDR events are speedily emerging, the capacity of Cas9 to discriminate a single nucleotide once editing *via* HDR occur is not clear.

Base editing (BE) is a newer genome-editing approach that uses components from CRISPR systems together with other enzymes to enable the efficient installation of point mutations into cellular DNA or RNA while avoiding double-stranded DNA breaks (Rees and Liu, 2018). A catalytically disabled nuclease associated with cytidine deaminase such as APOBEC1 and an inhibitor of base excision repair such as uracil glycosylase inhibitor (UGI) convert cytidines into uridines within a five-nucleotide window specified by the sgRNA. The approach, known as cytosine base editor (CBE) can readily mutate a G·C base pair to an A·T base pair (Komor *et al.*, 2016; Shimatani *et al.*, 2017). Differently, adenine deaminase (TadA) used in adenine base editors (ABEs) can convert an A·T base pair into a G·C base pair (Gaudelli *et al.*, 2017). Even though CBE and ABEs are providing an elegant method for single nucleotide substitution, it was recently discovered that CBE targets both DNA and RNA while ABE induces site-specific inosine formation on RNA (Zhou *et al.*, 2019). Disadvantages of the BE approaches regard the restriction to protospacer adjacent motifs (PAM) sequences, the limited type of substitutions that can be addressed, and the potential undesired multiple editing of the undesired nucleotide (if multiple instances are present) in an up to 9 bp wide window within the protospacer (Chatterjee, Jakimo and Jacobson, 2018; Grünewald *et al.*, 2019; Tan *et al.*, 2019).

In summary, in the context of single nucleotide editing, taking into account the limitations described above, both Cas9 editing *via* HDR and BE approaches can be applied in *in vitro* systems with the aim to create isogenic cell lines harboring the variant of interest to characterize the functional role in cancer predisposition and progression. Although genome editing technique is under continues evolution, its application in *in vitro* and *in vivo* remain still challenging.

1.5 Study specific background and rationale

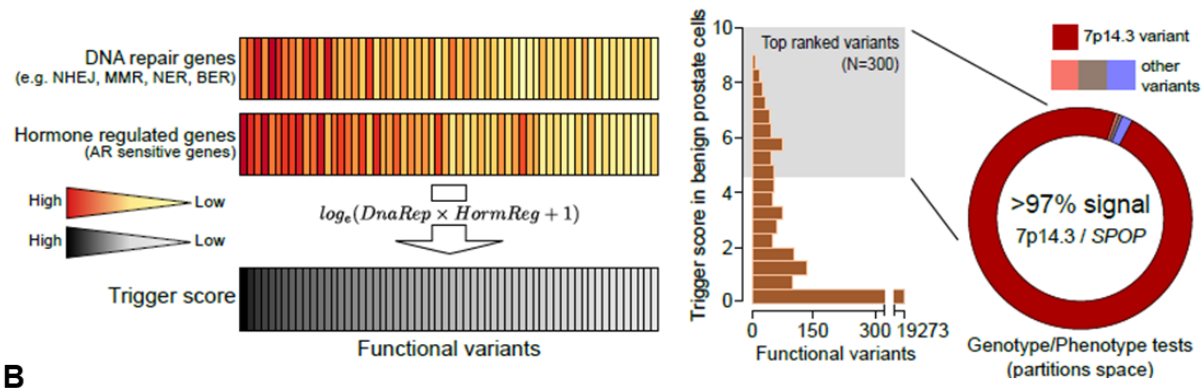
When I started my PhD training in 2015, members of our laboratory were focused on a project that stemmed from the initial hypothesis that inherited variants may be linked to early somatic prostate cancer specific genomic events. To test this hypothesis inherited variants within regulatory regions of the human genome, as defined by enhancer or promoter regions characterized through ENCODE ChIP-seq experiments, and selected common early genomic PCa events were considered, including TMPRSS2-ERG rearrangement, *SPOP* and *FOXA1* mutations. Comprehensive genomic data from more than 480 PCa patients with genetic and somatic annotations (blood DNA and tumor tissue available) and from more than 3,700 controls were mined for germline-somatic associations. They used a ranking score and benign prostate transcriptomes to identify a non-coding polymorphic regulatory element at 7p14.3 (rs1376350) that associates with DNA repair and hormone-regulated transcript levels and with the early recurrent prostate cancer-specific somatic mutation in the *SPOP* gene (**Figure 3A** and **B**). The SNP has a Minor Allelic Frequency (MAF) of about 2% in the Caucasian population (**Figure 3B, Table 1**) and increased frequency in SPOP mutant PCa. TRANSFACT based analysis identified an AR motif (M00962) where our SNP represented the nucleotides in 6th position and a CEBP family motif (M007720) where our SNP represented the 1st nucleotide. Interestingly, the prediction of binding affinity in presence of the minor allele A was significantly higher than in presence of the Major allele (**Supplementary Figure 1**, **Table 1**) for AR binding while it was unaltered for CEBP family binding (Romanel *et al.*, 2017).

To query whether there is a relationship between the polymorphic locus genomic sequence features, characterized by the proximity of AR and CEBP family TFBSs, and the somatic genotype of *SPOP* mutation in PCa, a member of our laboratory (Davide Dalfovo) recently investigated the frequency of this same co-occurrence genome-wide. Briefly, coupling updated ChIP-seq ENCODE data for regulatory elements consensus with Transcription Element Search System (TESS) (Schug, 2008) results, he identified about 2.7M and 5.7M sites for AR and for CEBP (Cohen *et al.*, 2018), respectively, encompassing a total of 196K and 399K SNPs with minor allele frequency > 1%. Interestingly, when then testing the association between the SNPs genotypes and SPOP mutation status in the same cohort from our original study, 7 out of 145 (4,8%) SNPs within co-occurrent TFs versus 9 out of 548 (1,6%) within non-co-occurrent TFs showed association (p-val = 0.0319, Fisher Exact test). When combining the 7 SNPs, the strongest SPOP association signal was found with the rs1376350 combined with a variant on chr9q31.1. This analysis supports the link between sequence specific inherited variants and the distinct PCa SPOP subclass.

The *in silico* prediction of enhancer activity at the locus was supported by *in vitro* experiments with luciferase assay. The binding affinity for CEBPβ and AR was supported by ChIP-qPCR data from human prostate carcinoma cells (PC-3). The SNP resides within a desertic region, where the nearest genes, with no biological connection to the prostate gland, reside ~500 kb up stream and ~250 kb downstream (**Figure 3C**). Specifically, Bone Morphogenetic Protein-Binding Endothelial Cell Precursor (*BMPER*) codify for BMP-binding endothelial regulator protein, 5'-Nucleotidase, Cytosolic 3A (*NT5C3A*) encodes a member of the 5'-nucleotidase family of enzymes that catalyze the dephosphorylation of nucleoside 5'-monophosphates and Bardet-Biedl Syndrome 9 (*BBS9*) coding for Bardet-Biedl syndrome 9 protein part of BBSome complex.
No association was detected between the genotype and the total number of somatic single-nucleotide variants (SNVs) in the human tumor data, although an increased somatic genomic burden in men with the minor allele associated with SPOP mutant prostate cancer was observed.

Based on this data, I started my PhD project with the overall aim to eventually characterize the locus by creating the most relevant system(s) to model the human data. As no available prostate cells harboring the germline variant exist, my work consisted on the establishment of isogenic cells via CRISPR/Cas9 with the aim to validate our hypothesis by using the homozygous (A/A) and/or heterozygous (A/G) model. During this challenging process, I also developed alternative strategies to help test our initial hypothesis on the link between germline variants and prostate cancer subclasses.

A



B

Table 1 Association signal of variant 7p14.3 with SPOP somatic phenotype

Cohort.	AA+AG Carriers	GG Carriers	MAF	Comparison with SPOP wt. Tumors		Comparison with tyrol control (N=1,014)		Comparison with Tyrol Extended Controls#		Comparison with 1000 Genome (N=2,504)		Comparison with all Controls## (N=3,795)	
				OR	P	OR	P	OR	P	OR	P	OR	P
Adenocarcinomas Discovery	24	217	0.052	5.75	3.0e -4	10.47	1.1e-0.7	10.2	9.1e-8	5.44	3.0e-5	6.38	4.5e-6
Adenocarcinomas Validation	23	217	0.048	4.45	4.1e- 3	8.22	8.8e-0.6	7.9	9.1e-6	4.04	1.5e-3	4.73	4.0e-4
Adenocarcinomas Complete	47	434	0.050	4.83	6.7e-6	9.20	1.54e-10	8.9	7.4e-11	4.72	3.0e-7	5.54	1.22e-8
Adenocarcinomas EUR only	42	373	0.052	4.96	3.7e-5	10.07	1.3e-0.9	9.75	7.4e-10	7.86	1.1e-7**	8.44	8.8e-10**
Validation Korean	19	61	0.206	5.84	4.0e-2	-		-	-	4.78	4.3e-2***	-	-

C

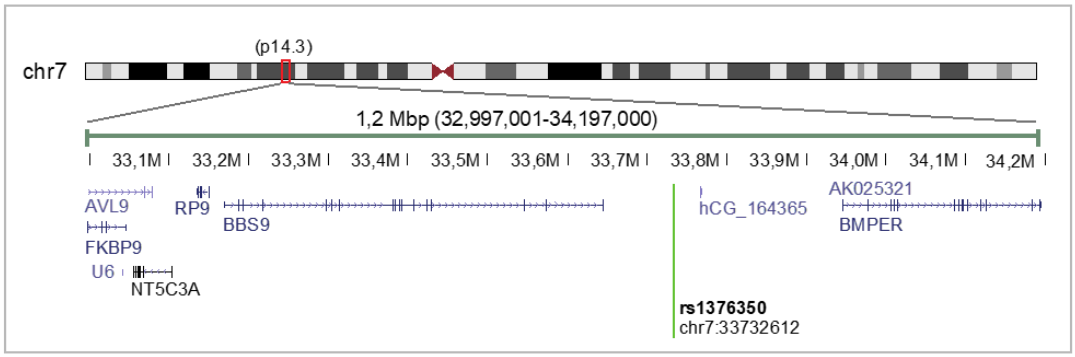


Figure 3. *In silico* prediction of an association between 7p14.3 with SPOP mutant prostate cancer. A) The number of DNA repair and hormone-regulated genes from healthy prostate cells that are modulated by a functional variant are combined into a ranking score that measures the likelihood to observe a prostate-specific early somatic event. Trigger score distribution (left) across all considered functional variants; top ranked variants are highlighted. Genotype/phenotype analysis (right) is performed on random partitions of the data set into discovery and validation sets for three early recurrent prostate cancer lesions (SPOP mutations, FOXA1 mutations, and TMPRSS2-ERG rearrangement). A 7p14.3 variant associated to SPOP was implicated in 97.4% of all collected associations (187 of the 192 partitions for which association signal was detected, red portion of the ring plot). No variants in the partition space for FOXA1 and TMPRSS2-ERG lesions were identified. B) Results refer to logistic regression analysis using dominant model corrected for age and PSA. First three rows show data from a random partition (discovery and validation) and the complete data set; columns include signal upon data set extension to controls from the Tyrol PSA Screening Cohort and the 1000 Genomes Project individuals' collection. Data is also reported for EUR descent individuals only and for an independent cohort of Korean patients (EAS from 1000 Genomes Project collection included as controls). D) SNP rs1376350 reside in the short arm of chromosome 7 (7p14.3). In 1Mb window the nearest genes are BMPER, NT5C3A and BBSome complex. #ETS positive/SPOP.wt tumors and controls; ##analysis not corrected for age and PSA; **EUR individuals only included (N = 503); ***EAS individuals only included (N = 504); Bone Morphogenetic Protein-Binding Endothelial Cell Precursor (BMPER); Cytosolic 5'-nucleotidase 3 (NT5C3A), Bardet-Biedl syndrome 9 protein (BBS), Retinitis Pigmentosa 9 (RP9), Late secretory pathway protein AVL9 homolog (AVL9), Peptidyl-prolyl cis-trans isomerase 9 (FKBP9).

2 Methods

2.1 Cell culture material and methods

2.1.1 Cell lines

All prostate cell lines were grown in RPMI medium supplemented with 10% FBS, 2 mM L-Glutamine, 10 U/ml penicillin, and 10 µg/ml streptomycin. Cells were maintained in 37°C with 5% CO2 incubator. Sex hormone depletion (androgens and estrogens), prior to DHT (Sigma-Aldrich) treatments, was achieved by growing the cells in medium without phenol red (Euroclone, Celbio), supplemented with 10% charcoal/dextran treated FBS (Hyclone, Celbio) for 48 h. The cell lines were purchased from ATCC (American Type Culture Collection, LGC Standards). PC-3 (derived from prostate cancer metastatic site; bone) and LNCaP (derived from prostate cancer metastatic site; lymph node) are GG at rs1376350 (GSM888588, GSM888346). The primary epithelial cell line Hs578Bst (derived from normal breast tissue) is AG at rs1376350. It was maintained in DMEM medium supplemented with 10% FBS, 2 mM L- Glutamine, 10 U/ml penicillin, 10 µg/ml streptomycin and 30 ng/ml EGF Recombinant human protein (Life Technologies). 293T/17 cells were obtained from (Anna Cereseto's laboratory) and maintained in DMEM supplemented with 10% FBS, 2 mM L-Glutamine, 10 U/ml penicillin, and 10 µg/ml streptomycin.

2.1.2 Transfection of prostate cells

Fugene (Promega), Lipo2000 (Invitrogen), TransIT-LT1 and TransIT-X2 (Mirus) were initially tested on PC-3 and LNCaP to maximize the transfection efficiency. For transfection of plasmids, 300.000 PC-3 cells were seeded on 6 well plate and transfected with 2.5 µg DNA and FuGENE HD (Promega). Transfection was scaled up to 1 x 10^6 cells on a 10 cm dish and transfected with 7.5 µg of DNA. DNA template and transfection reagent were mixed with Opti-MEM (Invitrogen) and incubated for 15 min before being added to the cell dropwise. The day after, cells media was replaced. Selection was applied 48 to 72 hours after transfection. LNCaP transfection was performed as for PC-3 but Lipo2000 (Invitrogen) was used instead of FuGENE.

2.1.3 Lentiviral vector production and transductions

Lentiviral particles were produced by seeding 293T/17 cells into a 10 cm dish. The day after, two hours before transfection the medium was replaced with fresh DMEM. Cells were transfected with 5 µg of pLenti-Cas9-Blast vector or pAIP_AR together with 2,5 µg of envelope plasmid pMD2G and 2,5 µg of psPAX2 packaging vector by using $CaCl_2$ 2.5 M and 2x HBBS method (Jordan and Wurm, 2004). After overnight incubation, the medium was replaced with fresh complete DMEM and 48 hours later the supernatant containing the viral particles was collected, spun down at 500 g for 5 minutes and filtered through a 0.45 µm PES filter. Vectors stocks were stored at -80°C for future use. For transductions, prostate cancer cells were seeded in 10 cm dishes and the next day viruses were added into dishes. After overnight incubation, the viral supernatant was removed and cells were kept in culture for a total of 48 hours. Cells were selected with Blasticidin (50 µg/ml) (Life Technologies).

2.1.4 Electroporation of PC-3 with Cas9-RNP complexes

Ribonucleoproteins (RNP) complexes were prepared by using Lonza SF cell line 4D- NucleofectorX, kit S recommended for PC-3 cells and purified Cas9 protein according to manufacturer's instruction. Briefly, RNP complexes were obtained by mixing: a) Tracer and crRNA annealing Mixture (combined 1:1 and heated at 95°C x 5 min, and cooled at room temperature (RT) for 20 min); b) 2.4 µl of 50 µM of donor DNA (PS-ssDNA or ssDNA- sequence reported in **Supplementary Table 14**, pre-warmed at 95°C x 5 min and immediately cooled on ice); c) 61 µM Cas9 protein (from Anna's Cereseto laboratory) and sgRNA 150 pMol (from IDT), incubated at RT x 20 min; d) 1.2 µl of Electroporation Enhancer. 2 x 10^5 cells PC-3 cells were trypsinized and resuspended in 20 µl of Nucleofector solution buffer and combined with 5 µl of RNPs complex and mixture of donor DNA and electroporation enhancer. The cell

suspension was then transferred in appropriate cuvette in the Lonza's 4D-Nucleofector System and electroporated. Cells were then recovered and plated in a 24 well plate with pre-warmed media conditioned with 25 µM Homolougous directed recombination enhancer (HDR). Once enough cells were present, DNA was extracted as described in **section 2.2.6 c**) and editing efficiency was assessed as described in **section 2.4**

2.2 Molecular Biology methods

2.2.1 Plasmids

pEGFPN1 (Clontech) and pHR_SIN_CSGW (Demaison *et al.*, 2002) plasmids were used to test transfection efficiency; SgRNA expressing plasmid under the control of U6 promoter harboring a Neomycin resistance cassette pGSsgRNA_NeoR was derived by GenesScript; pUC19 carrying sgRNAs was a gift from Anna Cereseto; High fidelity Cas9 expressing vector pX330-eSpCas9(1.1) was a gift from Yuichiro Miyaoka (Addgene # 108301); Cas9 expressing vector harboring a Blastomicin resistance cassette lentiCas9-Blast plasmid was a gift from Feng Zhang (Addgene #52962), Androgen Receptor expressing lentiviral backbone pAIP_AR was cloned in house using pAIP from Jeremy Luban (Addgene # 74171); VSV-G envelope expressing plasmid pMD2G (Addgene #12259) and 2nd generation packaging vector psPAX2 were a gift from Didier Trono (Addgene # 12260).

2.2.2 sgRNAs design

sgRNAs used for the macrodeletion (>730 bp) spanning the 7p14.3 locus were selected using the GPP Web Portal (http://portals.broadinstitute.org/gpp/public/) that ranks candidates according to their predicted on-target and off-target activity. sgRNAs for the microdeletion (50 bp) and for the disruption of DNA binding motifs around the SNP of interest were designed based on PAM 5'-NGG-3' sequences nearby the locus. sgRNA_C was designed with 21 bp instead of classic 20 bp to increase the transcription efficiency of the sgRNA scaffold were the sgRNAs was cloned (see **Supplementary Figure 2, Table 3** and **Supplementary Table 5**). All sgRNAs utilized in this work were characterized for predicted off-targets using Cas-OFFinder tool (Bae, Park and Kim, 2014) (**Supplementary Table 6**).

2.2.3 Cloning of sgRNAs

The cloning of desired sgRNAs into pUC19 and pX330 plasmids was performed by initially annealing oligonucleotides with appropriate overhangs for BbsI cutting sites. One µl of each of 100 nM stock oligos were mixed with annealing buffer and the hybridization was carried out on a thermal cycler program starting with 2 minutes at 5°C and then a gradient from 95°C to 25°C with the cooling of 1°C each 90 second. Annealed products with appropriate overhangs were cloned into the destination vectors (pUC19 and pX330 plasmid) previously digested with the BbsI restriction enzyme (NEB) so that they would be inserted between a U6 promoter-driven expression cassette and the guide RNA constant portion. Ligation was performed a RT per 1 h in a final volume of 20 µl by mixing 1 µl of Annealing product (1:500) with 40 ng of plasmid digested with BbsI in presence of 1 µl of T4 DNA ligase and its buffer (NEB). The ligation reaction was used to transform competent cells (as described in **section 2.4.4**).

2.2.4 Bacterial transformation

Competent cells were thawed on ice and mixed with 1 µl of plasmid or the whole ligation mix. After 30 min of incubation on ice and 45 sec heat shock at 42°C, 300 µl of Luria broth (LB) medium were added and incubated at 37°C on a shaking platform for 1 h. The suspension was then plated on LB-agar dishes containing antibiotics.

2.2.5 Isolation of plasmid DNA from bacteria

For plasmid isolation the PureYield™ Plasmid Midiprep (Promega) and Miniprep (Machery_Nagel) were used following the manufacturer's instructions. Briefly, bacterial suspension grown overnight was pelleted at 8000 rpm for 10 min at 4°C. The supernatant was discarded and the pellet was resuspended in P1 buffer (with RNAase). Bacterial

cells were lysed by incubation with Buffer P2 for 3 min before neutralization with buffer N3. The bacterial lysate was centrifuged at 8000 rpm for 15 min at 4°C to pellet the debris and the DNA containing supernatant was transferred into columns carrying a nucleic acid binding membrane. After centrifugation columns were washed twice with buffer PE to remove residual salt and ethanol carryover. DNA was eluted with elution buffer and then quantified with Nanodrop instrument.

2.2.6 DNA extraction

a) **QuickExtract DNA extraction** solution (EpicentreB). Cells were washed twice with PBS. For a 96 well 40µl of QuickExtract DNA extraction solution were added to lyse the cells, then transferred in a 0.2 ml microtube and incubated at 65°C for 15 min to facilitate protein digestion. A step of 2 min at 95°C was added to inactivate the Proteinase K.

b) **Phenol/Chloroform DNA Extraction**. Cells were trypsinized and pelleted. Pellets were lysed with DNA extract buffer (Tris HCl 0.01M pH8.5, 0.5mM EDTA 0.2% SDS, 0.2M NaCl) with Proteinase K for 1 hour at 65 °C or 37 °C overnight. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was then added, followed by a centrifugation (5 minutes at 16,000 g) to separate the aqueous from the organic phase. The upper aqueous phase was transferred to a fresh tube. DNA was then precipitated by adding 5 volumes of ethanol 100% and 1/10 of Sodium acetate (3 M) and 1 µl of Glicogen blu and incubated at –20°C overnight or at –80°C for at least 1 hour. The DNA was pelleted by a first centrifugation at 4°C for 30 minutes at 16,000 g. followed by a second centrifugation after a wash with 70% ethanol. The DNA pellet was then dried and then re suspended in water.

c) **DNA extraction columns**. DNA extraction was performed using NucleoSpin® kit (Macherey-Nagel) following manufacturer's instruction. Briefly, cells were washed with PBS, then lysis is achieved by incubation at 70 °C for 10 minutes of the sample in a proteinase K/SDS solution (buffer T1 and buffer B3). Appropriate conditions for DNA binding to the silica membrane in the NucleoSpin® Tissue Columns are achieved by the addition of chaotropic salts and ethanol to the lysate. The binding process is reversible and specific to nucleic acids. Contaminations are removed by subsequent washing with two different buffers (BW and B5). Pure genomic DNA is finally eluted under low ionic strength conditions in a slightly alkaline elution buffer (EB buffer).

2.2.7 Agarose gel electrophoresis

Depending on DNA fragment size, a solution of 1-2% agarose in 1X TAE was prepared. The solution was heated to promote the melting of the agarose and poured, with the addition of DNA staining Xpert green DNA stain (10000x, Grisp), into a gel chamber with combs. Samples were loaded into the wells of the polymerized gel. 10 µl of 1 kb DNA ladder (NEB) was loaded next to the samples and allowed size determination of the DNA. The gel was run at 100-120 V for one hour and then DNA was visualized using a UV transilluminator.

2.2.8 PCR amplification with SuperFi 2x Master Mix

To evaluate the editing, 25 µl PCR reaction was performed using 12.5 µl Platinum SuperFi Green PCR Master Mix (Invitrogen), 50 ng DNA template and 0.5 µM forward and reverse primers (**Supplementary Table 7**) Predicted PCR bands were further purified and sent for Sanger sequencing.

2.2.9 RNA isolation and quantification

For isolation of total RNA from cultured cells the RNeasy kit (Qiagen) was used following manufacturer's instructions. Briefly, cells were washed twice with PBS and lysed in RLT buffer with β-mercaptoethanol (β-ME). In order to increase lysis efficiency, pellet was passed at least 5 times through a blunt 20-gauge needle (0.9 mm diameter) fitted to an RNase-free syringe. One volume of 70% ethanol was added to the homogenized lysate to promote nucleic acids precipitation, and mixed by pipetting. The mixture was then transferred to an RNeasy spin column carrying a membrane that binds the RNA and then centrifuged for 1 min. After removing flow-through, DNA-ase treatment was performed for 15 min by incubating at RT in order to remove DNA carry over. Washes were performed to eliminate salts. The columns were

then transferred to a new 1.5 ml tube and RNA was eluted using 40 µl of Elution buffer and quantified using nanodrop. All centrifuges were done at ≥8000 g (≥10,000 rpm).

2.2.10 cDNA synthesis

Total RNA was transcribed into complementary DNA (cDNA) using first strand cDNA synthesis kit (Thermo Scientific) according to manufacturer's instruction. Briefly, 1 µg of RNA was retrotranscribed using Random hexamers in the presence of RNAse inhibitor RT ribo block, RT- retrotranscriptase, dNTPs and the appropriate buffer in a total volume of 20 µl. Retrotranscription was carried out for 5 minutes at 25 °C, to allow the annealing of random hexamers, followed by 1 hour at 42 °C. Reaction was terminated by heating 5 minutes at 70 °C and then chilled at 4 °C. cDNA was diluted 3 folds prior to Real Time quantitative PCR (RT-qPCR) testing.

2.2.11 Real Time quantitative PCR (RT-qPCR)

RT-qPCR was applied on cDNA samples for a relative quantification of mRNA levels and on DNA samples immunoprecipitated by ChIP. Samples were tested in triplicate using KapaSybr (Resnova) in a final volume of 10 µl. The reactions were set up using 5 µl of Kapa sybre Master Mix (2x), 0.4 µl of each primer (10 µM) and 2 µl of diluted CDNA or 2 µl DNA from ChIP for each well. The measurements were carried out with the c1000Thermal cycler (Bio-Rad) qPCR System using the following thermal cycling profile: 95°C x 3 min, followed by 40 cycles 95°C x 20 sec, 60°C x 20 sec, 72°C x 20 sec. Melting curves of the generated amplicons were obtained by presetting the thermal cycler at 72 °C and then increasing it incrementally (0.5°C each 30 second) as the instrument continues to measure fluorescence.
Analysis of relative mRNA expression was performed using the ΔΔCt method with GAPDH (glyceraldehyde 3-phosphate dehydrogenase) as reference gene (**Supplementary Table 11**) while Antibodies specific recruitment was calculated as enrichment respect to the IgG according to the ΔCt method.

2.2.12 Nuclear protein extraction

Nuclear protein extraction was performed using EpiQuik Nuclear Extraction Kit I (Epigentek) accordingly to manufacturer's instruction. Briefly, cells were washed twice with cold PBS, then scraped and centrifuged for 5 minutes at 250 g. The pellet was then resuspended in buffer NE1 containing DTT and a proteinase inhibitor cocktail at the concentration of 10^6 cell / 100 µl of buffer and incubated on ice for 10 minutes. Samples were then vortexed vigorously for 10 seconds and centrifuged for 1 minute at 11000 g. The supernatant was then removed and the nuclei pellet was lysed with 10 µl of buffer NE2 per 10^6 cells on ice for 15 minutes with vortex every 3 minutes. The suspension was then centrifuged for 10 minutes at 12000 g at 4°C. The nuclear extract was then transferred on a clean vial and quantified using BCA protein assay (Pierce).

2.2.13 Western blots

Cells were lysed in RIPA buffer (0.05 M HEPES pH 7.9, 0.14 M NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Sodium deoxycholate, 0.1% SDS) supplemented with 1% of protease inhibitor cocktail III (Calbiochem). Cell lysates were quantified using Pierce BCA protein Assay kit (Thermo Scientific) and separated by SDS-PAGE on 4-12% bis tris plus, Bolt Invitrogen (Thermofisher Sc) with Bolt MES SDS running buffer 20X for 45 minutes at 150V. After electrophoresis, samples were transferred to 0.45 µm PVDF membranes (Amersham Hybond) for 1h30 at 300 mA. The membranes were incubated with mouse anti-FLAG (Sigma) for detecting SpCas9, with a rabbit anti AR (5153, Cell Signaling technology), mouse anti CEBPβ (Sc-7962, Santa Cruz Biotechonlogy), rabbit anti Gr-1 (12041, Cell Signaling technology) or with mouse anti-α-tubulin (Sigma) for a loading control and with the appropriate HRP conjugated goat anti-mouse or goat anti-rabbit secondary antibodies (1:8000, Sigma) for ECL detection. Images were acquired using the UVItec Alliance detection system.

2.2.14 ChIP assay

PC-3 cells with and without AR stable overexpression were seeded into 150 mm Petri dishes in RPMI medium without phenol red, supplemented with 10% charcoal/dextran treated FBS. Cells were then treated with EtOH or DHT (100 nM) for 16 h and then subjected to ChIP with 3 µg of the following rabbit polyclonal antibodies: Anti Histone H3 mono methyl K4 (Abcam#8895), Anti Histone H3 acetyl K27 (Abcam #4729), Anti histone H3 trimethyl K4 (Abcam #8580), anti Histone H3 (Abcam #1791), anti Histone H3 Trimethyl LYS27 (Millipore #07-449), or a normal IgG (CS200581), using the MagnaChIP HiSens ChromatIP Kit (17-10461 Upstate, Millipore). Briefly, the experiment procedure included chromatin crosslinking with formaldehyde, nuclei isolation and lysis followed by chromatin shearing for 45 minutes with a duty factor of 10% in the Adaptive Focused Acoustic (AFA) ultrasonicator Covaris M220. Thirty µl of magnetic beads (Dynabeads) for each antibody were washed with BSA-PBS (5mg/ml) to block beads aspecific binding and then incubated overnight at 4°C on rotating wheel before adding the chromatin equivalent of 1.5 million cells for each Histone marks. After 6 hours at 4°C on a rotating wheel, samples were extensively washed 3 times with 3 increasingly stringent washing buffers, for a total of 9 washes. DNA was eluted in TE then treated RNAse A (20 µg/ml) and Proteinase K (200 µg/ml) to revert the crosslink; DNA was then purified with classic Phenol/Chloroform protocol (as described in **section 2.2.6 b**) and then resuspended in 30 µl of water. Precipitated DNA was analyzed by real-time qPCR (as previously described in **section 2.2.11**). FKBP5 promoter and enhancer regions were used as examples of AR responsive genes (Nelson *et al.*, 2002), NeuroD2 enhancer-promoter region was used as positive control for H3K27m3 and as negative control for H3K27 Ac enrichment (Stelloo *et al.*, 2018), GAPDH and nucleolin (NCL) promoter regions were used as references. Enrichments of these regions were compared with enrichments of the region of interest surrounding the 7p14.3 variant. The primers used are reported in (**Supplementary Table 13)** Antibodies specific recruitment was calculated as enrichment respect to the IgG according to the ΔCt method.

2.2.15 Biotinylated DNA pull-down assay

Twenty-eight nucleotides long oligos were designed to harbor the canonical DNA full site binding sequence for AR (Shaffer *et al.*, 2004), or to encompass our SNP of interest carrying the major allele G or the minor allele A, or a sequence lacking AR binding sites as predicted by an in house developed tool (by Davide Dalfovo). Biotinilated oligos' sequences are reported in **Supplementary Table 12**. Oligos were ordered as single strands; sense oligos were designed as carrying a biotin group at the 5', while their reverse complement were unmodified. Oligos were dissolved in deionized sterile water at the final concentration of 1 µg/µl then, for each pair, an equal quantity of sense and antisense biotinylated oligonucleotide solution, was mixed and hybridization step on the thermal cycler was performed (5 minutes at 100 °C and then slow cool down to RT). Nuclear extracts (NE) were prepared has previously described in **section 2.2.12**. For each pull down reaction 75 µg of NE were used with 40 µl magnetic beads (Dynabeads M-280 Streptavidin, Invitrogen) and 4 µg of double strand biotinylated DNA probe. Beads were resuspended and washed twice with cold PBS for at least 5 minutes on a roller. Tubes were placed on a magnet for at least 1 minute, supernatant was discarded. Beads were then pre-incubated on a roller, at RT, with Buffer B (5 mM HEPES, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 26% glycerol (v/v), pH 7.9) containing 0.2 mg/ml of sheared Salmon Sperm DNA Solution (ssDNA , Invitrogen) and 0.1 mg/ml of bovine serum albumin (BSA, Albumin Fraction V, Roth), in the presence of a cocktail of proteinase inhibitors (PI, Protease Inhibitor Cocktail Set III, Calbiochem) to eliminate nonspecific binding. After 30 minutes this blocking buffer was removed and NE and 4 µg of biotinylated dsDNA were added to the beads. After an overnight of incubation with gentle rotation at 4 °C, the precipitated matrix was washed five times at 4 °C with 500 µl of Buffer B (+ 50 mM NaCl+ PI+ ssDNA+BSA). The washed matrix was boiled in SDS-PAGE sample buffer and analyzed by western blot as previously described. Proteins pulled down using DNA sequences were then probed by western blot for the presence of AR, CEBPβ and GR-1 (as previously described in **section 2.2.13**).

2.3 Methods for assessment of genome editing

2.3.1 Tools for *in silico* prediction of editing efficiency

Two tools were used for the *in silico* prediction of editing efficiency. The Tracking of Indels by Decomposition, **TIDE**, web-tool was used for *in silico* prediction of editing efficiency (https://www.deskgen.com/landing/tide.html#/tide). Based on the quantitative sequence trace data from two standard capillary (Sanger) sequencing reactions, TIDE quantifies editing efficacy and identifies the predominant types of insertions and deletions (indels) in the DNA of a targeted cell pool (Brinkman *et al.*, 2014). For the detection and assessment of HR-mediated genome editing of our target locus by CRISPR_Cas9, we used the TIDER web-tool (https://tider.deskgen.com/). Based on the quantitative sequence trace data from two standard capillary (Sanger) sequencing reactions, plus a reference sequence, the TIDER software predicts the frequency of template small mutations in a pool of cells (Brinkman *et al.*, 2018).

2.3.2 Assessment of genome editing by Sanger sequencing

For clones screening, single cell dilution method was used in order to have one cell per well in a 96 well plate. During the first week 50 µl of fresh medium was added in each well every 2 days. After 7 days, 100 µl of fresh medium was added and cells were checked under the microscope in order to select wells having only one cell. After one month, confluent wells were splitted into new 96 well plates. One plate was used for DNA extraction while the sibling plate was kept into the incubator. DNA was extracted using Quick Extract DNA extraction solution (as described in **section 2.6 a**). Two µl of DNA extracted was used for PCR amplification with superfi 2x Platinum Master Mix with Fclon and Rclon primers. PCR product was gel-purified and sent for Sanger sequencing with Fscreening primer sitting 200 bp up stream to the locus (all PCR primer sequences are reported in **Supplementary Table 7** and **8**). TIDE was used to estimate the editing efficiency. Positive clones were transferred from 96 well plate into new 24 well plate. Once grown, cells were used for further experiments.

2.3.3 Assessment of genome editing by PCR Digestion with MfeI

The presence of the minor allele A creates a new restriction site for the MfeI enzyme (CAATTG) that is absent in the presence of the major allele G. The local sequence is C**G**ATTG, where the SNP of interest is the second nucleotide of the restriction site. For this reason, the enzymatic digestion of the PCR product obtained with primers Fclon-Rclon can be used as tool to assess successful single nucleotide editing.
Digestion of 500 ng of purified PCR product of edited cells were digested with 0.6 µl MfeI-HF (NEB) in the presence of 4 µl of Buffer Cut Smart 10x in a final volume of 40µl. Digestion was performed for 2 hours at 37°C, followed by 20 minutes at 65°C to inactivate the enzyme. Digested products were run on an agarose gel. Non edited cells (carrying the G allele) showed a band of 1603 bp while edited cells showed 2 bands of 1000 bp and 603 bp. respectively.

2.3.4 Assessment of genome editing by TOPO TA cloning of PCR products

The Sanger sequencing of PCR products encompassing our region of interest performed on pool of edited cells often showed a mix of different genomic events. To identify the sequences of the single genomic events, we used TOPO® TA Cloning® Kits (Invitrogen) for cloning of PCR products from selected clones, following manufacture's instruction. One µl of PCR product was incubated with 1µl of salt solution (1.2 M NaCl, 0.06 M MgCl2), 1 µl TOPO vector and 4 µl of H2O for 20 min in RT. Mixture was transformed in DH5-alpha bacterial competent cells as described in **section 2. 2.4.** Colonies were picked from LB-agar dishes selective for Ampicillin, grew overnight in 2 ml of LB with ampicillin and then plasmid DNA was isolated with a miniprep kit (as described in **section 2.2.5**), and sent for Sanger sequencing.

2.3.5 Assessment of genome editing by droplet digital PCR (ddPCR)

To design the appropriate primers and probes for rare event detection by ddPCR we used Primer3plus (http://primer3plus.com) so that annealing temperatures (Tm) of Forward and Reverse primers perfectly matched while being 1-3 degree lower than the Tm for the probes. Recommended length for primers/probe was between 17-25 nucleotides, optimal amplicon length between 55–100 nucleotides. In designing the allele specific probes, the following were considered: 1) WT (G allele) and mutant (A allele) probe Tms needed to be a perfect match; 2) WT (G allele) probe needed to be conjugated at the 5' with the weaker fluorophore (Hexachlorofluorescein, Hex), while the rare event detection probe for mutant (A allele) needed to be conjugated with a stronger fluorophore (6-carboxyfluorescein, FAM). Both probes were conjugated with a Black hole quencher (BHQ) in 3' to reduce the background noise; 3) Avoid repeats of Gs or Cs longer than 3 bases; 4) Probes should not have a G at the 5' end as it would result in fluorophore quenching; and 5) Design the probe to anneal to the strand that has more Gs than Cs (so the probe contains more Cs than Gs). These criteria allowed the design of these probes: FAM-BHQ1 (FAM_A): MUT TGAGAGCAATTGTGGCA 55.8°C; HEX-BHQ2 (HEX_G): WT TGAGAGCGATTGTGGCA 55.8°C.
PCR optimization and probe specificity were tested using DNA extracted from the Hs578Bst cell line that is heterozygous at the rs1376350 locus (A/G), as control. Optimal annealing temperature was determined using a c1000 Touch thermal cycler with a gradient feature to test temperatures in the 55-65 °C range.
PC-3 cells were seeded in a 96 well plate with a density from 5 to 1 cells/well. Once confluent cells were split in two new 96 well plates. One plate was used to keep the cells in culture while the other was used for DNA extraction with 40 µl of QuickExtract DNA extraction solution (as described in **section 2.2.6 a**). Two µl were used as template for ddPCR. ddPCR reaction was assembled in a final volume of 20 µL with: ddPCR SuperMix for Probes (1x, Bio-Rad), forward primer (900 nM), reverse primer (900 nM), Reference and mutated probes (HEX and FAM respectively, 250 nM). All ddPCR assays were analyzed using the QX200 droplet reader and Quantasoft software version 1.7.4 (Bio-Rad). Standard ddPCR thermal cycling conditions were used for the SNP assay, with an annealing temperature of 55.8°C.
(Program of PCR: 95°c x 10 min, Ramp 2°C/sec, 94°C x 30 sec, Ramp 2°C/sec, 58.8°C x 1 min for 40 cycles. Then 98°C x 10 min, Ramp 2°C/sec and hold 12°C).

2.4 Next Generation Sequencing (NGS) data generation and analysis

2.4.1 Libraries preparation for DNA targeted sequencing at the edited locus

For sequencing of the DNA around the locus of interest, three sets of primers were designed by using Primer3 in order to cover 500 bp around the cutting site of Cas9. Primers efficiency was tested in non edited PC-3 with Q5 high fidelity DNA polymerase (NEB) for 25 cycles (95°C x 20 sec, 61°C x 30 sec, 72°C x 30 sec). The optimized protocol was then applied on DNA of PC-3 cells edited using the Cas9_RNP technology as described in **section 2.1.4.** For each edited pool we obtained three PCR products with three sets of primers (Supplementary Table 10). After quantification with Qubit dsDNA HS (high sensitivity) Assay Kit, 4 ng were run in Caliper, which gives information of purity of PCR product. Before purification with beads, PCRs of each sample were pooled at equimolar concentration in a final volume of 50 µl. To the 50 µl mixture 40 µl (0.8x ratio) of AMPure XP paramagnetic beads (Beckman Coulter) were added on a ratio of 50:40 to remove contaminants as dNTPs, salts, primers and primer dimers. DNA bound beads was washed and then eluted in TRIS HCL 10 mM pH 8 and quantified with Qubit dsDNA HS (high sensitivity) Assay Kit. A second PCR was performed for the indexing of our PCR products (Nextera, Illumina). Selected index combinations were used with these conditions: for 10 cycles (95°C x 30 sec, 55°C x 30 sec, 72°C x 30 sec) and this mix: 0.5 ng of PCR product, 5 µl of index 1, 5 µl of index 2, 5 µl of Q5 Buffer, 1 µl Q5 polymerase, 0.4 µl of 10 µM dNTPS to a final volume of 50 µl. PCR purification with 0.8x ratio of AMPure XP beads was performed to select the desired size products. DNA was then quantified with Qubit dsDNA HS (high sensitivity) Assay Kit and pooled before MySeq 2 × 250 bp sequencing (Illumina).

2.4.2 Analysis of Targeted DNA sequencing data

Analysis of the sequencing data for the quantification of both non-homologous end joining (NHEJ) and homologous dipendent repair (HDR) occurrences in targeted CRISPR-Cas9 experiments was performed using CRISPResso (Pinello *et al.*, 2016). Briefly, low quality sequences are filtered and adapters are trimmed (SLIDINGWINDOW:4:15, MINLENGTH:36). The sequences are aligned to the reference amplicons, (3 different amplicons with the ancestral locus and the respective amplicons with the modified locus) using the CRISPRessoPooled algorithm (min identity score=60, HDR homology percentage=98%). Then, the software quantifies the fraction of HDR and NHEJ outcomes and frameshift/in frame mutations.

2.4.3 Libraries preparation for RNA-seq

For libraries preparation, RNA was quantified using Qubit RNA BR (broad-range) Assay Kit and the total RNA integrity was assessed through electrophoresis with Agilent Bioanalyzer RNA 6000 Nano Kit. All samples showed an RNA Integrity Number (RIN) ≥ 9.90. RNA libraries were prepared according to the manufacturer's protocol using Illumina TruSeq Stranded mRNA Kit. Briefly, 1 µg total RNA was diluted to a final volume 50 µL using nuclease-free ultrapure water and purified with RNA purifying beads. The RNA was subsequently fragmented and eluted on the thermocycler with the following protocol modification: 4 minutes at 94°C. The cleaved RNA fragments were subsequently reverse transcribed using Invitrogen SuperScript II Reverse Transcriptase (10,000 U, Invitrogen) in order to synthesize the first cDNA strand. The RNA template was then removed and a second cDNA strand was synthesized generating ds cDNA. After purification of the cDNA using Beckman AMPure XP Beads, the 3' ends of the blunt fragments were added with one adenine nucleotide (A-Tailing). This allowed the following ligation of the adapters containing a complementary thymine at the 3' ends. Each adapter (Illumina IDT for Illumina TruSeq RNA UD Indexes) contains a unique 6 nt sequence, an index, which was assigned to each sample. The ligation process occurred at 30°C for 10 minutes. After ligation, the fragments underwent two clean-up rounds using AMPure XP beads. A following enrichment step was performed using polymerase chain reaction: a primer cocktail annealing to adapters' ends allowed the selective amplification of those fragments having adapter molecules at both ends. The PCR was optimized to 12 cycles. A final clean-up step with AMPure XP beads allowed the wash away of unwanted fragments. The library was finally quantified using Qubit dsDNA HS (high sensitivity, Thermo Fisher Scientific) Assay Kit and delivered to the NGS facility for the following quality controls. Single end (100 bp) sequencing was performed on a HiSeq 2500 (Illumina).

2.4.4 Analysis of RNA-seq data

Pre/processing and differential analysis Analysis of RNA sequencing data was performed as follows: quality of raw data was controlled using fastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and visualized using MultiQC (Ewels *et al.*, 2016). Trimmomatic (Bolger, Lohse and Usadel, 2014) was used to trim adapters from sequences, to filter out low quality sequences (<15 in a sliding window of 4bp) and sequences that resulted too short after the trimming (<50bp). Trimmed and filtered fastQ files were aligned to the hg38 reference human genome using STAR (Dobin *et al.*, 2012). Read counts were generated through the STAR quantMode function and rearranged in a single expression matrix through a custom R script. Differential expression analysis was performed using edgeR (Robinson, McCarthy and Smyth, 2009), that includes an internal normalization step to generate the pseudo counts used in the actual differential expression analysis. Differentially expressed genes (DEGs) were filtered based on False Discovery Rate (FDR) (<0.05). Normalized counts per million reads (cpm) were used exclusively for visualization purposes. Previously published RNA-seq data were pre-processed as reported in (Romanel *et al.*, 2017) and here analysed for differential expression analysis using edgeR (starting from published read counts data) for consistency purposes.

False positives biological effects filtering. To control for biologically different control samples and biological replicates of the edited cells, a step of differential expression analysis between the control samples was performed to exclude putative false positive results. By default, edgeR uses the most complex dispersion model for each transcript in the expression matrix. As more replicates were present for the treated cell lines with respect to the controls, a common dispersion (i.e. squared coefficient of variation) was used instead of the default to avoid overestimation of DEGs in control vs control analyses. Specifically, we used 0.2 in the control vs control comparisons and the automatic dispersion model detection in the control vs treated comparisons (McCarthy, Chen and Smyth, 2012). DEGs from control vs control

comparisons were blacklisted and subjected to functional and upstream prediction analyses to nominate biological effects related to the variability in the control samples.

Functional and upstream regulation analysis. Functional characterization of DEGs lists was carried out using clusterProfiler (Yu *et al.*, 2012). For the prediction of upstream regulators, we used QuaternaryProd (Franceschini *et al.*, 2012; Fakhry *et al.*, 2016) that for a given a set of DEGs computes the significance of upstream regulators in the network by performing causal reasoning using statistical scores. We applied the Ternary Dot product Scoring Statistic (Ternary Statistic) that exploits signed and unambiguous edges in the STRING database (Franceschini *et al.*, 2012) to build the regulatory network.

2.4.5 Gene lists

All gene lists used throughout the study, including DNA repair genes (N=180) and hormone regulated genes (N=330), are detailed in (Romanel *et al.*, 2017). Briefly, from the Human DNA Repair Genes database and an additional curated list (http:// sciencepark.mdanderson.org/labs/wood/dna_repair_genes.html, **Supplementary Data 1** in (Romanel *et al.*, 2017). The list of hormone-regulated genes was obtained from LNCaP cells treated with small interfering RNA (siRNA) targeting AR in (Wang *et al.*, 2009). Transcripts associated with the non-coding polymorphic regulatory element at 7p14.3 in human data are obtained as in (Romanel *et al.*, 2017) upon alignment to hg38.

2.4.6 Human genotype and transcript data

The human data originally used for the identification of the association between the polymorphic locus and the SPOP mutant phenotype are detailed in (Romanel et al., 2017). Here we used rs1376350 genotypes and transcripts of prostate cancer patients specifically from N = 63 benign prostate tissues and N = 319 prostate cancer tissues.

2.4.7 Code and data manipulation of sequencing

Statistical analyses, data manipulation and data visualization were performed using R (R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.). FastQ and alignment files were manipulated using samtools (Li *et al.*, 2009). All the analyses were implemented using custom shell scripts by Dr. Yari Ciani, a computational post/doc fellow in the Laboratory of Computational and Functional Oncology, CIBIO Department, University of Trento.

3 Results

Based on the study specific background, my PhD work evolved through four major sets of experiments, all aimed at disrupting or altering the locus of interest in prostate cancer cells via CRISPR/Cas9 editing, followed by the characterization of the transcriptomes of edited pool or single clones (**Supplementary Figure 2, Table 3** and **Supplementary Figure 7, Table 4**). Specifically, the initial choice was to induce a large deletion spanning the locus (731 bp, referred to as macrodeletion) to first confirm in pooled cells whether an overall transcriptional deregulation would follow. Subsequently, as part of this first approach, I also screened for clones positive for the macrodeletion. The second approach consisted of the induction of a 50 bp deletion spanning the locus (referred to as microdeletion). The third and the fourth approach regarded a more granular editing of the sequence of interest; the disruption of two TF motifs recognized through *in silico* prediction on and around the locus (AR and CEBPβ) and finally the single nucleotide editing to create isogenic cells harboring the heterozygous (G/A) or the homozygous genotype (A/A) for the minor allele at rs1376350.

The choice of the experimental model to use in my project was crucial. Provided that all the available PCa cell lines (about 10) have homozygous genotype for the reference allele at the locus of interest and none harbor the characteristic *SPOP* mutation, we reasoned in terms of the following aspects: 1) the background experiments were all performed in PC-3 cells and, importantly, experiments testing allele specific enhancer activity in a Luciferase assay showed higher signal for the risk allele, suggesting this cell line as a good candidate for my challenging project; 2) given the observation that both AR and CEBPβ would bind on and around the variant, we reasoned that being able to control the most variables would be beneficial; PC-3 cells are almost AR negative and only express CEBPβ but not CEBPα, features not shared with any other PCa cell line. Therefore, despite recognizing that the best model would be of benign prostate cells as representative of the state where we hypothesized the variant would exert its main role, we opted for the PC-3 cells as experimental model.

3.1 The 7p14.3 locus is functional based on macrodeletion of 731 bp in PC-3

3.1.1 Setting up condition of transfection in PC-3

The majority of the protocols involving Cas9 system has been set up in HEK293T cells (derivative of human embryonic kidney 293 cells and containing the SV40 T-antigen) as they are highly transfectable. Of note, the delivery of the editing system could be cumbersome as the Cas9 cassette (4 kb) is inserted in large plasmids of more than 6 kb of size, overall resulting in very large vectors for transient transfection. In order to successfully apply CRISPR/Cas9 in prostate cells, I first tested four transfection reagents known to be compatible with our prostate cells using two plasmids with different sizes, each harboring the EGFP expression cassette: pEGFPN1 (4.7 kb) and pHR_SIN_CSGW (9.7 kb). Tests were run in PC-3 cells for each transfection reagents and each vector. Percentage of EGFP positive cells, calculated 72 h after transfection using Tali™ Image-Based Cytometer, showed higher efficiency for the FuGENE HD (Promega) compared to other transfection reagents. The analysis indicated that almost 90% of cells expressed EGFP (green) when transfected with pEGFPN1 (4.7 kb) compared to 40% when transfected with pHR_SIN_CSGW (9.7kb). The PC-3 cells results are in line with the literature (Hornstein *et al.*, 2016), where the larger the size of the plasmid carrying the Cas9 the lower is the transfection and the editing efficiency by Cas9 enzyme (**Figure 4**).

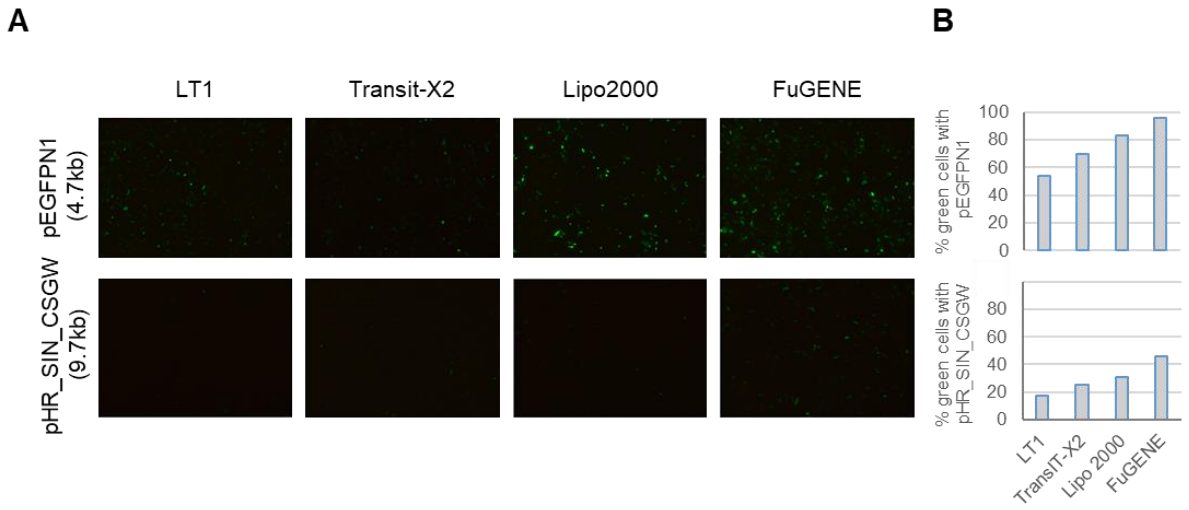


Figure 4. **Testing transfection efficiency in PC-3**. Four transfection reagents were tested using two plasmids with different size, both harboring the EGFP sequence. A) PC-3 were transfected with pEGFPN1 (4.7 kb) (upper panel) and pHR_SIN_CSGW (9.7 kb) (lower panel). B) Quantification of EGFP expressing cells 72 h post transfection with pEGFPN (4.7 kb) (upper plot) and pHR_SIN_CSGW (9.7 kb) (lower plot). Images and analysis were produced with Tali™ Image-Based Cytometer.

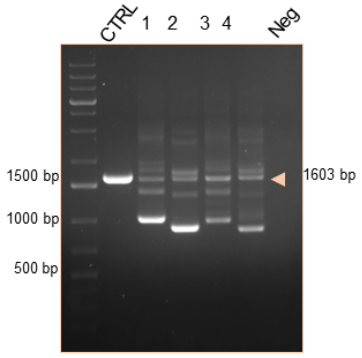
3.1.2 Deletion of 731 bp around the locus with Cas9 endonuclease in PC-3

Our first approach to investigate the functionality of the locus was to induce a macrodeletion of more than 600 bp surrounding the 7p14.3 variant. Once transfection conditions were optimized, we tested four combinations of sgRNAs

A

Table 2. Combinations of sgRNAs used for deletion of the 7p14.3

	Combination	Expected size (bp)	Deleted size (bp)
1	Cas9+1st combination	978	625
2	Cas9+2nd combination	872	731
3	Cas9+3rd combination	955	648
4	Cas9+4th combination	848	755
CTRL	Not transfected	1603	0



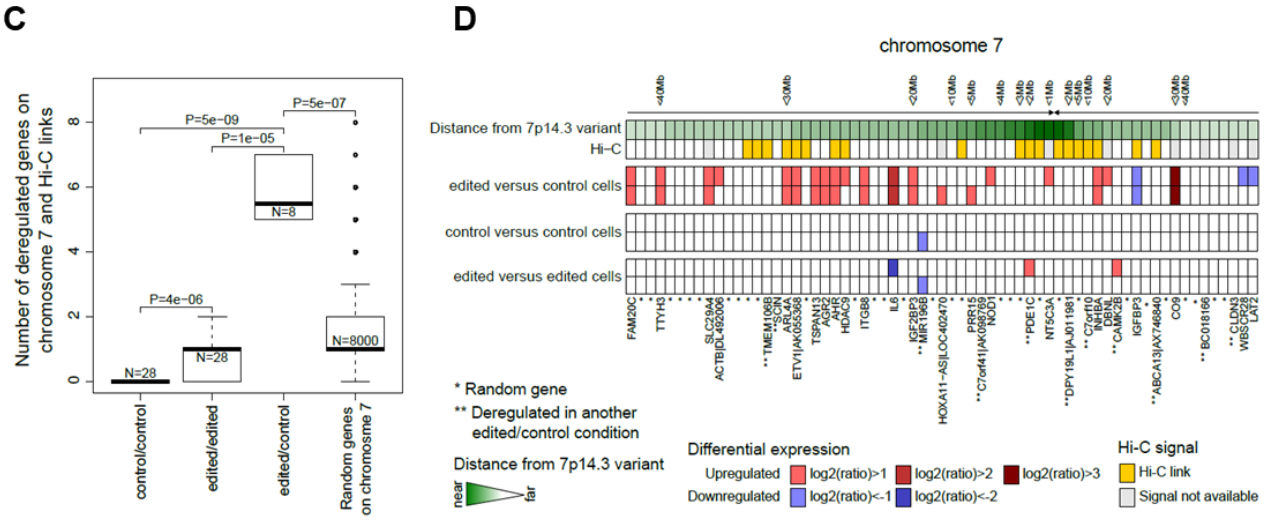


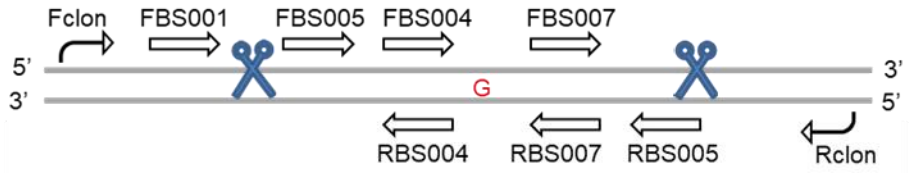
Figure 5. Detection of deletion and RNA-seq analysis of pool of edited PC-3 cells. A) Table 2 lists: in the first column the four combinations of sgRNAs upstream and downstream the locus, tested with eSpCas9(1.1); in the second column the expected sizes of PCR amplicons for each combination and in the third column the corresponding sizes of the deleted portion of DNA surrounding the locus variant. B) PCR products for each combination of sgRNAs with eSpCas9(1.1), in pool of edited cells, were analyzed in 1% agarose gel. C) Deregulation of transcripts on chromosome 7 with respect to prostate cells Hi-C identified links. Enrichment is shown by comparing the level of deregulation in edited *vs.* control cells, in edited *vs.* edited, and in control *vs.* control cells. Further, enrichment is shown by comparing the level of deregulation in edited *vs.* control cells with simulated data computed by generating, for each tested combination, 1000 random selections of genes at chromosome 7 with size equal to the observed deregulated set. P-values are computed using Mann–Whitney test. D) Visual representation of deregulation patterns in edited *vs.* control cells at chromosome 7 within a 40 Mb window around the 7p14.3 variant. Representative experimental conditions of edited *vs.* control cells are shown and random combinations of edited *vs.* edited and control *vs.* control cells are shown (Romanel *et al.*, 2017).

(**Figure 5A, Table 2**) either with wt SpCas9 or with enhanced specificity variant eSpCas9(1.1). eSpCas9(1.1) is a variant of wt Cas9 developed with the aim to reduce off targets events (Slaymaker *et al.*, 2016). Since the efficiency of editing analyzed with PCR revealed a similar profile of editing for both Cas9 (data not shown), the PC-3 cells were transfected with each combination of sgRNAs and eSpCas9(1.1) to minimize off targets. After selection with Puromycin, DNA was extracted and analyzed by PCR. Combinations 1 and 2 resulted in higher editing efficiency compared to combinations 3 and 4 (**Figure 5B**). RNA-seq analysis of pooled cells demonstrated an enrichment of transcriptome deregulation in edited *vs.* non-edited cells (control cells) when compared to the same analysis in edited *vs.* edited and control *vs.* control (P = 1e-04 and P 3e-05 Mann–Whitney test, respectively). The fractions of deregulated genes that showed upregulation resulted on average in 70%, 65%, and 62%, respectively. Selected transcripts were validated by real-time. A focused analysis along chromosome 7 in edited *vs.* control cells, but not edited *vs.* edited or control *vs.* control, showed significant concordance with genes predicted to physically interact with the 7p14.3 locus by previously generated Hi-C chromosome conformation capture data from benign prostate cells (Rickman *et al.*, 2012) (**Figure 5C** and **D)**. These data represent my contribution to the manuscript Romanel A, Garritano S, et al, Nature Communications 2017 (Romanel *et al.*, 2017).

3.1.3 Screening of clones positive for deletion

To better characterize the signal of the transcriptome deregulation due to the excision of the locus, we decided to interrogate the transcriptome of single clones positive for the deletion. From the bulk of edited cells obtained with combination 2 (**Figure 5, Table 2**), single clone dilution was performed in order to have one cell per well in a 96 well plate. To select clones positive for the deletion of the locus we employed PCR analysis using six pairs of primers annealing outside (Fclon-Rclon), inside (FBS005-RBS005; FBS004-RBS007) or one inside and the other outside (FBS001-RBS007; RBS004-Rclon) the deleted region (**Figure 6A** and **B, Table 3**). Of the 45 clones screened in this way, 10 were homozygous for the deletion of 7p14.3 locus variant since no amplification of PCR occurred with pairs of primers annealing only inside (FBS005-RBS005; FBS004-RBS007) or inside and outside (FBS001-RBS007; RBS004-Rclon) the locus. However, in these 10 clones, amplification with primers outside the intended deleted genomic region (Fclon-Rclon) gave two bands of different size (1000 and 872 bp) suggesting that the repair after editing was different in the two alleles. The results were further confirmed with Sanger sequencing showing that one allele presented a re-end joining with small indels, while the other allele had an incomplete deletion (140 bp segment was retained 587 bp downstream of sgRNA_*up* (2) and 55 bp upstream of sgRNA_*down* (3)). The profile of incomplete deletion is shared by clone 16, 29 and 32 (**Figure 6C** and **D)** (profiles of other 7 clones are not shown). The remaining 35 clones demonstrated deletion of one allele as confirmed by a 872 bp PCR product obtained with primers annealing outside the deletion (Fclon-Rclon) while the second allele presented either a translocation or an inversions. Clone 8 is an example of the translocation event as suggested by the absence of amplification with a pair of primers annealing one outside and the other inside the deletion (FBS001-RBS007) and successful amplifications with primers annealing inside the deletion (FBS005-RBS005; FBS004-RBS007) with primers annealing one inside and the other outside the deletion (FBS004-Rclon) (**Figure 6A** and **B, Table 3**). Clone 26 is an example of inversion because no amplification was detected with pairs of primers annealing outside and inside the locus (FBS001-RBS007) whereas amplification with both reverse primers (RBS004-Rclon) and primers annealing inside the locus (FBS005-RBS005; FBS004-RBS007) were successful (**Figure 6C**). The profile of inversion was detected in other clones (data not shown) with the same procedure as for clone 26.

A

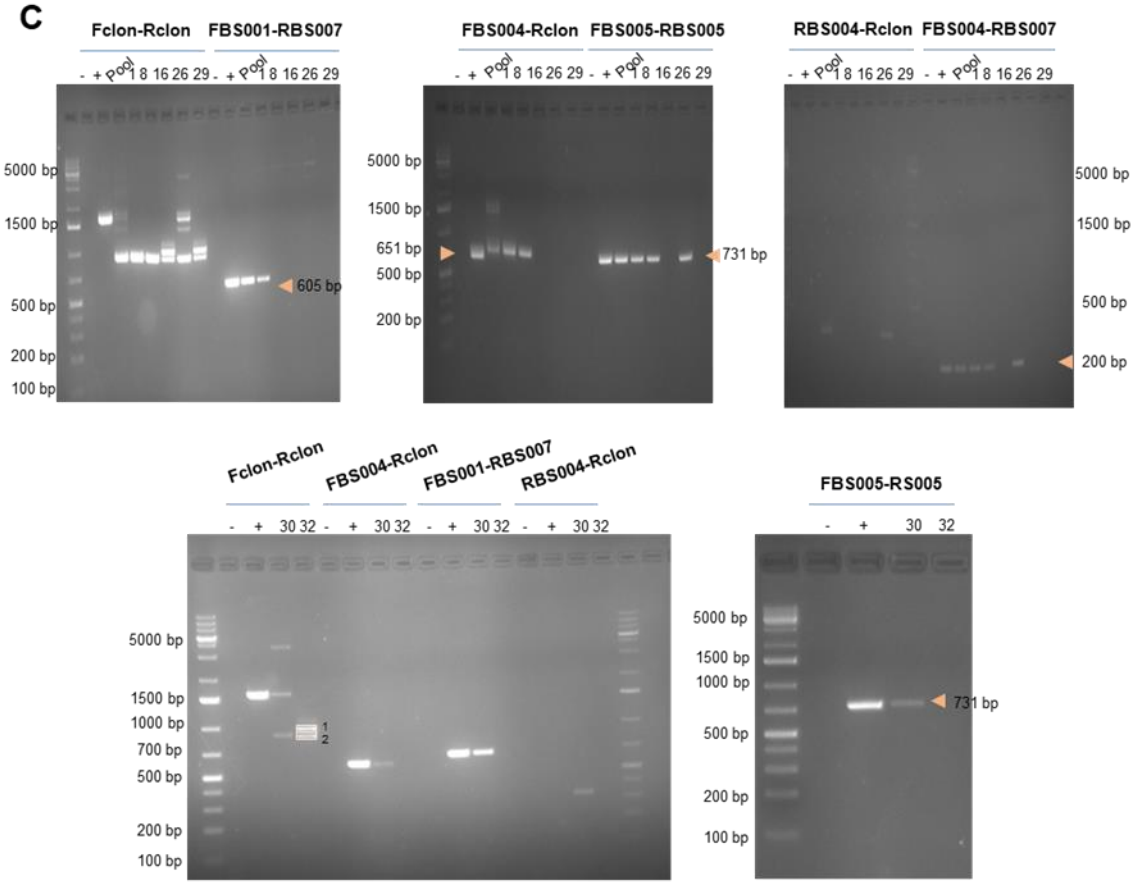


B

Table 3. Combination of primers and PCR screening of clones positive for 731 bp deletion

Combination of primers	Expected size of PCR product (bp)		Observed size of PCR products (bp)			
	Wt	Del/Del	Del/Del	Del/Invertion	Del/Transloc	Total
Fclon-Rclon	1603	872	1000/872	872/1603	872	
FBS001-RBS007	605	-	-	-	-	
FBS004-RBS007	200	-	-	-	200	
FBS005-RBS005	731	-	-	-	731	
FBS004-Rclon	651	-	-	-	651	
RBS004-Rclon	-	-	-	651	-	
Number of clones			10	25	10	45

C



D

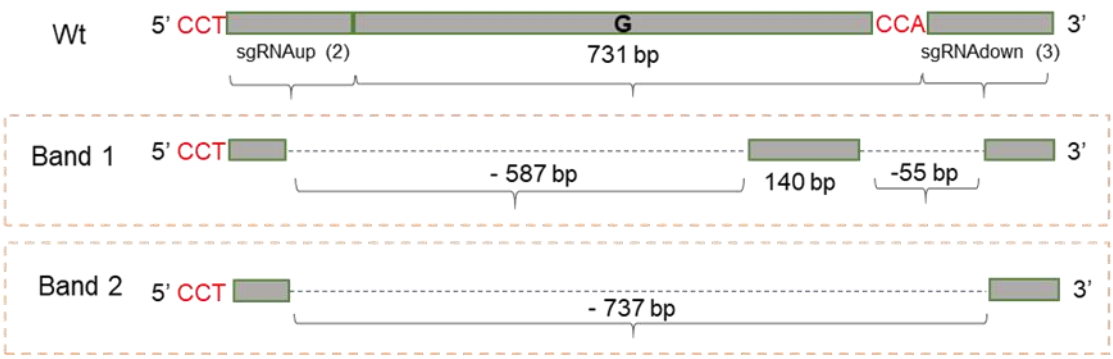


Figure 6. Screening for clones positive for macrodeletion of the locus via PCR and Sanger sequencing. A) Clones were screened for the deletion of 731 bp with PCR using the six pairs of illustrated primers; primers outside the intended deleted genomic region (Fclon, FBS001 and Rclon), pairs flanking the precise breakpoints of the intended deletion from both *up* and *down* sgRNAs (FBS005-RBS005), and primers inside the intended deletion (BS004 and RBS007). B) Table 3 shows the size of each product of PCR in non edited and in edited PC-3 cells. The first two columns list the predicted amplification sizes if no deletion occurred in PC-3 transfected with sgRNA_scramb (+) (column 1) or if deletion occurred in both alleles (column 2). The next columns report the events observed by PCR analysis with each pair of primers. Combination of primers outside and inside the deleted locus (FBS001-RBS007; FBS004-Rclon) gives information on deletion and translocation events. RBS004-Rcloning pair suggests possible inversion of the deleted portion of genome. C) PCR product of PC-3 transfected with sgRNA_scramb (+), pool of edited cells from second combination of sgRNAs and single clones were run in 1% agarose gel. Arrows indicates bands observed in PCR for those clones positive for each event described above; not all single clone PCR products were sequenced (exact boundaries of deletions not known for some clones). D) Band 1 and 2 of PCR product (Fclon-Rclon) of clone 32 were purified from agarose gel and sequenced. The schema represents the sequence for both alleles compared to the wt. In red PAMs for both sgRNAs are shown, dash lines indicate the deleted portion of DNA. (-) corresponds to mix of PCR without template.

Different patterns of translocation and inversion, due to the activity of Cas9 endonuclease, have been recently observed by other members of our group in similar experiments in different cell lines. Based on the 45 screened clones, no evidence of heterozygous deleted cells emerged (the second allele always showed evidence of structural rearrangement), suggesting that these combinations of sgRNAs with eSpCas9(1.1) endonuclease worked with very high efficiency. Indeed, the intensity of the lower band was higher in pool of edited cells, supporting the idea that in almost all cells one allele was deleted and repaired with indels, while the other allele presented different rearrangements. Clones with deletions spanning the SNP on both alleles (such as for clone 32) were used for the analysis of the transcriptome. Results are presented in **section 3.4.**

3.2 Fine-tune editing of the 7p14.3 functional locus

3.2.1 Microdeletion of 50 bp surrounding 7p14.3 locus in PC-3

To further investigate the implication of the locus in terms of deregulation we decided to delete 50 bp spanning the 7p14.3 locus in PC-3 with Cas9 endonuclease and sgRNA_A/E (**Figure 7A**). For this purpose, we followed the same procedure performed for the macrodeletion of 731 bp. PC-3 cells were co-transfected with peSpCas9(1.1), pUC19_sgRNA_A and sgRNA_E with FuGENE HD (Promega) and selected for 5 days with Puromycin (2 µg/ml).

Figure 7. Microdeletion in PC-3. A) Design of sgRNAs based on PAM sequences surrounding the locus. Blue indicates the sequence predicted to be deleted with sgRNA_A and E; red indicates the ancestral allele G of rs1376350. Coordinates of genome sequence are from GRCh37/hg19, consistently to the macrodeletion work previously shown. B) Deletion of approximately 50 bp in PC-3 were analyzed in 2% agarose gel. Sample 1 represents non transfected PC-3 cells, sample 2 corresponds to PC-3 transfected with peSpCas9(1.1), pUC19_sgRNA_A and E. Pool of edited sample (2) show evidence of a lower band corresponding to the deletion of ~50 bp, a middle band corresponding to not edited cells and an upper band potentially corresponding to rearrangements. (-) corresponds to mix of PCR without DNA template.

Editing efficiency was assessed with PCR amplification analyzed in 2% agarose gel (**Figure 7B**). We were interested in comparing the transcriptomes of pools of PC-3 cells edited with eSpCas9(1.1) and sgRNAs for the macrodeletion (731 bp) and the microdeletion (50 bp) in order to identify similarities and differences.

3.2.2 Disruption of AR and CEBPβ Motifs around the polymorphism in PC-3 with CRISPR/Cas9

3.2.2.1 Cutting efficiency of three sgRNAs in PC-3 cells

We designed sgRNAs nearby the 7p14.3 locus variant, taking into account PAM sequences and types of Cas9 endonucleases to be combined in order to achieve the highest efficiency of editing. There are three accessible PAMs around the rs1376350 locus to be used with three Cas9 endonucleases, from left to right. sgRNA_B was designed including a G nucleotide to the starting site of 5' sequence in order to increase transcription efficiency by U6 promoter. For the same reason, sgRNA_C was designed with 21 bp by adding a G in 5' (no matching with DNA sequence). Fifteen nucleotides downstream the germline variant a PAM sequence allowed for the design of another sgRNA with 20 bp (including a G nucleotide in 5' of guide sequence), sgRNA_A (**Figure 8A).**

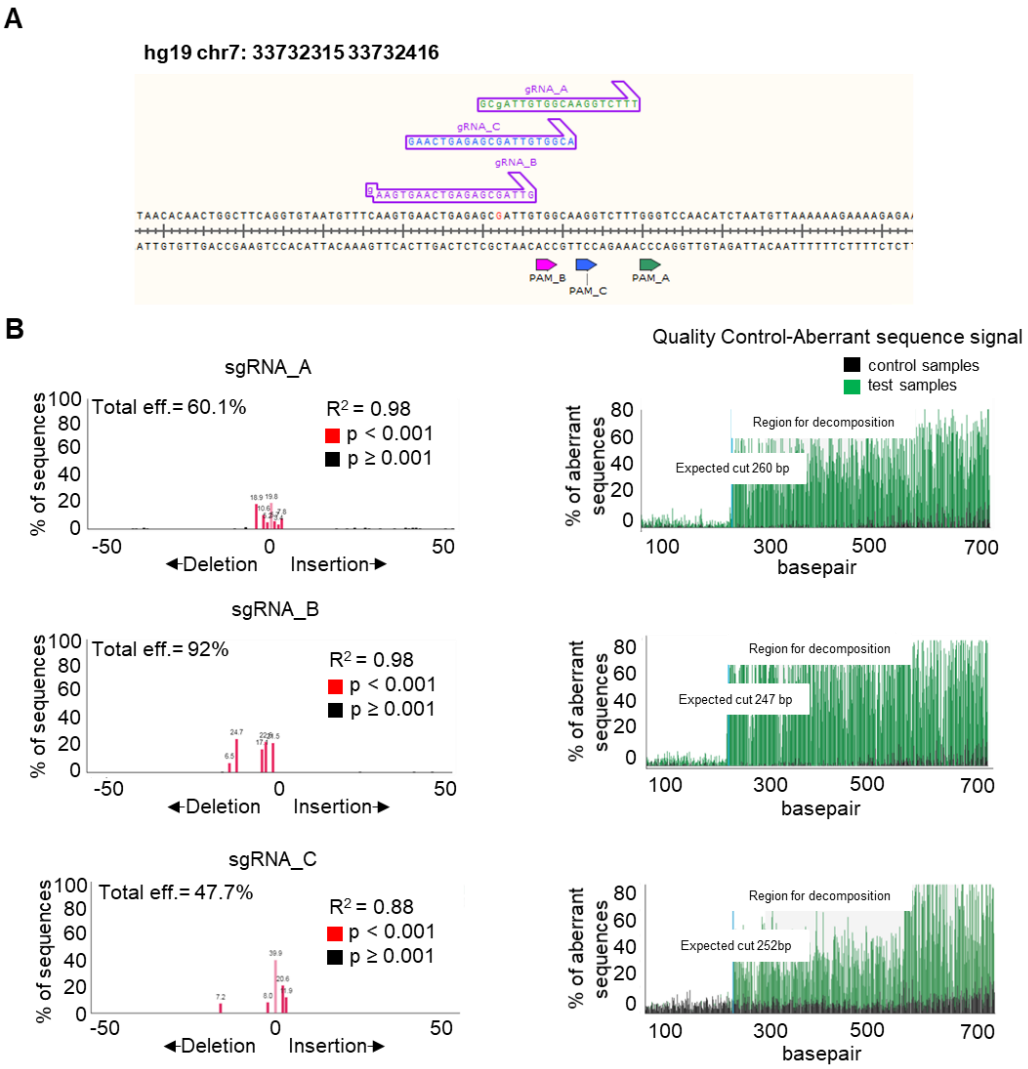


Figure 8. **Testing cutting efficiency of wt Cas9 with three sgRNAs in PC-3**. A) Design of sgRNA A, B and C nearby the locus variant based on three PAMs indicated in the scheme. B) TIDE quantification of cutting efficiency with sgRNA_A, sgRNA_B and sgRNA_C. On the left, graphic representations of possible insertions and deletions due to editing. Red and black bars indicate, respectively, statistically significant and non significant events. On the right, corresponding graphic representations of aberrant nucleotide signal of the sample (green) compared to that of the control (black).

Unlike the other two sgRNAs close to the SNP, sgRNA_A is permissive to be used with the high fidelity evoCas9 (Casini *et al.*, 2018) that does not tolerate guides longer than 20 bp or containing mismatched nucleotides at the 5'of the spacer sequence, similar to other high fidelity Cas9 variants. These requirements do not apply to wt SpCas9, which has been shown to be more flexible in terms of length and mismatches (Casini *et al.*, 2018). Wt Cas9, SpCas9, and evoCas9 were used in the reporter system (see **section 3.3.1**). The TIDE software (Brinkman *et al.*, 2014) was used to predict the efficiency of editing within a 50 bp range to query for indels created as a response to the Cas9 activity. TIDE results showed that all sgRNAs with wt Cas9 cut at least around 50% of the times (**Figure 8B**).

3.2.2.2 Disruption of Motifs for AR and CEBPβ in PC-3

In silico prediction and ChIP-qPCR experiments performed in PC-3 (G/G) ((Romanel *et al.*, 2017) and additional motifs search analyses) suggested that two TFs binding sites, one for AR and one for CEBPβ reside nearby the SNP. Based on the overall transcriptomic results from the cells edited with the macrodeletion, we were interested in understanding if the disruption of the motifs for AR and/or CEBPβ can also alter the transcriptome. In an effort to characterize the locus we decided to disrupt the binding motifs for AR and CEBPβ by using CRISPR-Cas9 system. PC-3 expressing stable Cas9 were transfected with pGSsgRNA_B and C, conferring resistance to Neomycine, and with px330sgRNA_A conferring resistance to Puromycin. The rationale was to take advantages of the editing profile, predicted from TIDE, induced by these two sgRNAs to achieve a partial or complete disruption of AR (**Figure 9A**) and/or CEBPβ motifs (**Figure 9B**). The right column of Figure 9 depicts possible scenarios of resulting edits on AR and on CEBPβ motifs compatible with TIDE results (**section 2.3.1**). The pool of the transfected cells was then diluted and plated in 96 well plate to obtain single cell clones to further test them by PCR and Sanger sequencing.

Of all possible indels events, we were interested in the ones spanning more than 7 bp because they would likely disrupt the entire motif. Since each allele has a different profile of repair, Sanger sequences were difficult to interpret. To bypass this issue, DNA from clones predicted with TIDE software to have a deletion of more than 7 bp were amplified with PCR (FBS005-RNS005), cloned into TOPO-TA vector and transformed into bacterial cells. This allowed the analysis of the single editing event separately, as each bacterial colony carried only one amplicon-carrying plasmid (**Figure 9C**). Ten bacterial colonies for each clone were picked and sent for sequencing. Giving that overall the two alleles are repaired in a different manner after cleavage by Cas9, it was difficult to find clones positive for disruption of both motifs in both alleles. Indeed, of the 70 screened clones only one fitted our requirements (**Figure 9D** and **E, Table 4**). Clone B4 obtained with editing via sgRNA_B presented the deletion of 14 and 24 bp, as reported in TIDE analysis. The list of clones selected and screened with the TOPO-TA system is presented in **Table 4** (**Figure 9E).** We selected clones positive for the disruption of either AR or CEBPβ in one or both allele and separately, in different clones or together (in the same clone) and analyzed their transcriptomes.

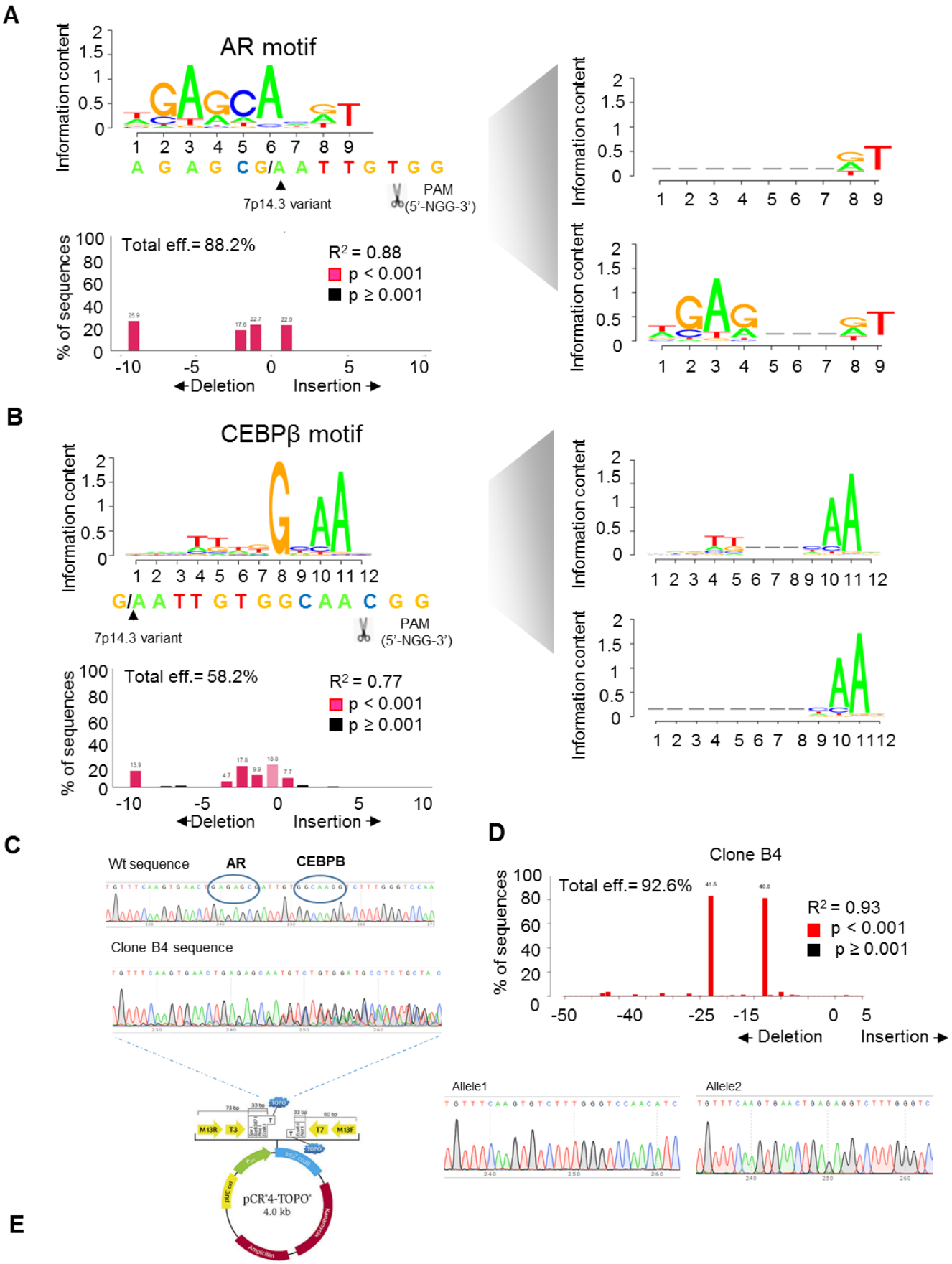


Table 4. Profiles of clones positive for editing of TFs binding sites

allele 1 allele 2	AR/CEBPβ AR/CEBPβ	-/CEBPβ AR/CEBPβ	-/- CEBPβ/-	CEBPβ/- CEBPβ/-	-/AR CEBPβ/-	-/AR -/AR	-/- -/-	Total
Number of clones	49	9	3	0	4	3	1	70

Figure 9. Schematic representations of disruption motifs with CRISPR-Cas9 system and screening for clones positive for disruption motifs. A) Disruption motif for AR was induced by transfecting PC-3_LvCas9 with pGSsgRNA_B. TIDE inference of indels created by Cas9 to predict possible scenarios as partial or complete disruption of the motif (right panel). B) Disruption motif for CEBPβ was induced by transfecting PC-3_LvCas9 with pGSsgRNA_C. As in the previous scenario, partial or complete deletion of the locus allow for the disruption of the CEBPβ motif (right panel). C) PCR product of clones selected from TIDE prediction were cloned into a TOPO-TA vector. Plasmid extracted from bacterial colonies were sent for sequencing and profiled alleles were analyzed. D) Clone B4 showed disruption of both motifs in both alleles. E) List of clones screened in this series of experiments.

3.3 Single nucleotide editing

Although recent advances in genome-editing are offering new approaches that may increase genetic studies, substitutions of single nucleotides relevant to inherited variants and to acquired point mutations still remain hard to achieve.

When this project started in 2015, insertion of a donor DNA *via* HR was the most plausible strategy for single nucleotide editing. One year later Komor et al. (Komor *et al.,* 2016) introduced BE as a new approach of genome editing that enables the direct, irreversible conversion of one target DNA base into another in a programmable manner, without requiring dsDNA backbone cleavage or donor template. The technique ensured high editing efficiency with low percentage of indels but the activity of BE was limited within a window of approximately 9 nucleotides. The low specificity of the system (able to modify more than one nucleotide) and/or lack of PAMs availability nearby the 7p14.3 locus (see **section 1.4**) forced us to follow the classic approach of single base editing *via* HR. In 2018 new variants of BE with more permissible PAM sequences and restrict window of activity from 2 to 4 nucleotides were reengineered. This prompted us to use both editing strategies in parallel. We obtained isogenic cell lines at 7p14.3 locus *via* HR before testing the activity of APOBEC1-xCas9 (D10A) in our cells. In this section we describe the set up protocol for single nucleotide editing via CRISPR/Cas9 and HR and describe few technical issues we faced during the development of the project.

3.3.1 Reporter system to assess the efficiency of editing of Cas9 in presence of allele A or G

One limitation to successful substitution by HDR is the possible re-cutting of the locus by Cas9. To test the discriminatory efficiency of a single nucleotide by Cas9 endonuclease, we designed a reporter system made by the pcDNA3 vector harboring two stop codons and an EGFP cassette preceded by the P2A peptide sequence. Forty bp of the genomic sequence, including the rs1376350 locus, with the ancestral allele G or with the minor allele A were cloned inside the reporter vector (**Figure 10A**). Once verified for the insertion of the genomic DNA, pcDNA3 together with wt SpCas9 and pUC19_sgRNAs were transfected in HEK293T cells. Briefly, the reporter system works as follow: i) in the presence of the G allele, the sgRNA guided Cas9 recognizes and cuts the sequence, the stop codon is bypassed as result of the repair mechanism and the EGFP cassette is expressed. This results in the detection of EGFP protein by western blot; ii) if the presence of the A allele prevents Cas9 from recognizing the locus, the stop codon is read and, in turn, no EGFP protein is detected by western blot. The reporter system was tested for three Cas9 and three sgRNAs (**Figure 10B**). Specifically, wt Cas9 was tested with three sgRNAs, eSpCas9(1.1) was tested with sgRNA_B and sgRNA_C, and evoCas9 was tested with sgRNA_A only. The results showed that the best discrimination between the two alleles occurs with peSpCas9(1.1) in combination with sgRNA_C (**Figure 10B**). We don't know why in this experiment eSpCas9(1.1) was not working with sgRNA_B since there was no evidence of EGFP protein in presence of both alleles. The EvoCas9 data showed lower efficiency, whereas the wt Cas9 demonstrated higher performance but lower discrimination. Therefore, we decided to proceed with eSpCas9(1.1) in combination with sgRNA_C in the following experiments aiming for single nucleotide editing.

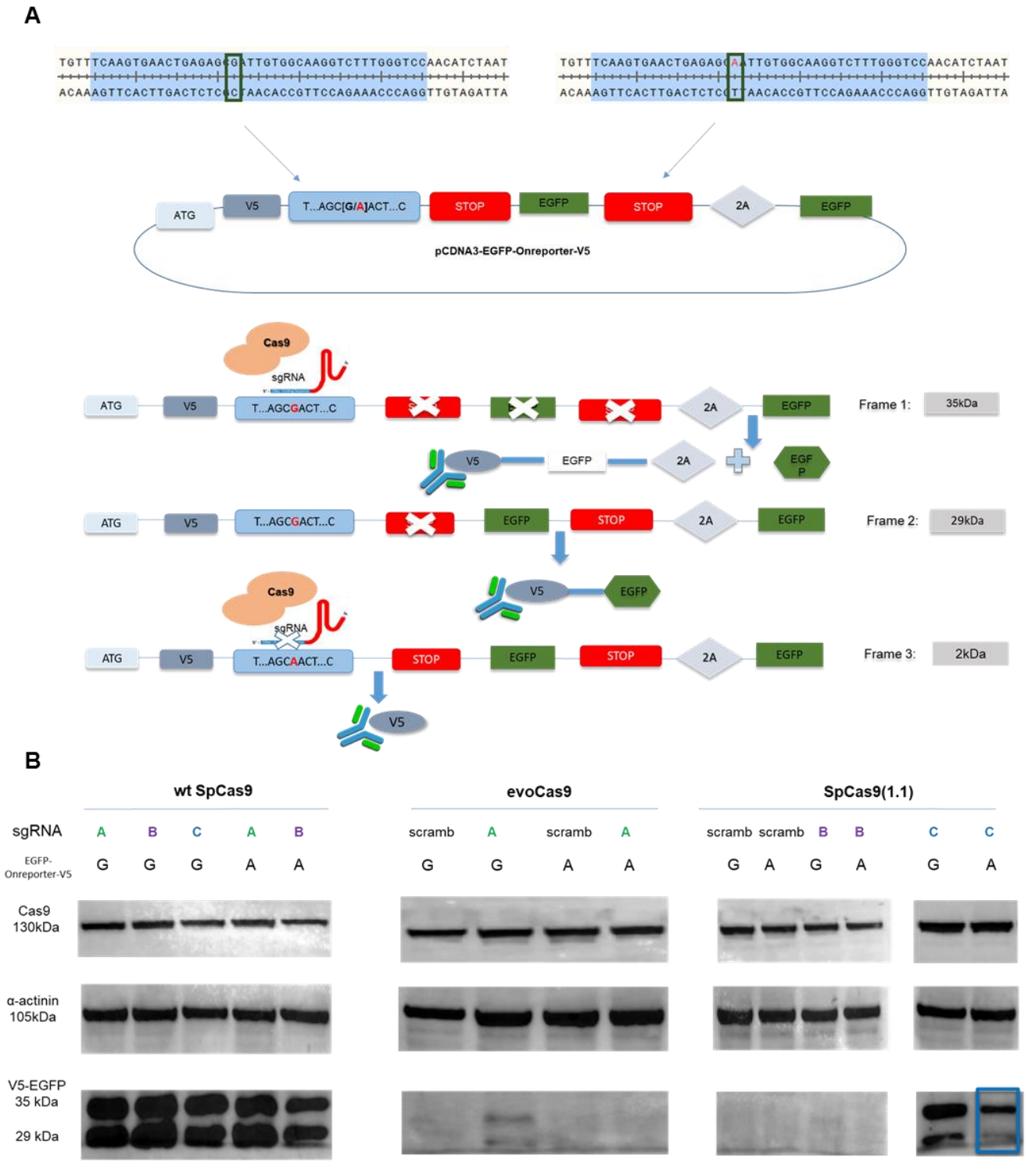


Figure 10. Reporter system to assess the efficiency of editing of sgRNAs and Cas9 in presence of allele A or G. A) Two sequences of 40 bp with allele G or A were cloned into a report system constructed as reported in the cartoon. HEK293T were transfected with pCDNA3, plasmid harboring Cas9 and sgRNAs. The report system allowed for the discrimination of events of editing and/or no editing based on the expression or not of EGFP protein, detectable with western Blot. B) Wt Cas9 was tested with three sgRNAs for allele G and two sgRNAs for allele A (last sample was lost during preparation). EvoCas9 was tested with sgRNA_A, while eSpCas9(1.1) was tested with sgRNA_B and C. Experiment was run only once in HEK293T cells.

3.3.2 Editing of single nucleotide with CRISPR/Cas9 system: strategies for screening

Isogenic cell lines with the rs1376350 minor allele are crucial for the functional characterization of the locus of interest and for this study. Although CRISPR/Cas9 system remains the most powerful tool in our hands, inducing single nucleotides substitutions via HDR was challenging due to low frequency of such precise and sophisticate repair mechanism. For this reason, we considered and tested a series of screening strategies summarized below (**Supplementary Figure 3A**). Furthermore, an appropriate strategy for accelerating the screening procedure was considered during the design of the experiment.

1. We observed that the minor allele at the locus of interest creates a MfeI digestion site, upon HDR editing (**Supplementary Figure 3A**). To verify if this could be exploited, we used pgl14 plasmid harboring 1,603 bp of the genomic region including our locus with either the A or the G allele. One µg of plasmid was digested with MfeI and run in agarose gel. Plasmid harboring the allele G was not digested by MfeI while the pgl14_A was digested as expected (data not shown).
2. We also considered the ddPCR technique which is a highly specific tool for the detection of rare events (Miyaoka *et al.*, 2014; Findlay *et al.*, 2016). The *in house* BioRad QX100 ddPCR reader offers two unique channels for monitoring fluorescence emission intensities (channel 1 for FAM and channel 2 for HEX labeled probes). We therefore designed a common primer pair and two allele-specific TaqMan probes conjugated with different fluorophores and co-respective quenchers following the manufacture's guidelines. Probes differ from each other only for the presence of minor (FAM_A) and major allele (HEX_G) (sequences of probes in **Supplementary Table 9**). As the primers were designed outside and inside the donor DNA, the detection with the corresponding probes was specific to the genomic region (with no amplification of the donor sequence) for either allele. However, this system could not detect NHEJ events and therefore the fraction of positive events for allele A is not representative of the events that occur into cells after editing. We first optimized the PCR annealing and extension temperature by using DNA extracted from Hs 578Bst (ATCC® HTB-125™), cells with heterozygous genotype (A/G) at the locus of interest. A gradient for the annealing temperature was decided from 65°C to 55°C. The optimal annealing temperature is the one that results in the largest fluorescence amplitude differences between the positives and the negatives droplets and that avoids non-specific amplification. Both FAM_A and HEX_G presented a good separation of fluorescent signal (green and blue) from background (black) starting from 58.8°C (**Supplementary Figure 3B**); this separation was consistent with the lowest temperature, although rainy patterns were evident from well F to H suggesting not specific events that may occur with lower temp of annealing.
3. Last, to accelerate the screening in the context of the low frequency of HDR events, we decided to follow the protocol of Mijaoka et al., published in Nature Methods in 2014, by subdividing the cell pool in subpopulations until a cell with the intended phenotype would be identified either via ddPCR or via MfeI digestion (**Supplementary Figure 3C**).

3.3.3 First test of single nucleotide editing in PC-3 with ssDNA and PS-ssDNA

Based on the results obtained with the reporter system (**section 3.3.1**), we decided to use eSpCas9(1.1) and sgRNA_C for the single nucleotide editing in PC-3. Although the transfection efficiency of the plasmid harboring Cas9 is low (**Figure 4**), we opted for this strategy to avoid the main drawbacks that arise with the use of a stable Cas9 expression, namely the increased off target effects and the re-editing of the locus upon HDR occurrence. In terms of donors, as previously described, some important features needed to be considered to increase the HDR efficiency, including single strand instead of double strand donor, length of the HA and chemical modifications. According to the information in the literature we decided to test two single strand donors: a shorter donor corresponding to the antisense DNA sequence with a phosphorothioate group at both 5' and 3' (PS-ssDNA) and one without chemical modification and longer HA (ssDNA) (Renaud *et al.*, 2016) (**Figure 11A**). PC-3 were transfected with peSpCas9(1.1), pUC19 sgRNA_C and donors DNA harboring the minor allele A. After 3 days from the transfection and 5 days of Puromycin selection (2 µg/ml), DNA was extracted and Sanger sequencing of the PCR product was used for the estimation of editing with TIDE. In this experiment, the cutting efficiency was estimated around 16.7% (R2= 0.98) in presence of ssDNA and around 32.4% (r2= 0.93) in presence of PS-ssDNA; both results were lower compared to the results of the same test performed without any donor DNA, where we estimated total efficiency up to 50% with sgRNA_C (**Figure 8B**). This decrease in editing

efficiency may be attributable to toxicity caused by the donor DNA. Based on the results from this test, we considered the PS-ssDNA as the best option for further experiments (**Figure 11B**).

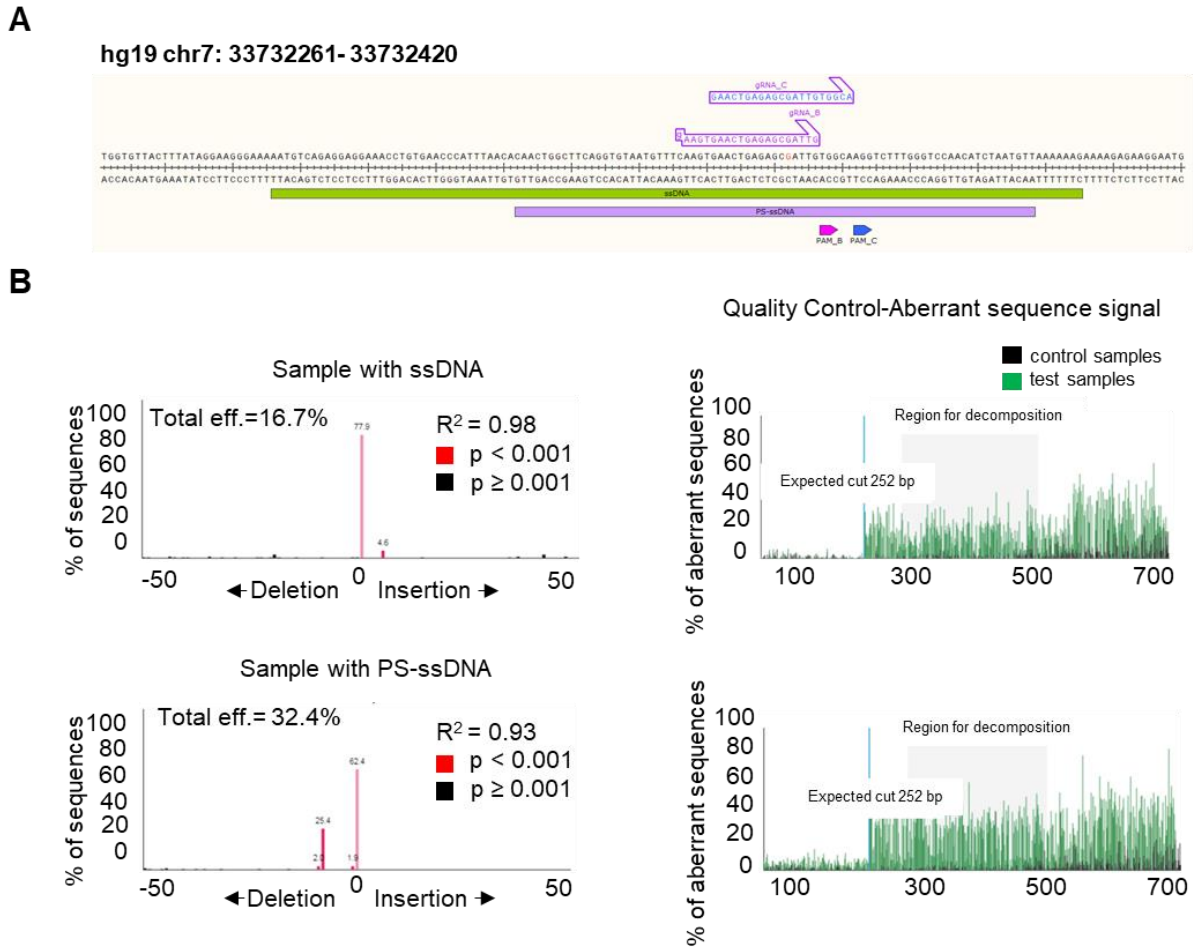


Figure 11. Editing of 7p14.3 with eSpCas9(1.1), sgRNA_C, ssDNA and PS-ssDNA. A) Graphic representation of ssDNA (in green) and PS-ssDNA (in purple) flanking the region on chromosome 7. Red indicates the reference base of the SNP under study, substituted with A in the donors. B) Tide predicted 17% of editing in PC-3 transfected with peSpCas9(1.1), pUC19 sgRNA_C and ssDNA and 32% of cutting efficiency when transfected with peSpCas9(1.1), pUC19_C and PS-ssDNA. Red and black bars in the left graph indicate statistically significant and non significant events. On the right, corresponding graphic representations of aberrant nucleotide signal of the sample (green) compared to that of the control (black).

3.3.4 Screening of HDR events frequency in pool of edited cells with ddPCR and Sanger sequencing

DNA extracted from pool of cells edited with eSpCas9(1.1), sgRNA_C and PS-ssDNA were used for ddPCR. The frequency of events was calculated as (FAM_A/(FAM_A+HEX_G))*100 (**Figure 12A**); importantly, this frequency is not representative of the percentage of HDR *per se*. In fact, A and G probes will bind only to perfect matching A or G sequences and will not bind in case of indels, this is why the ddPCR screening is not able to discriminate between NHEJ and HDR. It is likely that the ~10% of allele A frequency from the screening in pool of cells with sgRNA_C and PS-ssDNA (sample 4) is an overestimation for HDR. To test this, we decided to go through sib-selection by seeding in the first 96 well plate 2-4 cells/well. Once enough cells were available, we split the cells in two new plates, one was used for DNA extraction and for ddPCR analysis while the other one was kept in the incubator. Wells with higher percentage of frequencies were further selected for Sanger sequencing. Most of wells positive in ddPCR presented DNA with an insertion of 45 bp of donor DNA immediately after the cutting site (**Figure 12B** and **C**), compatible with an insertion *via* NHEJ. This result pointed to the need of increasing both transfection efficiency and editing efficiency if we want to

increase our chances of isolating cells with HDR events. Furthermore, a second screening method should be paired with ddPCR rare event detection in order to avoid false positive results in our context.

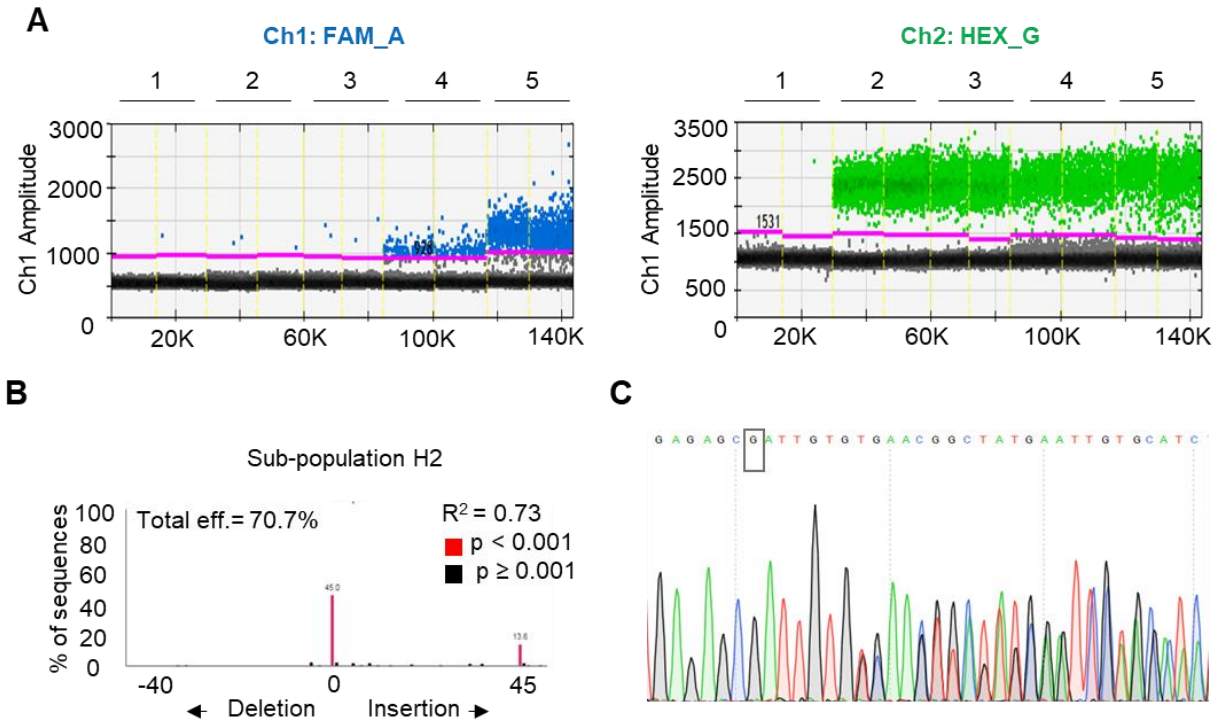
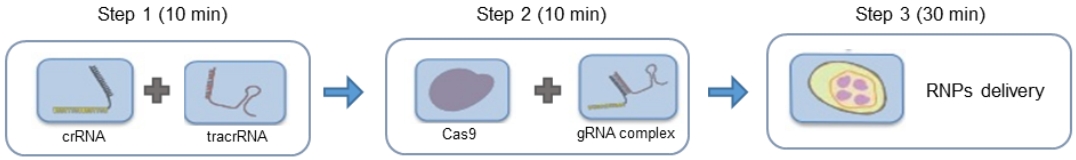


Figure 12. Editing of 7p14.3 with peSpCas9(1.1), sgRNA_C, and PS-ssDNA. A) ddPCR screening of bulk of edited cells transfected with peSpCas9(1.1), sgRNA_scramb (3) or sgRNA_C (4) and PS-ssDNA. Blue dots represent positive events for minor allele A (FAM_A) while green dots indicate positive events for ancestral allele G (HEX_G). PC-3 (G/G) (2) were used as negative control for allele A, HS578Bst (5) were used as heterozygous for the locus (A/G). Well 1 corresponds to sample without template. B) TIDE analysis for sub-population of cells positive for ddPCR identified insertion of 45 bp of donor after the cutting site. C) Insertion of donor DNA via knock-in was further validated with Sanger sequencing. Black box indicates the ancestral allele G of rs1376350; 6 bp after the sequence becomes unclear due to indels created by Cas9 cleavage and donor insertion via knock-in.

3.3.5 Editing of single nucleotide with RNPs and TIDE/TIDER analysis

Low transfection efficiency is a limiting factor for editing efficiency and therefore also for HDR events. For this reason we decided to consider an alternative approach to pursue HDR: the delivery of Ribonucleoprotein (RNP) complexes by nucleofection Cas9-RNP complexes have attracted enormous interest since they can be easily and flexibly reprogrammed to target any desired locus for genome engineering and gene regulation applications. Briefly, an *in vitro*-transcribed or chemically synthesized gRNA is complexed with Cas9 protein and transferred into cells either with classical transfection or *via* electroporation (**Figure 13A**). We first tested the RNP method in PC-3 cells with wt Cas9 protein complexed with three sgRNAs targeting the 7p14.3 locus variant of interest (described in **section 2.1.4**). For this first RNP experiment, the sgRNA guides were transcribed in *vitro* using HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB). Results from TIDE showed that the locus was not edited at all suggesting that either the transfection still remained a major limit in our hands and/or the sgRNAs were not well transcribed (data not shown) or rapidly degraded. We next decided to try with a more sophisticated electroporation instrument and with synthetic sgRNAs chemically stabilized to increase half-life in the cells (sgRNAs purchased from IDT). Cells were electroporated following the protocol reported in Methods (**section 2.1.4**) by using the Nucleofector 4D system (Lonza). Combination of RNPs complex harboring wt Cas9 protein, sgRNA_B or sgRNA_C, ssDNA or PS-ssDNA are listed in Table 5 (**Figure 13B**).

A

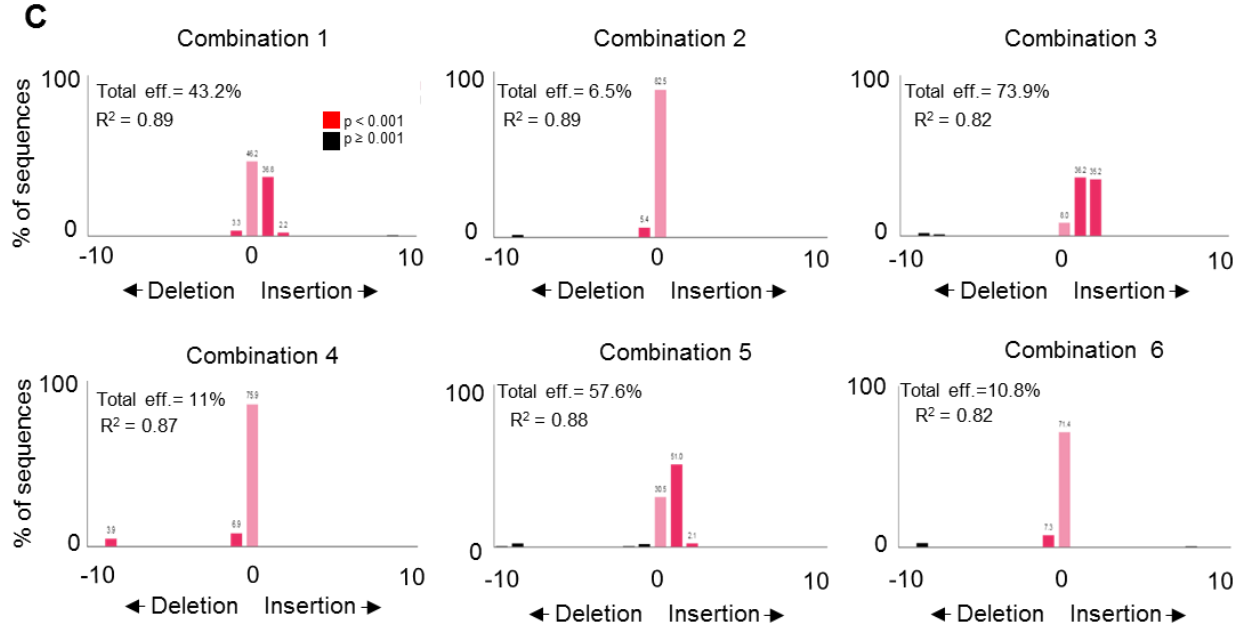


Modified from Jacobi et al., 2017

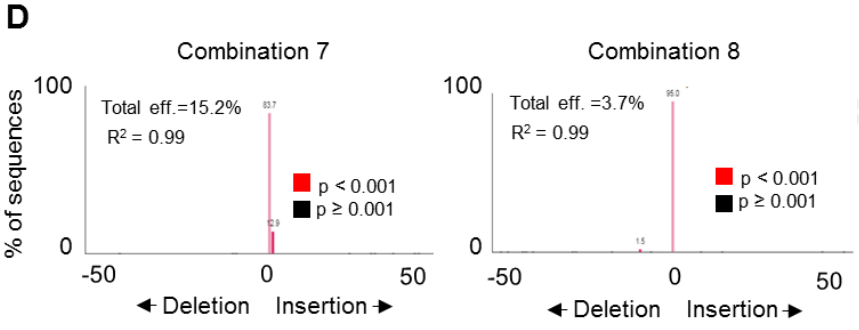
B Table 5. Combinations of RNPs complex with donor DNA and HDR

	1	2	3	4	5	6	7	8
Cas9 protein	+	+	+	+	+	+	+	+
sgRNA_B	+	-	+	-	+	-	+	-
sgRNA_C	-	+	-	+	-	+	-	+
ssDNA	-	-	+	+	-	-	-	-
PS-ssDNA	+	+	-	-	+	+	-	-
HDR molecule	+	+	+	+	-	-	-	-
Electroporation enhancer	+	+	+	+	+	+	+	+

C



D



E

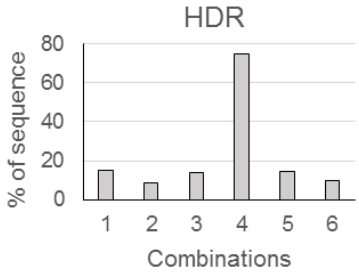


Figure 13. Editing of 7p14.3 with RNPs system in PC-3. A) Graphic representation of RNP method. Once crRNA and trcRNA are linked together, a complex with Cas9 protein is produced and transfected into cells, either with transfection reagents either by electroporation. For this experiment we used synthetic made-to-order sgRNAs. B) In table 5 are listed 8 combination of Cas9 protein, sgRNA, donor DNA, electroporation enhancer and HDR molecule. HDR molecule was used from combination 1 to 4 while electroporation enhancer was used in all samples. C) TIDE was used to predict the percentage of cutting efficiency for each combination, including combination 7 and 8 in which no donor DNA was added. D) Graphic representation of percentage of HDR predicted with TIDER for 6 combinations. Note that TIDER is designed to analyze HDR events in contrast to TIDE.

Due to technical issues, combinations with RNP and sgRNA_B and C with donor ssDNA in absence of HDR molecule was not included in this single experiment. We decided to also test the use of a small molecule reagent that effectively diverts repair pathway towards HDR, to enhance overall HDR efficiency (purchased from IDT). Experiments were run with (combinations 1-4) or without (combinations 5-8) the small molecule (**Figure 13, Table 5)**.

Five days post electroporation, DNA was extracted and amplified and PCR products were sent for Sanger sequencing. TIDE and/or TIDER (**section 2.3.1**) were used for the estimation of editing efficiency and\or of HDR (**Figure 13C, D** and **E**). To test the cutting efficiency, we included in the experiment two combinations of Cas9 protein and synthetic sgRNAs (B and C) without any donor (sample 7 and 8, respectively) (**Figure 13C**). Surprisingly, TIDE estimated an efficiency of editing of 15.2% with sgRNA_B and 3.7% with sgRNA_C, significantly lower than the one obtained using the same sgRNAs delivered as plasmids by transfection with FuGENE (efficiency of editing 92% and 47%, respectively, **Figure 8**). The predicted editing efficiency with TIDE suggested a higher activity of Cas9 protein with sgRNA_B compared to sgRNA_C. (43.2%, 73.9%, 57.6% *vs* 6.5%, 11% and 10.8%) (**Figure 13C**). When we analyzed HDR frequency, we noticed a range of values for combinations 1, 2, 3, 5, 6 with slightly higher values for the ones obtained with sgRNA_B (combination 1, 3 and 5) (**Figure 13D**). For combination 4 the HDR frequency was ~70%, however the TIDE suggest 11% of cutting efficiency (this is supported by ddPCR data, see **section 3.3.6**). Altogether, we concluded that further analysis is necessary to determine the best condition for higher occurrence of HDR events (see next **section**).

3.3.6 ddPCR to test HDR events in electroporated cells

In order to confirm the results suggested from TIDER analysis, we decided to evaluate the percentage of positive signal for HDR events in ddPCR following the same procedure reported in **section 3.3.3**, for each combination from **Table 5**. Signal for FAM_A in channel 1 and HEX_G in channel 2 are indicating positive events for the presence of allele A in the locus (**Figure 14A**) and the fraction of (FAM_A/(FAM_A+HEX_G))*100% quantifies the percentage of allele A

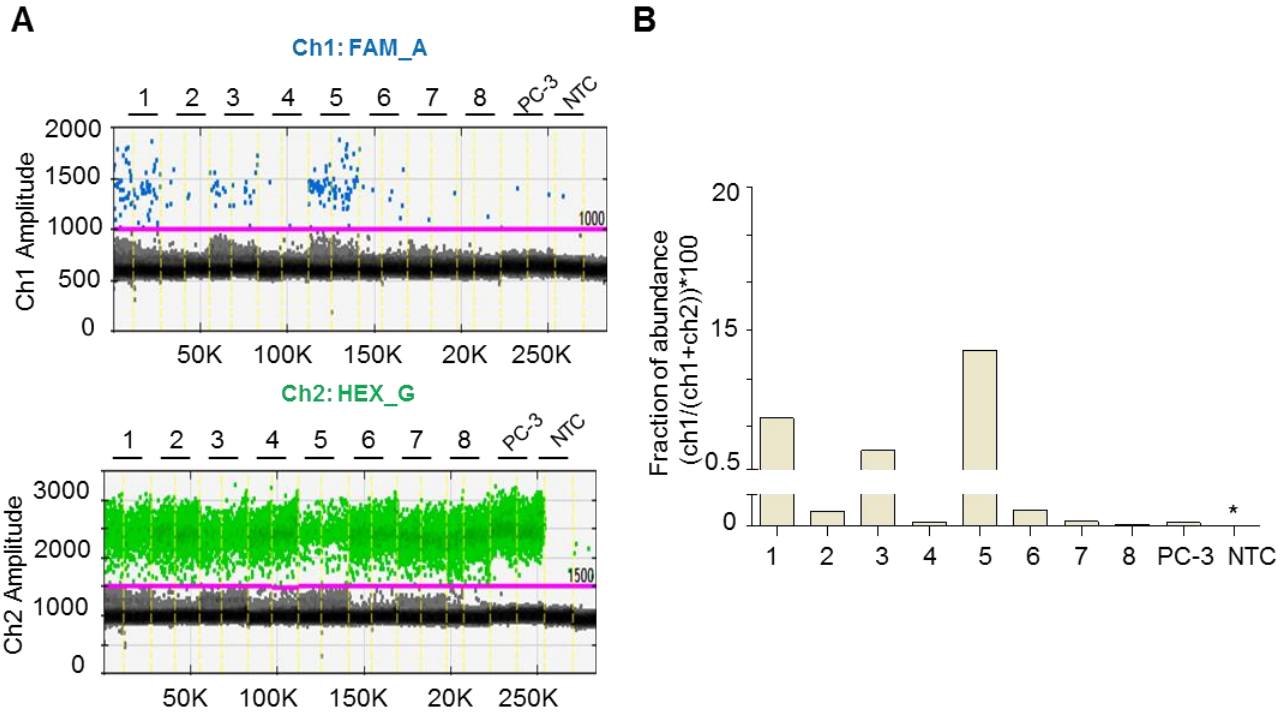


Figure 14. ddPCR to assess events of HDR in samples electroporated with RNPs complex. A) Graphic representation of signal from channel 1 and channel 2 corresponding to FAM_A (blue) and HEX_G (green) respectively from samples electroporated with RNPs (**Figure 13B, Table 5**). Each sample was run in duplicate. Vertical yellow lines distinguish different wells. Horizontal purple line indicates the signal cutoff, manually selected based on positive droplets and intensity of background signal. Grey and black dots represent negative signal due to absence of template inside each droplet. B) Mean frequencies are calculated for the two wells of each sample (ch1/(ch1+ch2)). *Negative control (NTC) includes mixture of probes and primers without template; the NTC fraction is not informative due to few counts from contamination and therefore is not reported.

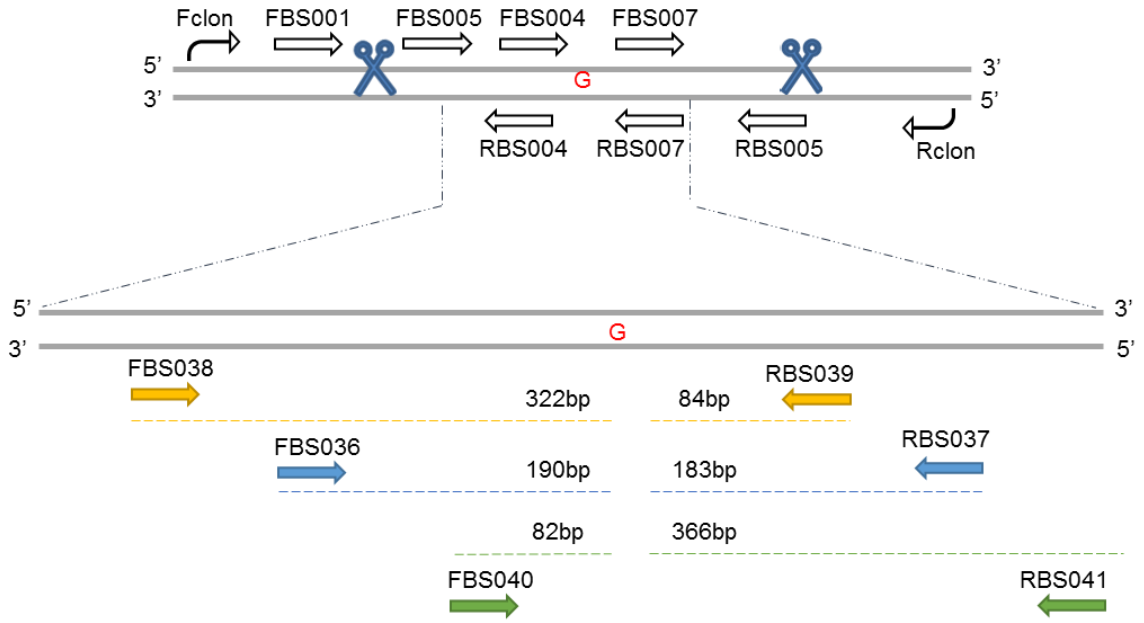
without discerning possible indels that may occur after the editing (see **section 3.3.4**). Fractions for combination 1, 3 and, more so, 5 were considerable (more than 2% for the first round of screening). No significant results were obtained for combinations 2 and 6 (transfected with sgRNA_C) (**Figure 14B**). For combination 4, TIDER suggested a substitution of one allele via HDR in 70% of the population, but no positive result was provided by ddPCR, confirming an over estimation of the algorithm of TIDER software. As expected, for combinations 7 and 8 (lacking the donor DNA), no evidence of HDR events was detected. Negative control sample (-) ratio of 20% is an aberration due to the number of droplet randomly appearing as background contamination (ie 2 A droplet vs 8 G droplets).

3.3.7 Design of targeted ultra-deep sequencing experiment to calculate HDR events from electroporated samples DNA

To exclude that positive A probe signal in ddPCR was an artefact due to a partial insertion of the donor segment instead of a true HDR event, we needed to confirm these results with an alternative approach. We used a targeted ultra-deep sequencing strategy, which allowed the quantitative discrimination between HDR and NHEJ frequency. Moreover, the precise estimate of the percentage of HDR positive events for allele A facilitated the decision of the screening approach to follow. In case of HDR events frequency < 1%, we would have proceed by sib-selection (Miyaoka *et al.*, 2014). Furthermore, ultra-deep sequencing provides useful to quantify possible knock-in events around the locus. We selected a genomic region of ~600 bp spanning the locus of interest to be amplified with three pairs of primers. Primers were designed *up* and *down*-stream the 7p14.3 locus variant using the Prime3 tool (**Figure 15,** sequences in **Supplementary Table 10**). PCR amplification was optimized and performed separately for each set of primers and then the three mix of PCR were sent for sequencing. Mix1: FBS036-RBS037; Mix2: BS038-RBS039; Mix3: FBS040-RBS041. DNA libraries were prepared as described in **section 2.4.1**. from DNA obtained from combinations 1, 3 and 5 (Cas9 with sgRNA_B, with ssDNA or PSssDNA,), chosen because of the promising estimation performed by TIDER and TIDE, while DNA from combination 2 (obtained with Cas9 with sgRNA_C, and PSssDNA) was chosen to generate target sequencing data that allowed comparison between sgRNA-B and C. Mix 1, Mix2 and Mix3 of PCR from combination 1, 2, 3 and 5 were sequenced and analyzed by using CRISPResso, a computational pipeline able to qualitatively and quantitatively evaluate the outcomes of genome-editing experiments of deep sequencing of the target loci (Pinello *et al.*, 2016). Based on the wt sequence with the G allele and the alternative sequence with the A allele, the software calculated the percentage of reads flanking the wt sequence (percentage unmodified), the percentage of reads with indels (percentage NHEJ events) and the percentage of reads corresponding to the alternative sequence (percentage HDR). The average coverage obtained across the samples was >56000X (range: 48729X, 67619X).

The results showed good concordance between mixes covering the locus, indicating the same amount and profile of indels. The most frequent event was the insertion of 1 nucleotide after the cutting site followed by ~10 bp deletion with a lower attendance. No larger indels were observed. The experiment confirmed that sgRNA_C cuts the locus with a lower efficiency compared to sgRNA_B (unmodified ~75%, % NHEJ ~25%, no HDR detected) (Combination 2, **Figure 15**). Instead, combination 1 and 3, in which the only difference was the donor DNA, presented lower HDR events (3,3% and 1% respectively) compared to combination 5 (5,4%). Meanwhile, in combination 1 the percentage of NHEJ (57,6%) was lower comparing to combination 5 (71,5%) suggesting that the HDR molecule added to combination 1 and presence of PS-ssDNA donor successfully inhibited NHEJ events. In combination 3 the NHEJ (70,3%) was higher compared to combination1 but lower compared to combination 5. The most frequent event occurring in three samples was the insertion of 1 nucleotide with a frequency of 46% in combination 5, 44% and 34% in combination 3 and 1, respectively. These results suggest that PS-ssDNA increased the chance of positive events for allele A. Results obtained with Mix 2 and Mix 3 are shown in **Supplementary Figure 4**.

A


Fclon
FBS001
FBS005
FBS004
FBS007
5'
3'
G
3'
5'
RBS004
RBS007
RBS005
Rclon
5'
3'
G
3'
5'
FBS038
322bp
84bp
RBS039
FBS036
190bp
183bp
RBS037
82bp
366bp
FBS040
RBS041

B

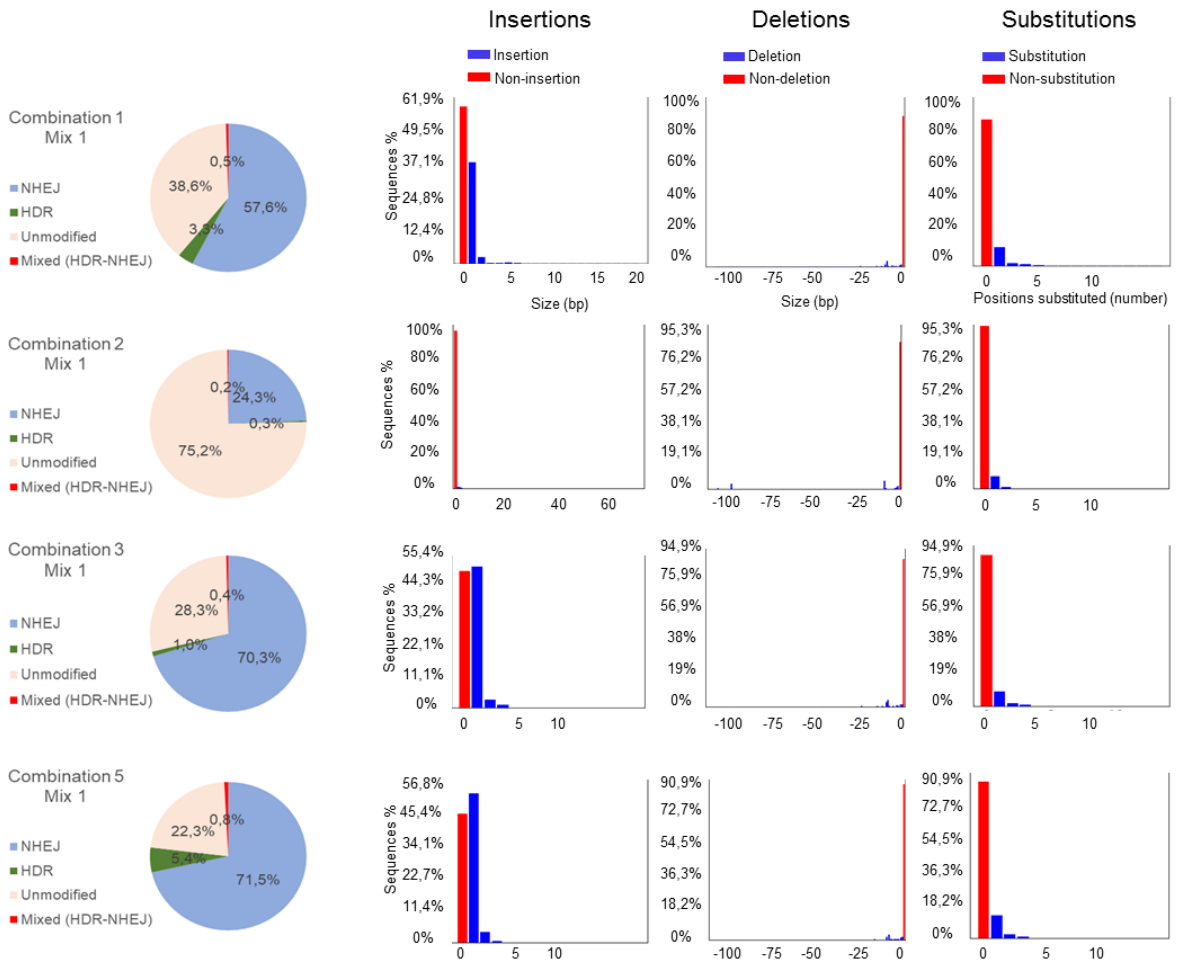
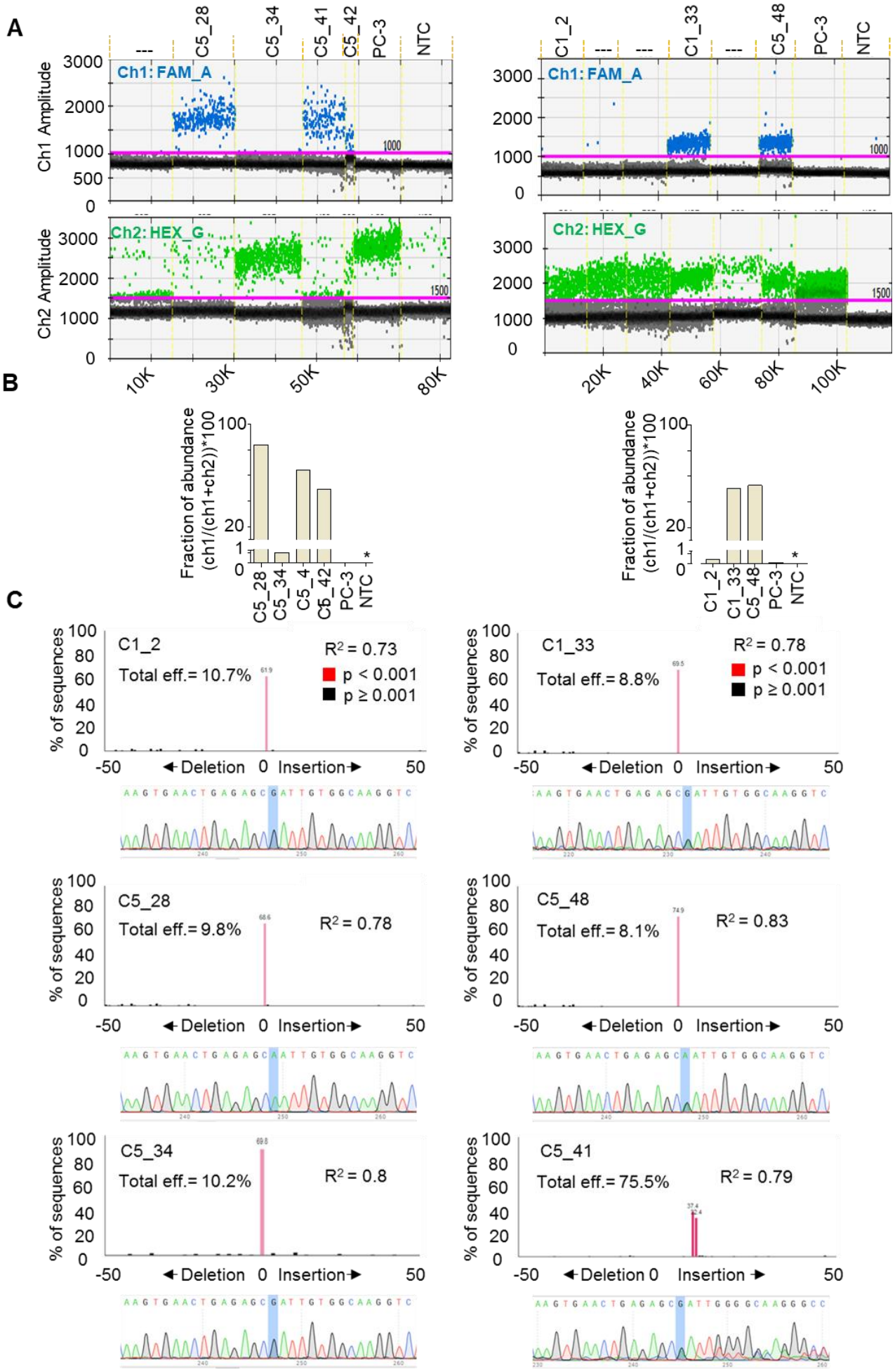

Insertions
Deletions
Substitutions
Insertion
Non-insertion
Deletion
Non-deletion
Substitution
Non-substitution
Combination 1
Mix 1
NHEJ
HDR
Unmodified
Mixed (HDR-NHEJ)
0,5%
38,6%
57,6%
3,3%
Combination 2
Mix 1
0,2%
24,3%
0,3%
75,2%
Combination 3
Mix 1
0,4%
28,3%
1,0%
70,3%
Combination 5
Mix 1
0,8%
22,3%
5,4%
71,5%
Sequences %
Size (bp)
Positions substituted (number)

Figure 15. Targeted sequencing of the locus edited with RNPs. A) Graphic representation of 1603 bp of genomic region on chromosome 7 across the locus of interest, with sets of primers used to screen the clones positive for the deletion as described in **section 3.1.3**. Three pairs of primers were designed within the portion of DNA corresponding to the Macrodeletion (blue scissors correspond to the cutting site from Cas9 endonuclease). More than 600 bp of genomic region was covered with three set of primers upstream and downstream the locus as represented in the figure. Each color corresponds to a pair of primers used for PCR amplification. **B**) CRISPResso calculated indels based on the allele A target sequence and the allele G control sequence for Mix1 (FBS036-RBS037). The pie chart shows the quantification of editing frequency as determined by the percentage and number of sequence reads. Bar plot represent insertion, deletion and substitution calculated for the represented samples. (Results from Mix 2 and 3 are presented in **Supplementary Figure 4).**

3.3.8 Screening clones for combination 1 and 5 with ddPCR

Targeted ultra-deep sequencing results suggested higher probability to find clones positive for the substitution of allele G with allele A in pool of combination 1 and 5 (3% and 5% respectively of HDR events, (**Figure 15**). We therefore decided to utilize a classical method of screening in which a serial dilution of bulk of cells would allow to have one cell/well in 96 well plate for these two combinations. Once confluent, wells were divided in two plates, one was used for DNA extraction and the other one was kept in the incubator. DNA extracted from one 96 well plate was used as template for ddPCR analysis (**Figure 16**). Wells that resulted positive by ddPCR were selected for Sanger sequencing. In a total of 90 screened clones (40 for combination 1 and 50 for combination 5), we found one homozygous genotype for the alternative allele (well C5_28, A/A) and 2 heterozygous genotype (C1_33 and C5_48 A/G) coming from bulk of cells of combination 5 and 1, respectively. As negative control (G/G), we selected clones from both combinations in which editing didn't occur (example C5_34 and C1_2). For some of the clones with FAM_A positive signal but HEX_G negative, we noticed by Sanger sequencing and TIDE that one allele (FAM_A) was positive and the other allele harbored indels that did not allow HEX_G to recognize the locus and to release fluorescent signal. For example, clone C1_5 presented a possible insertion of 1 or 2 nucleotides and C5_42 presented insertion of 5 or 6 nucleotides as a result of the repair upon editing. As previously mentioned, allele G substitution with allele A creates a restriction site for MfeI. Clones positive for ddPCR were analyzed through digestion of 1603 bp of PCR amplicon derived from Fclon-Rclon (**Supplementary Figure 5**).


A

C5_28
C5_34
C5_41
C5_42
PC-3
NTC
Ch1: FAM_A
Ch1 Amplitude
Ch2: HEX_G
Ch2 Amplitude
1000
1500
10K
30K
50K
80K
C1_2
C1_33
C5_48
20K
40K
60K
100K
B
Fraction of abundance (ch1/(ch1+ch2))*100
C5_4
C1_4
C
% of sequences
C1_2
Total eff.= 10.7%
R² = 0.73
p < 0.001
p ≥ 0.001
C1_33
Total eff.= 8.8%
R² = 0.78
C5_28
Total eff.= 9.8%
R² = 0.78
C5_48
Total eff.= 8.1%
R² = 0.83
C5_34
Total eff.= 10.2%
R² = 0.8
C5_41
Total eff.= 75.5%
R² = 0.79
-50
Deletion
0
Insertion
50

Figure 16. Screening single clones with ddPCR. A) Only well C5_28 and C5_41 and C5_42 show positive signal for FAM_A within a very low signal for HEX_G. Well C5_34 seems to have a similar profile to wt PC-3. B) Fraction of FAM_A signal is calculated and presented in the lower graph (percentage presented as an average of two wells for each well). C5_28 have 84% of positive events for allele A, C5_41 and C5_42 have 64% and 49% of positive events for FAM_A respectively. Samples are sorted as in panel A. C). The chromatogram for each clone is shown below the TIDE graph. The nucleotide of interest is shown in blue. *Negative control (NTC) includes mixture of probes and primers without template; the NTC fraction is not informative due to few counts from contamination and therefore is not reported. Yellow lines shows delineate each well and black dash lines are indicative of other samples not reported in the figure.

3.4 The transcriptomes of 7p14.3 edited cells

Upon establishment of the isogenic cell lines homozygous and heterozygous for allele A at rs1376350, we generated RNA-seq to study the effect of the variant on the cells transcriptome. The analysis included comparison of deregulated genes (DEGs) as detected for each edited cell (or pool of cells) and their control across all, including macrodeletion single clones, microdeletion pool of cells, single clones from the disruption motifs experiment and clones selected positive for the heterozygous or homozygous genotype at rs1376350, the latter with or without AR overexpression (detailed information in **Supplementary 7, Table 4A and B**). Although we are unaware of the real binding affinity of each TF in the selected clones upon the disruption motif experiments, we classified all clones upon disrupted motifs in groups based on an *in silico* prediction for the binding, to obtain four groups. As first step of analysis, while we recognizing that DEG detection on different cell types included variable numbers of clones and/replicates affecting the power of the tests, we qualitatively compared the total number of detected and of shared DEGs (**Figure 17A**). The macrodeletion DEG set was the second largest, including more than 800 genes; this is in line with a remarkable effect on transcriptome due to possible implication of other motifs residing in the deleted portion of DNA beyond the polymorphic locus (Pritchard *et al.*, 2016). High numbers of DEGs were also detected for the various groups of cells with combinations of disrupted motifs, with significant intersections with the microdeletion set. For instance, 154 genes out of 796 from the AR.neg.CEBPβneg DEG set are shared with the macrodeletion set. Similarly, one fourth of the DEG list from the SNP.AA clone (26 out of 103) are shared with the macrodeletion set. Of note, the test that led to the smallest signal is the heterozygous clone SNP.AG, with 9 altered transcripts. The considerable number of DEGs for DisMotif:AR.pos.CEBPβhet and DisMotif:AR.pos.CEBPβhet can be explained by the fact that different controls were used for the normalization since some of clones had a different *history*.

The take-home message from this first analysis is that the closer we go to the locus of interest, the more precise information we obtain strictly regarding the variant of interest. We observed a lower DEGs number in heterozygous cells (A/G) compared to (A/A) but in complex the change of a single nucleotide in both alleles is associated with a considerable number of DEGs (100 genes).

A

	Macro	Micro pool	DisMotif.ARneg. CEBPβneg	DisMotif.ARhet. CEBPβneg	DisMotif.ARpos. CEBPβhet	DisMotif.ARpos. CEBPβneg	DisMotif.ARpos. CEBPβpos	SNP.AG	SNP.AA	SNP.AR.EtOH	SNP.AR.DHT	DEGs (N)	up-DEGs (N)	down-DEGs (N)
Macro	857											857	388	469
Micro pool	75	95										95	44	51
DisMotif.ARneg.CEBPβneg	154	31	796									796	395	401
DisMotif.ARhet.CEBPβneg	56	14	135	279								279	170	109
DisMotif.ARpos.CEBPβhet	116	23	234	219	678							678	280	398
DisMotif.ARpos.CEBPβneg	26	5	22	10	14	268						268	97	171
DisMotif.ARpos.CEBPβpos	173	33	288	200	497	23	1249					1249	307	942
SNP.AG	4	0	2	3	3	1	4	9				9	5	4
SNP.AA	26	7	30	18	32	4	53	4	103			103	82	21
SNP.AA.AR.EtOH	9	1	14	12	15	3	24	4	32	46		46	24	22
SNP.AA.AR.DHT	12	1	16	20	25	3	34	4	29	38	55	55	26	29

* DEGs selected by FDR <0.05

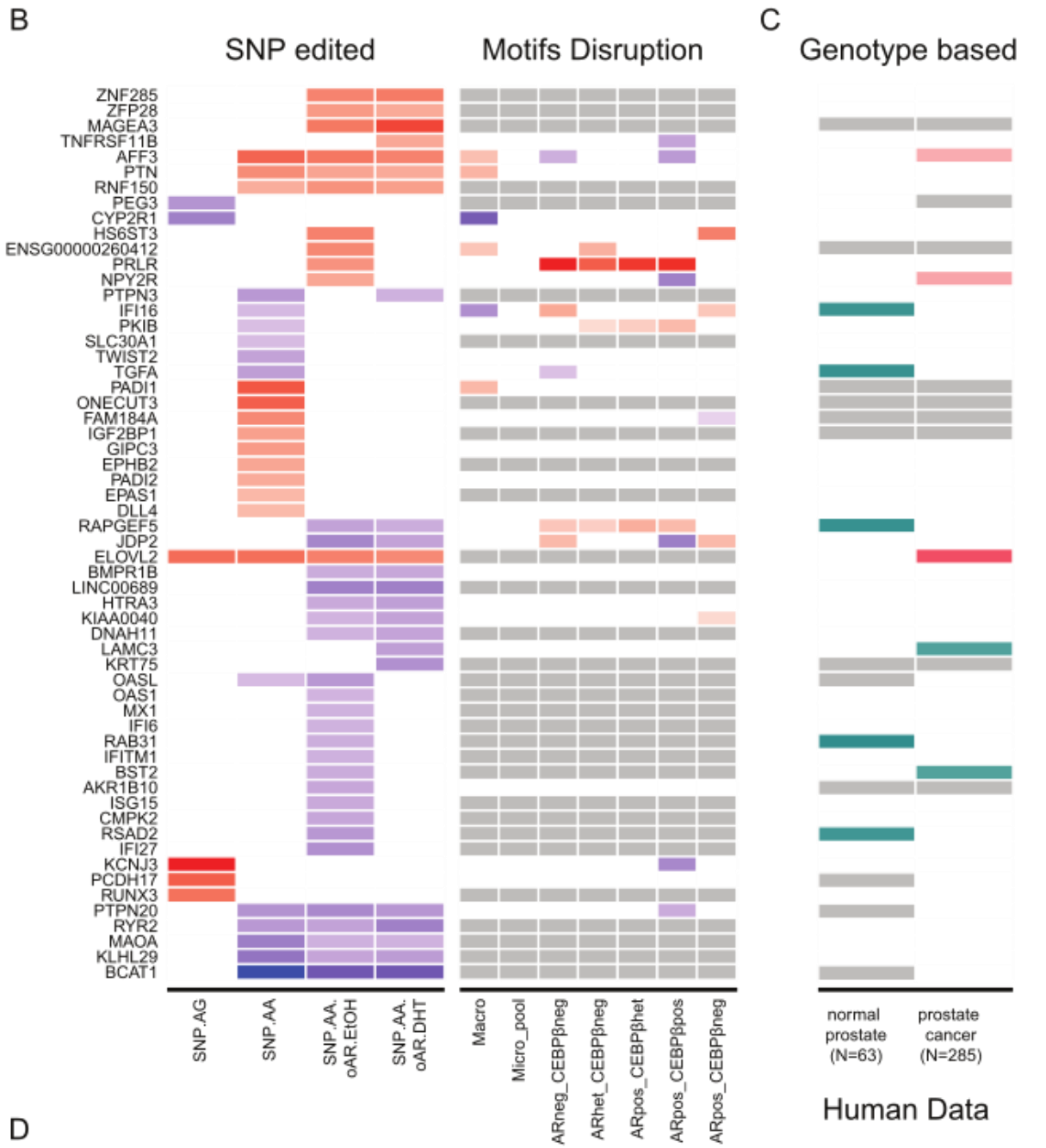

B
SNP edited
Motifs Disruption
C
Genotype based
ZNF285
ZFP28
MAGEA3
TNFRSF11B
AFF3
PTN
RNF150
PEG3
CYP2R1
HS6ST3
ENSG00000260412
PRLR
NPY2R
PTPN3
IFI16
PKIB
SLC30A1
TWIST2
TGFA
PADI1
ONECUT3
FAM184A
IGF2BP1
GIPC3
EPHB2
PADI2
EPAS1
DLL4
RAPGEF5
JDP2
ELOVL2
BMPR1B
LINC00689
HTRA3
KIAA0040
DNAH11
LAMC3
KRT75
OASL
OAS1
MX1
IFI6
RAB31
IFITM1
BST2
AKR1B10
ISG15
CMPK2
RSAD2
IFI27
KCNJ3
PCDH17
RUNX3
PTPN20
RYR2
MAOA
KLHL29
BCAT1
SNP.AG
SNP.AA
SNP.AA. oAR.EtOH
SNP.AA. oAR.DHT
Macro
Micro_pool
ARneg_CEBPβneg
ARhet_CEBPβneg
ARpos_CEBPβhet
ARpos_CEBPβpos
ARpos_CEBPβneg
normal prostate (N=63)
prostate cancer (N=285)
Human Data
D
TPTE
TPTE2
TNS3
EFNB3
IL12B
CXCR4
CASP3
EFNA1
EFNA3
CXCR2
PLCG1
PTEN
IL23A
EFNA4
ENSP00000426741
EFNA2
SRC
EFNA5
RASA1
PIK3CA
ENSP00000274335
SMARCA2
SMARCA4
ASH1L
TAF1L
ENSP00000276072
RPTOR
EIF2S2
IRF1
STAT2
IRF9
log(fold change)
−10
10
NA
-log(p)
1.8
2.8
down
up
-log(p)
20
9
down
up

Figure 17. The transcriptome of cells edited at and around rs1376350. A) Comparison of the lists of deregulated genes (DEGs) from edited versus non edited cells. The table lists the total number of DEGs (*up* and *down*-regulated *vs* controls) and the number of shared DEGs for each comparison. All the DEGs are selected based on FDR <0.05. B) The heatmaps show information for a list of DEGs identified in the single clone single nucleotide editing experiments based on FDR <0.05. Rows are sorted per correlation based clustering (pearson correlation of log(fold change)). Genes differentially expressed between single clone control samples are omitted. The four most left columns report data from single nucleotide editing experiments, followed by the macrodeletion, microdeletion and disruption motifs data. Grey indicates genes with coefficient of variance >0.9 in deletion and motif editing controls. Color scale indicates logFC in every comparison, only significant differentials are colored. C) Differential expression analysis results for a set of human prostate date, for genes as in panel B. The heatmap shows the significance of differential expression (log(p), Mann-Whytney test, one tail, sign indicates direction) in human tissues comparing individuals with A/A or A/G genotypes respect to individuals with ancestral genotype (G/G). D) Prediction of upstream regulators. The heatmaps show the significance (-log(p)) of the predicted upstream regulators based on DEGs identified in the single clone single nucleotide editing experiments.

Next, we wanted to examine in details which are genes affected by the single nucleotide substitution (with or without AR overexpression). For DEGs selected from the SNP experiments (**Figure 17B**), we interrogated their profiles in the transcriptomes of the DisMotifs groups by using strict criteria. When we apply coefficient of variance >0.9 in deletion and motif editing samples, only *PRLR* remain in the list of DEGs in almost all subclasses of edited cells. The number of concordant deregulation profiles was higher with less stringent thresholds, but in the absence of replicates we opted for a conservative output.

To test whether the deregulation signal detected in the edited cells is relevant to human prostate cells, we interrogated a large set (N>300) of human data with known genotype and transcriptome data, both using normal and tumor prostate tissue samples (**Figure 17C**). We observed that *AFF3* and *ELOVL2* are upregulated in tumor samples harboring the allele A ((A/G) and (A/A)) as in SNP.AA with and without AR overexpression. While *NPY2R* is upregulated in tumor samples and cells with minor allele A, no differences are observed in A/A cells upon AR overexpression. Few genes, *RAPGEF5* and *IFI16*, resulted down-regulated in normal prostate tissues (as in samples with allele A), upregulated in cells with disrupted motifs, and unaltered in tumor samples. While *TGFA*, *RAB31* and *RSAD2* are down-regulated in both normal prostate samples and in cells carrying the allele A, but their expression is unaltered in human prostate cancer samples cohort.

Last, we asked what are the possible upstream regulators based on the detected DEGs in SNP edited experiments and then queried those in the upstream regulators from the DisMotifs class of cells (**Figure 17D**). A set of regulators was inferred uniquely to the SNP.AA cells, whereas two subsets of activated and repressed regulators were shared among multiple edited cells. Examples of activators include ASH1L, SMARC2 and SMARC4, two members of the SWI/SNF complex. Members of Interferon pathway (IRF1 and IRF9) and STAT2 were nominated as shared repressors in particular when AR is activated with DHT and in DisMotifs subclasses.

These initial results obtained from the analysis of the transcriptomes of edited cells suggested a role for the SNP of interest, augmented by the comparative analysis with the disrupted motifs results and more so by the human data. Although preliminary, the up-stream regulators analyses evidenced two important pathways of chromatin conformation and interferon pathway as deregulated in presence of minor allele of the variant (with or without AR overexpression) or in presence of disrupted motifs suggesting an implication of the locus in a 3D spatial contact with other portion of the genome under still unknown mechanism. RT-qPCR validation of selected genes supported this data (see **section 3.4.1**).

3.4.1 RT-qPCR validation of genes selected from RNA-seq analysis

To validate the RNA-seq analysis data, we selected few genes to be tested by RT-qPCR by using the same RNAs used for library preparation (**section 2.2.11**). We measured the levels of 6 selected genes in homozygous C5_28 (A/A), heterozygote C1_33 (A/G) and control C1_2 (G/G) cells. The analysis confirmed the trends of de-regulation that emerged from RNA-seq analysis. Specifically, *MAGEA3*, *NPY2R* and *PTN* transcripts are up-regulated in C5_28 (A/A) cells compared to heterozygote C1_33 (A/G) and control cells C1_2 (G/G). *ELOVL2* expression was upregulated in both C1_33 and C5_28 clones compared to the C1_2 with a higher level of expression in homozygous cells (C5_28), while *PRLR* was up-regulated in both C1_33 (A/G) and C5_28 (A/A).

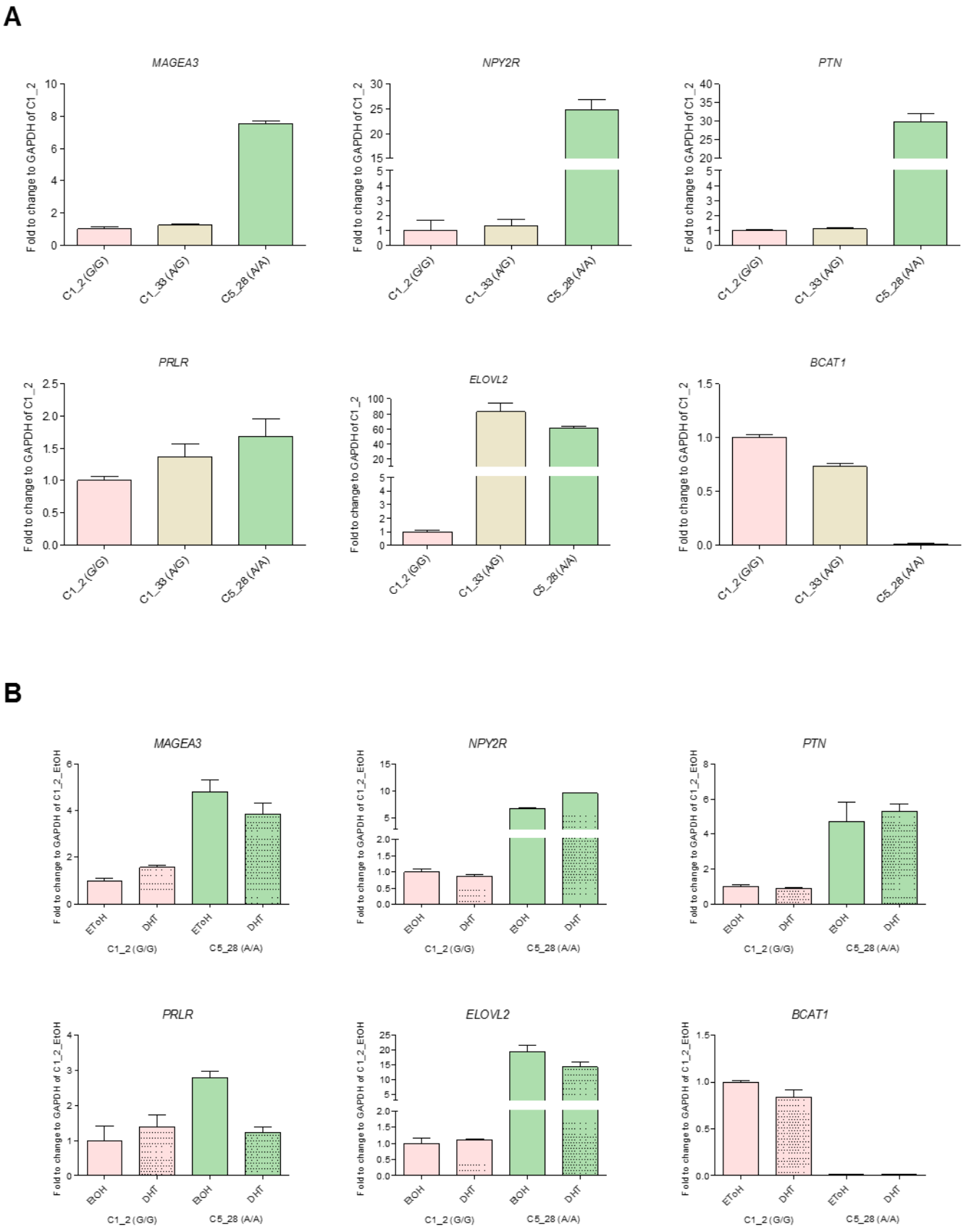


Figure 18. RT-qPCR validation of RNA-seq analysis in single nucleotide edited clones. A) Validation of six genes by RT-qPCR from the same RNA used for library preparation of sample C1_2 (G/G); C1_33 (A/G); C5_28 (A/A). B) The same list of genes were validated in sample C1_2 (G/G) and C5_28 (A/A) with AR overexpression upon EtOH or DHT treatment. All mRNA levels are normalized to housekeeping GAPDH gene expression in C1_2 (G/G) cells. Error bars represent the standard error of the mean of three technical replicates.

When we checked the levels of *BCAT1*, we confirmed the result of RNA-seq with reduced expression of the gene in C5_28 (A/A) cells compared to C1_2 (G/G) (**Figure 18A**). The results were further confirmed on a second biological replicate (**Supplementary Figure 6**).

We interrogated the same transcripts in samples C1_2 (G/G) and C5_28 (A/A) overexpressing AR with LvpAIP upon EtOH or DHT treatment (**Figure 18B**). The levels resulted in line with the data without AR activation, suggesting that for this set of genes AR does not play main role in their regulation. However, biological replicates and further experiments needed on a number of heterozygous and homozygous clones for the locus of interest.

3.4.2 ChIP-qPCR for histone marks in PC-3 cells and in clones positive for allele A

The 7p14.3 locus was initially nominated as a putative functional region based on data mining of human variants (polymorphic sites) that would lie in regions defined as functionally active by ENCODE histone mark ChIP-seq data (Consortium *et al.*, 2012). In general enhancers are associated with high H3K4me1 and low H3K4me3 whereas promoters with low H3K4me1 and high H3K4me3 (Heintzman *et al.*, 2007). The functional activity of the locus was tested by *in vitro* Luciferase assay in wt PC-3 cells with genotype G/G where an enhancer signal was detected (Romanel *et al.*, 2017). Furthermore, ChIP analysis performed in the same cells confirmed that the 7p14.3 locus is marked by histone specific of enhancer regions. Specifically, nuclear proteins extracted from PC-3_LvpAIP_AR were immunoprecipitated with H3K4me1 or H3K4me3 antibodies and qPCR of GAPDH_promoter and KLK3_enhancer regions were used as a positive control. Results showed that in wt PC-3, H3K4me1 is high whereas H3K4me3 is low comparing to GAPDH (**Figure 19A**). Low levels of recruitment of H3K4me1 in KLK-3 enhancer known to be a target of AR are explained by the fact that this gene is only slightly expressed in PC-3 cells.

H3K4me1 is widely distributed and generally covers substantially broader regions than the underlying genetic enhancer elements. It has been shown that H3K4me1 is lost or reduced at enhancers after they are disengaged from the transcriptional machinery (Calo and Wysocka, 2013). H3K27ac is a mark associated with active enhancers, on the contrary, enhancers and regions enriched for H3K27me3 are considered inactive (Zentner, Tesar and Scacheri, 2011). We further investigate the activity of the locus by testing for the enrichment of H3K27ac and H3K27me3 in our isogenic cell line with A/A and G/G genotype upon AR overexpression. We used primers for the FKBP5 enhancer, known to be induced by AR, primers for the NCL promoter as positive controls for H3K27Ac and primers for a gene not regulated by AR, NEUROD2 as positive control for H3K27me3 (Stelloo *et al.*, 2018). Results showed a robust enrichment of the signal for H3K27Ac in the positive control regions FKBP5 enhancer and NCL gene. Although the percentage of input of the occupancy of modified histone in 7p14.3 calculated in C1_2 (G/G) and C5_28 (A/A) is much lower compared to the percentage of input obtained for the positive control, a consistent difference of enrichment is observed between C1_2 and C5_28 clones. The signal for H3K27Ac enrichment in the locus, marker of active enhancer, is higher in C5_28 (A/A) compared to C1_2 (G/G) (**Figure 19B**). Conversely, the signal for H3K27me3 is lower in C5_28 (A/A) compared to C1_2 (G/G) (**Figure 19C**) supporting the active role of the locus as an enhancer and supporting our *in silico* data suggesting that the region is more active when the minor allele A is present. Comparable results obtained from nuclear extracts of PC-3 cells overexpressing AR and stimulated with DHT suggest that AR activation does not induce histone modifications at the 7p14.3. Take together these results show that functional activity of the locus is higher for the minor allele A compare to the major allele G, independently from AR activation.

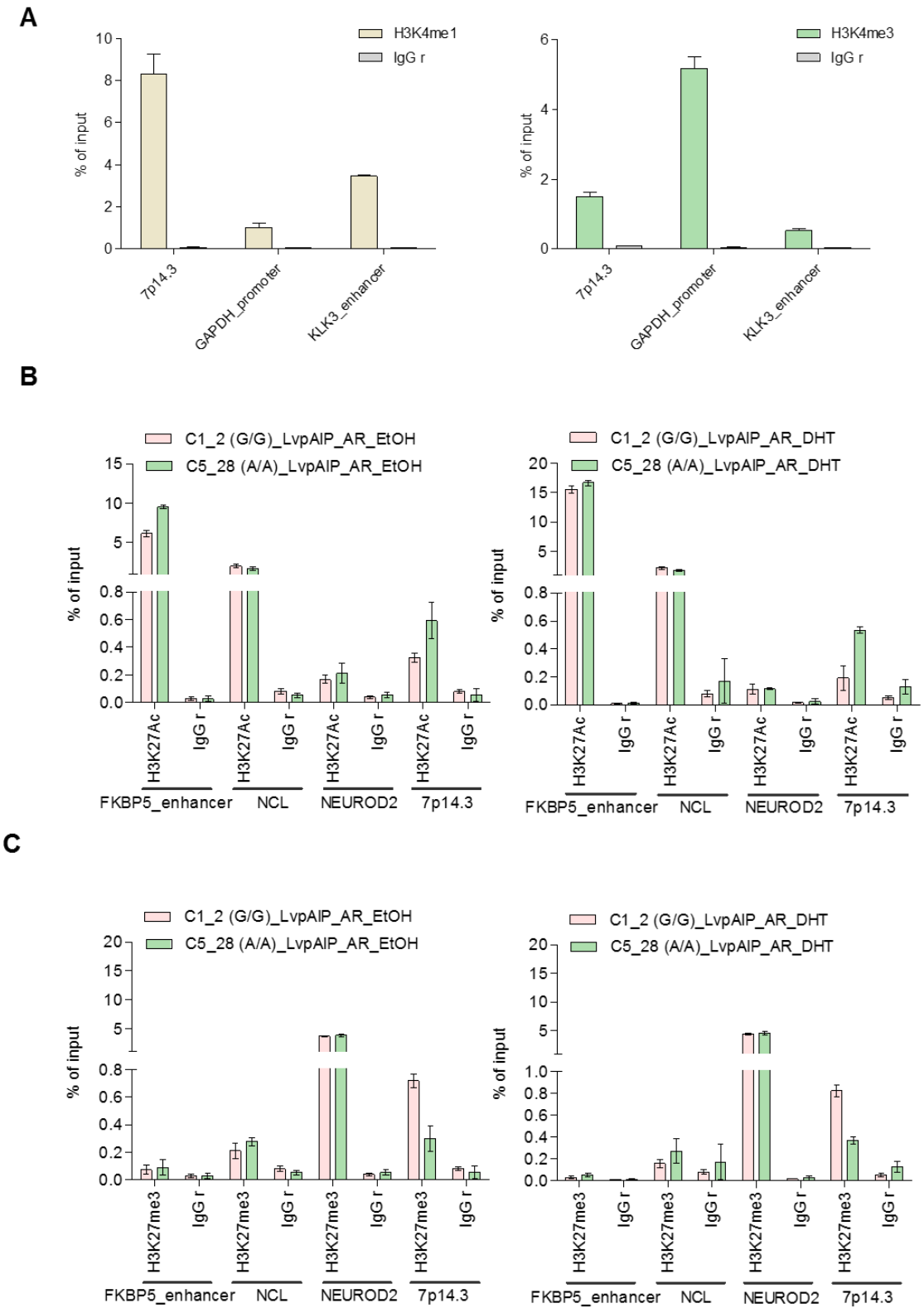


Figure 19. ChIP-qPCR of histone marks in PC-3 and single clone edited PC-3_LvPAIP_AR. A) ChIP-qPCR data for H3K4me1 and H3K4m3 in PC-3_LvPAIP_AR. GAPDH promoter and KLK3 enhancer represent positive control for promoter and enhancer regions respectively. B) ChIP-qPCR at H3K427Ac in C1_2 (G/G) and C5_28 (A/A) overexpressing AR upon EtOH or DHT treatment. FKBP5 enhancer and NCL are shown as positive controls whereas NEUROD2 and IgG as negative controls. Error bars represent the standard deviation of the mean of three technical replicates. C) ChIP-qPCR for H3K4me3 cells as in B. FKBP5 enhancer and NCL are shown as negative controls whereas NEUROD2 as positive control. Error bars represent the standard deviation of the mean of three technical replicates.

4 Discussion

4.1 Editing of single nucleotide via HDR

The overall aim of my work consists in the functional characterization of a germline variant located in a non-coding region on 7p14.3. This variant was identified as associated with an early somatic event that occurs in about 10% of PCa patients with localized disease. In order to functionally validate the 7p14.3 locus, we worked on the establishment of isogenic cell lines harboring the minor allele by using CRISPR/Cas9 system. Aware of the challenges of the approach, including delivery and low efficiency of HDR events in *in vitro* and *in vivo* systems, the rationale was to pursue multiple CRISPR/Cas9 based strategies in parallel to the single nucleotide editing, in order to increment the probability of a successful relevant editing. Specifically, this included three main set ups, two for deletions, a macrodeletion and a microdeletion of 731 and of 50 bp (both surrounding the variant), and a third one to alter the binding sites of TFs that we had previously nominated through *in silico* analyses (for AR and CEBPβ). In this latter approach, we exploited the Cas9 endonuclease activity to induce the disruption of their motifs.

As a very first step, we assessed the efficiency of transfection with results in line with what is reported in the literature; the larger the size of the plasmid harboring the Cas9 endonuclease, the lower the transfection efficiency and, consequently, the less efficient the editing. Even though the activity of Cas9 might also depend from the accessibility of the locus, in general the use of stable Cas9 expressed into cells better allows for genome editing compared to the use of Cas9 transfection *via* plasmid. It the context of single nucleotide via HDR, however, the stable Cas9 solution is more complicated to adopt because of the low discriminatory capacity that might cause re-editing of the locus once the HDR occurred. For this reason, we tested re-editing event by three Cas9 enzymes, wt SpCas9, eSpCas9(1.1) and evoCas9, by using a reporter system. We observed that wt SpCas9 does not discriminate the difference of a single nucleotide between two targets and acts with the same efficiency on the two alleles; that evoCas9, tested with sgRNA_A only, does not cut the locus with a good efficiency, and that eSpCas9(1.1) better discriminates the two nucleotides when combined with sgRNA_C. This prompted us to use eSpCas9(1.1) with sgRNA_C transiently transfected in our PC-3 cells with the aim to substitute ancestral allele G with minor allele A at the rs1376350 locus.

Since single nucleotide editing requires the presence of a homologous template containing the desired change for HDR, we tested two single strand donor DNA: an ssDNA (120 bp) and a chemical modified donor PS-ssDNA (77 bp). The PS-ssDNA presents symmetric HA corresponding to the locus, while the ssDNA is designed with ~70 bp of HA to the left and ~47 bp to the right of the cleavage site (**Figure 11A**). TIDE calculates a higher editing efficiency in the sample in which the modified donor DNA was used (**Figure 13D**).

To detect the presence of allele A in pools of edited cells, we used ddPCR for both combinations of cells edited with eSpCas9(1.1) and sgRNA_C. What emerged from this experiment was that Cas9 is not only cutting with a higher efficiency in the sample with PS-ssDNA, but also that signal from FAM_A probe corresponding to the minor allele is higher compared to sample in which ssDNA was used. In our case, assessment of positive events with a combination of allele-specific hydrolysis probes for rare events detection with ddPCR detects HDR events without measuring NHEJ, since NHEJ would likely cause indels that prevent the G specific probe from binding even in absence of HDR. Conscious about the limit of the strategy, we decided to speed up the screening by doing sib selection in which DNA of sub-population of 5 cells/well were analyzed and tested by ddPCR. Wells positive for FAM_A signal in ddPCR were then amplified by PCR, sent for Sanger sequencing and analyzed using TIDE. This allowed the identification of an insertion of 45 bp of donor DNA (harboring the minor allele A) after the cleavage site as the origin of the positive A signal detected by ddPCR. Months after I obtained this result, Mijaoka et al. described how the limit of low discrimination of NHEJ events in ddPCR can be overpassed by an allele-specific combination of fluorescent signals which results in droplets that contain either HDR, NHEJ, or wt alleles, or combinations of these; these droplets in turn occupy distinct locations on the two-dimensional plot, allowing absolute quantification of discrete alleles (Miyaoka et al., 2018). As the previously discussed experiment was run only once, we must avoided speculations regarding the causes of knock-in *via* NHEJ of PS-ssDNA. The most plausible scenario is the formation of secondary structures of donor PS-ssDNA which turn into the insertion of a portion of the donor.

After the first test resulted in a NHEJ event, we started thinking about alternative strategies of editing. We observed that the efficiency of the deletion of 731 bp including the locus of interest was very high. Indeed, the lower band corresponding to the edited profile of cells was more intense compared to other combinations of sgRNAs (**Figure 5A**

and **B**). Taking into account this positive result, we decided to go through an alternative strategy of editing of a single nucleotide known as "Obligare" based on the idea of "forcing" cells to do precise insertion of a DNA segment via knock-in (Maresca et al., 2013). The strategy was to use clones positive for the deletion of 731 bp and to design a new sgRNA targeting the new sequence defined by the re-enjoining of the extremities. For such clones the ideal donors DNA were the PCR products corresponding to the deleted portion of DNA (731 bp), with both alleles (A/A or G/G). Testing for the identification of the perfect clone for the "Obligare" strategy reveled that a perfect re-end joining of two extremities without indels were very rare. We found clones positive for the deletion of the locus, but we didn't find heterozygous or homozygous clones with perfect re end joining of two extremities in both alleles without large indels. The most frequent scenario was the one in which one allele had the perfect re-end joining while the other had either an insertion of ~100 bp or an inversion or a translocation (**Figure 6**). We could have modified the "Obligare" method to use two sgRNAs targeting different re-end joining events on the 2 alleles, but we preferred to consider alternatives.

One of the most appealing alternatives was single nucleotide editing with BE method. Right when we were considering this strategy, several papers were published with new variants designed to increase the specificity and the sensitivity of the system and to resolve issues related to the delivery of the complex, the editing of undesired (same) nucleotides in a window of up to 9 bp, the PAMs flexibility and the type of nucleotide modifications. Aware of the challenges of the BE system, we tried to use APOBEC1-xCas9 (D10A) by targeting the antisense region, but we encountered similar problems to Zafra et al. (Zafra *et al.*, 2018) during the cloning step (no data are shown in this work). In parallel, we were working with the RNPs based method, which soon led to successful results for our region, prompting us to move on with that solution while abandoning the BE method. Despite our specific problems with the cloning of the lentivirus and the challenges that BE presents regarding the substitution of a single nucleotide, we recognize that fascinating work with new variants is emerging so fast to strongly support the power of BE and to widely impact the research community (Tan *et al.*, 2019).

At this point, the RNP system delivered by nucleofection was considered the best option for a successful editing. First of all, the Nucleofector 4D system can ensure a high-efficiency of delivery; second, with this method antibiotic selection for the plasmids carrying the sgRNA can be avoided, eliminating the risk of this affecting the transcriptome of edited cells. Third, Cas9 activity is limited due to short-term stability of the RNP which is catabolized by the cell itself, quick and transient activity of Cas9 protein can avoid re cutting of the locus once HDR occurs as it was observed in the reporter system test (**Figure 10**). By using synthetic sgRNAs, we avoided intracellular transcription of the guide and eliminated possible events of integration of the plasmid harboring the sgRNA expression cassette (Liang et al., 2015). Furthermore, RNP complexes seem to be less toxic than transient plasmid transfection. Taking into account all these positive aspects of the RNPs system, we used synthetic sgRNA_B and sgRNA_C (20 bp each) and two donor DNAs from previous experiments (ssDNA and PS-ssDNA) in order to reach our main goal. In addition, electroporation enhancer and HDR molecules were added to the mixture of transfection as they were suggested to increase the efficiency of transfection and editing. Each combination reported in **Table 5** was analyzed with TIDE, TIDER, ddPCR and targeted sequencing. Except for TIDER, all the methods of analysis we applied led to more or less the same frequencies of events and helped us to reach these conclusions:

Samples without donors are less edited by Cas9 than samples where donors were used, suggesting that oligos (ssDNA or PS-ssDNA) are bursting the delivery of the mixture carrying Cas9 protein and sgRNAs (Results of TIDE in **Figure 14**) (Richardson *et al.*, 2016).

The sgRNA_C performed differently in terms of cutting efficiency when provided to the cell within a plasmid (to then be transcribed) or as already mature RNA (synthetic guide). Note that the only difference was an extra G at the 5' end in the plasmid case (**Figure 12** and **Figure 13**).

Combinations including HDR molecules showed lower frequency of NHEJ events compared to combinations without HDR molecules (**Figure 14B**). These results point to the importance of controlling cell repair machinery in order to increase HDR events.

Although we have indications that PS-ssDNA, HDR molecules and cutting efficiency contribute to HDR events, our data do not allow us to disentangle the absolute contribution of each of those variables. Based on anecdotal examples, it appears that the higher the Cas 9 activity, the higher the chance to induce HDR events.

Altogether we can conclude from these results that the use of RNPs associated with particularly efficient delivery methods (Nucleofector 4D) provided us higher chance to succeed and we recommend it as the election system for editing in prostate cells. Prior to the screening of the clones, information such as frequency of indels or HDR events are important and can shape the decision of the best screening strategy; however, since neither TIDE nor ddPCR are able to discriminate large indels, targeted sequencing of the edited genomic region is the only tool that can give both a qualitative overview of the nature of the events and an accurate quantification of their frequencies.

4.2 Alteration of the region including 7p14.3 is recapitulated in a deregulation of the transcriptome of prostate cell line.

GWAS provides a powerful approach to identify polymorphic loci associated with increased susceptibility to cancer. GWAS studies nominated variants in both protein coding and non-coding areas of the human genome, with the former offering more straightforward understanding of their molecular effect, whereas the latter resulting more complex to be deciphered. Altogether, the identification and characterization of inherited variants associated with increased cancer risk could unravel relevant molecular mechanisms by which they exert their biological functions in cancer evolution and progression.

The rs1376350 resides in a noncoding region of the human genome and is associated with an early somatic event in PCa, *SPOP* mutation. Few papers (Gao *et al.*, 2018; Hua *et al.*, 2018) provided a contribution in terms of functional validation of germline variants residing in noncoding regions associated with PCa. They demonstrated that a functional region can act as an enhancer or promoter depending on a germline variant, recognized and bound with allele dependent affinity by TFs implicated in the regulation of nearby genes. In contrast to these two studies, our work is based on the functional characterization of a germline variant associated with a specific subclass of PCa that is molecularly defined. Of note, genetic data from PCa patients coupled with their transcriptomes didn't identify a role for rs1376350 on nearby genes, but rather a broader pleiotropic role on the transcriptome. These peculiarities make the validation of our hypothesis as interesting as difficult.

The enhancer activity of 7p14.3 locus has been supported with Luciferase assay (Romanel *et al.*, 2017) and with ChIP-qPCR analysis (in PC-3_LvpAIP_AR). ChIP-qPCR showed that H3K4me3 enrichment (characteristics of promoter regions) at 7p14.3 is lower compared to the enrichment of H3K4me1, marker of enhancer region (**Figure 19A**) (Heintzman et al., 2007; Cauchy, Koch and Andrau, 2017). In addition, H3K27Ac on isogenic cells with homozygous genotype for the major (C1_2 (G/G)) and the minor (C5_28 (A/A)) allele upon AR overexpression showed higher activity in presence of allele A than with allele G (**Figure 19B** and **C**). The low methylation state of H3K27 in C5_28 compared to C1_2 further confirmed the enhanced activity of the locus. The ratio of H3K27Ac to H3K27me3 enrichment to the locus is similar, independently from AR activation with DHT (**Figure 19B** and **C**). It would be interesting to investigate the state of the histone markers in heterozygous cells as well and to check if one allele is sufficient to determine the fate of the enhancer state or if the homozygous state for A is necessary. Further, an assessment in another prostate cell type such as LNCaP (expressing endogenous AR) of the minor and major alleles effect will help to elucidate the real implication of AR on histone modification and enhancer activation or silencing.

The deregulation of transcriptome of clones positive for macrodeletion further confirmed the functional activity of the region exemplified by the deregulation of more than 700 genes. The number of deregulated genes shared between the macrodeletion and the disrupted motifs cells are significant throughout the groups. Of note, the number of DEGs is much lower in clones positive for the single nucleotide editing suggesting that not only the locus has an enhancer activity but that the germline variant is sufficient to trigger a specific deregulation that encompass more than 100 genes across the genome. A lower number of DEGs in the clone heterozygous for the SNP (C1_33 (A/G)) might suggest a potential dosage effect of the minor allele although technical and biological replicates are needed to properly test it. Altogether to further support these promising results it is necessary to increase the number of samples, by including further clones selected positive for the allele A and to control for possible clone specific DEGs. When we interrogated the profile of genes along chromosome 7 in cells edited for the macrodeletion, we observed a deregulation of some of those genes, which is not observed in cells with disruption motifs and cells with the minor allele. However, these data already indicate that to characterize a functional variant the more appropriate option is in fact the use of an in vitro or in vivo model carrying the variant (plus adequate control models). Alternatively, if this is not possible, strategies supported by genome

editing techniques including the induction of small indels at the locus of interest by Cas9 endonuclease or the masking of the binding motif with dCas9 can help to elucidate the functional activity of the locus (Shukla and Huangfu, 2018).

Of particular interest, the ELOVL2 gene, located on chromosome 6, is upregulated in both clones presenting the disrupted motifs and in clones with the minor allele, including upon AR over-expression. ELOVL2 codifies for fatty acid elongase 2 enzyme implicated in the Polyunsaturated fatty acid elongation (PUFA) synthesis (Jakobsson, Westerberg and Jacobsson, 2006). It has been shown that in PCa patients with *SPOP* mutation, there is an upregulation of the transcriptome of ELOVL2. When we interrogate human genotype/transcript data, an upregulation of the transcriptome in patient is noticed while there is no evidence of deregulation of transcriptome in healthy tissues. This data suggests a possible implication of the germline variant in the lipid biosynthesis, precursors of androgens and estrogens hormones.

All tumor cell line models present a set of limitations. Specifically, the PC-3 cell line used for the functional characterization of rs1376350 is a fully transformed cell line, established from a bone marrow metastasis exhibiting a complex karyotype. This implies that the system is not representative of the early disease status in which we expect the variant to fully exert its role. A better model could have been a primary epithelial cell, but working with primary cells present many technical challenges and limitation in terms of versatility and experiment duration. Further characterization of the variant should be performed in other PCa cell lines with different genetic background and molecular phenotypes; as previously mentioned, this will include the LNCaP (derived from prostate cancer metastatic site; lymph node), AR positive cells for which we already generated transcriptomic data from pools of macro- and microdeleted cells. We are also aware that any single model does not recapitulate the diversity of human tumors (Aurich-Costa et al., 2001; Gao and Chen, 2015; Namekawa et al., 2019).

Despite the weaknesses of our model described above, it is interesting to point out some silver lining of the system that can encourage other studies aimed at similarly investigating functional regions through genome editing technology. For example, if we consider that clones selected positive for the editing of the locus are obtained with different editing formula (transient transfection, stable expression of Cas9 and RNPs, **Supplementary 7, Table 4A** and **B**), the intersection of the deregulated transcriptome can't be a product of off targets, but likely the result of the activity of the locus cooperating towards an yet unknown mechanism. Moreover, when we interrogate the profile of these DEGs in human data (healthy and tumor samples), a fraction of those genes validates the profile detected in edited cells (**Figure 17C**). For example, in human tumor data, AFF3, ELOVL2, NPY2R, LAMC3 and BST2 transcripts in the presence of the genotypes (A/A) and (A/G) recapitulate the scenario of transcriptome levels in isogenic cells with minor allele.

To better understand the implication of the locus in the deregulations of the transcriptome is important to find the protagonists cooperating together with the 7p14.3.

Traditionally, regulatory elements are classified as promoters or enhancers depending on their genomic locations, associated histone marks and regulatory effects. An additional profile of promoter activity is described through production of short bidirectional RNAs centered around transcription start sites (TSSs) (Heintzman et al., 2009). Of interest, it was shown that some intergenic and intragenic regions including enhancers may also be transcribed to produce short bidirectional transcripts, known as eRNAs. However, the function of these eRNAs transcribed in enhancer regions remains obscure in terms of regulatory transcriptional activity. Gene expression requires successful communication between enhancer and promoter regions, whose activities are regulated by a variety of factors and associated with distinct chromatin structures. To the classical scenario of physically interaction of enhancer–promoter by DNA looping and recruitment of transcriptional machinery, extragenic non-coding eRNAs produced by enhancer are adding another feature to the complex transcriptional machinery. This scenario is still under debate, since single cell analysis showed an irrelevant role of eRNAs in gene activation (Rahman et al., 2016). Even though we don't have information about the implication of our locus in this direction, TRANSFACT analysis supports the classical approach with AR and CEBPβ as plausible players (**Supplementary Table 1**). To interrogate the binding of these two TFs to the locus, we need to perform ChIP-qPCR experiments in our brand-new isogenic cells (I would like to point out that we obtained the cell lines harboring the alternative allele not earlier than March 2019). While we were working on generating our PC-3 with the variant, we designed a pull down strategy to check if 28 bp of oligo harboring the minor or the major allele were binding AR and/or CEBPβ in PC-3_LvpAIP_AR (DHT). The assay demonstrated that in this system neither in the presence of allele A nor in presence of allele G, AR binds the oligo of 28 bp. AR allele specificity binding is instead detected when we change a G with an A 2 bp upstream than the rs1376350 position (inserting a half site motif for AR) or when probes with abrogated motif for CEBPβ are added to the assay. These results suggest that a

physical interaction of AR and putatively CEBPβ exists, but we yet don't know if this interaction occurs directly at the locus or whether other co-factors are involved.

Computational methods have been developed to scan TF motifs across the entire genome to predict the capacity of a TFs to bind specific regions. However, the presence of TF motifs does not imply that the TF is actually capable of binding this region. For example, prediction of AR to ARE elements occur in a few million times throughout the human genome. However, only hundreds to tens of thousands AREs appear to be functionally active in a given context and occupied by AR in prostate cell lines (McNair et al. 2017, Stelloo et al. 2018) and tissue samples (Sharma et al. 2013, Chen et al. 2015, Pomerantz et al. 2015, Stelloo et al. 2015, 2018). Moreover, ChIP-seq data for AR in tumor samples and cell lines showed a different pattern of AR activity suggesting that although AREs are the same, AR acts differently from one cell type to another (Yang et al., 2018). Beyond the prediction of AR to sit into the locus, we might underestimate the role of CEBPβ which binds the locus independently from AR occupancy. We still don't have data for the binding of CEBPβ to the locus but we can speculate about the competition of these two TFs for the binding site flanking the SNP. In absence of ChIP-qPCR analysis for AR and CEBPβ in isogenic cells harboring the minor allele the results from the pull down assay are suggestive of an antagonism between AR and CEBPβ, but are inconclusive.

5 Future Perspective

It is fundamental to discern whether the functional germline variant within TFs binding motifs can directly affect gene expression and in cooperation with other functional variants or under epigenetic stimulation can facilitate the emergence of mutations in SPOP; or if in men harboring the inherited variant, SPOP mutation occurs independently and the two components then drive tumorigenesis and provide cells with fitness advantage. Towards this, next steps will include induction of SPOP mutation phenotype in isogenic cells harboring the minor allele to then query cells properties and cells transcriptomes. In addition, we will test the DNA repair machinery with and without SPOP mutation in A/A and G/G lines to measure NHEJ versus HDR upon DNA damage.

3C analysis experiments could be performed to confirm the physical interaction between our polymorphic locus and the genes which expression resulted frequently altered in the RNA-seq data upon the genomic perturbations we introduced. More broadly, one could envision Hi-C or capture Hi-C experiments with each of the A/A and the G/G genotype to agnostically catalog all shared and unique physical interactions genome-wide.

Mass Spectrometry on proteins enriched by using DNA oligos containing the rs1376350 as bait can help to elucidate which are the actors recruited to our locus and then frame the picture of the mechanism of action of rs1376350 with other proteins in *in cis* or *in trans* model. As of today, we do not have enough information supporting our initial hypothesis of AR having a direct role in driving tumorigenesis by binding with differential affinity to rs1376350. Nonetheless we cannot exclude an indirect and crucial role of other steroid hormone metabolites or analogues as related to the significant upregulation of ELOVL2 in the transcriptome of both patients and edited for the A/A or A/G genotype. The broader computational analysis on the genomic sequence features of the locus compatible with AR and CEBP family members binding and with polymorphic sites along the human genome (Davide Dalfovo in silico work) further supports the implication of these components with SPOP mutant prostate cancer and with PCa carcinogenesis.

6 Acknowledgment

I like to specifically acknowledge the work of Orsetta Quaini for the preparation of the RNA sequencing libraries, Francesca Lorenzin for the ChIP-qPCR experiments, and Paola Gasperini for the Biotinylated DNA pull-down assays. Contributors of the computational work were acknowledged when discussed.

I like to acknowledge the input regarding CRISPR/Cas9 technology from the Laboratory of Anna Cereseto.

The work included in this PhD thesis was partially supported by an AIRC Investigator Grant (IG 19221).

7 Appendix

7.1 Biotinylated DNA pull-down assay

In silico analysis identified 2 potential transcription factor binding motifs overlapping the rs137635 locus. To test this *in silico* prediction *in vitro,* we set up a Biotinylated DNA pull-down assay using as baits for AR 28 nucleotides long probes biotinylated at the 5' to encompass our SNP of interest carrying the major allele G or the minor allele A. A sequence of the canonical DNA full site binding sequence for AR (M08908 transfac) and a sequence lacking AR binding sites as predicted by a tool developed in house (by Davide Dalfovo) were used as positive and negative control, respectively (**Supplementary Figure 1A-C, Table 1** and **2**).

Unexpectedly, when nuclear protein extracts of PC-3 stably overexpressing AR treated with DHT 100 µM were precipitated with allele A or allele G probes, we did not observe any significant difference in AR enrichment while our AR canonical full binding site region was very efficient (**Supplementary Figure 1E**, lanes 1-2 vs C+). To assess whether CEBPβ binding to rs137635 could prevent AR allele specific binding, we mutated nucleotide 18th and 24th of our probes in order to abrogate CEBPβ binding to the locus. Again, we did not observe any significant difference in AR enrichment (**Supplementary Figure 1E,** lanes 3-4). To rule out the possibility that the essay was not sensitive enough to precipitate AR from an AR binding half site, we mutated the 4th nucleotide of the M00962 motif from G to A, in the presence of our SNP A or SNP G (in 6th position) as in motives M08908. With these probes (**Supplementary Figure 1E**, lanes 5-6), we did observe a significant difference in AR enrichment when allele A was present instead of allele G. Interestingly, probes that combined the mutations for abrogating CEBP binding and increasing AR binding did recruit more AR than with probes 3-4, but showed less allele specific binding compare to probes 5-6 suggesting that CEBPβ prevents recruitment of to the probe. These conclusion are further consolidated by the results obtained from the DNA pulldown performed with proteins extracted from PC-3 overexpressing AR and lacking CEBPβ (Knockout with CRISPR/Cas9) in which recruitment of AR is increased (**Supplementary Figure 1E**, lanes 7 and 8 vs 5 and 6).

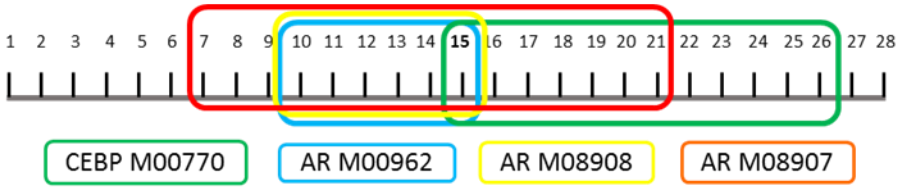
Since there is high overlap between AR and GR-1 binding motifs, we assesses GR-1 presence in the precipitated protein nuclear extracts bound to the same biotinylated probes (1-8). We treated PC-3 cells that endogenously express GR1 for 6 hours with dexamethasone to ensure GR1 nuclear translocation prior to nuclear proteins extraction. GR1 immunoblot detection mirrors AR detection suggesting that these nuclear receptors have the same affinity for the tested regions. Interestingly CEBPβ KO does not affect GR-1 binding as it does with AR (**Supplementary Figure 1E,** lanes 7 and 8 vs 5 and 6 in blots 2 and 4).

A

Supplementary Table 1. Score of predicted binding affinity ofr two TFs in the locus

	Ancestral allele G	Risk allele A	Ancestral allele G	Risk allele A
	AR consensus		CEBP family consensus	
La score	1.87	8.74	8.97	8.97
La/Lm score	0.18	0.82	0.77	0.77
p-value	0.08	**0.0009**	**0.0002**	**0.0002**

B

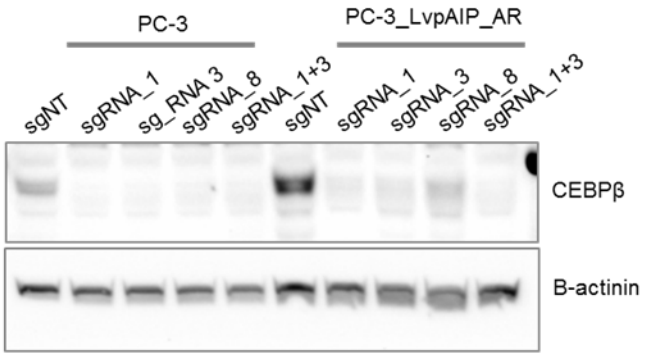


C

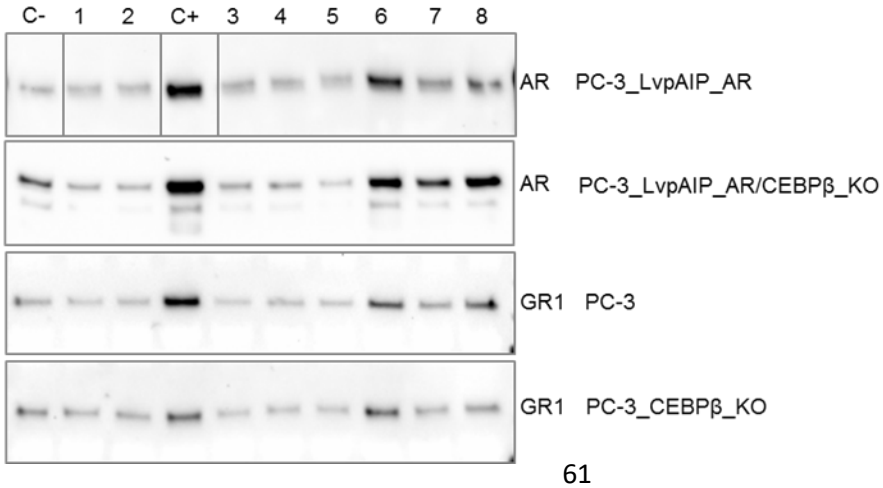
Supplementary Table 2. Oligos design for Pull Down experiment

Type of Modification	Line Band	Modified Sequence
AR_canonical full site	C+	AGT CCC AGA ACA TGA TGT TCT GGA TGG T
scramble	C-	AGC GTG TAA TTT GAC CCG CTA ACT TCT A
7p14.3_G	1	AGT GAA CTG AGA GCG ATT GTG GCA AGG T
7p14.3_A	2	AGT GAA CTG AGA GCA ATT GTG GCA AGG T
CEBP_mut _G	3	AGT GAA CTG AGA GCG ATG GTG GCG AGG T
CEBP_mut _A	4	AGT GAA CTG AGA GCA ATG GTG GCG AGG T
AR_half site_G	5	AGT GAA CTG AGA A CG ATT GTG GCA AGG T
AR_half site_A	6	AGT GAA CTG AGA A CA ATT GTG GCA AGG T
AR_half site+CEBP mut_A	7	AGT GAA CTG AGA A CG ATG GTG GCG AGG T
AR_half site+CEBP mut_G	8	AGT GAA CTG AGA A CA ATG GTG GCG AGG T

D



E



Supplementary Figure 1. **AR and GR-1 binding in presence or absence of CEBPβ on the 7.14.3 locus.** A) Schematic representation of the position of the full sites and half sites binding motifs for AR, GR1 and CEBP family in a 28 bp stretch centered on rs137635 in the 7p14.3 locus. B) DNA sequences of oligos used in a biotinylated DNA pull-down assay to enrich for AR, GR-1 and CEBPβ. In bold the nucleotides mutated in the different probes to abolish the affinity for CEBPβ or to increase the affinity for AR. C) Western Blot for CEBPβ expression in PC-3 cell line with and without AR overexpression after targeting the disruption of CEBPβ with CRISPRCas9 and 4 different sgRNA combinations (sgNT, sgRNA non targeting sequence). Actinin was used as loading control. The combination sgCEBP1+3 was used in the following experiment identified as CEBP KO. D) DNA-pull-down assays followed by Immunoblots for AR and GR-1. Nuclear proteins used for AR pull down were extracted from PC-3 cell line overexpressing AR (PC-3LvpipAR) and from cell line overexpressing AR where CEBPβ was knocked out (PC-3LvpipAR_CEBPβ KO) after stimulation with DHT for 16 hours. Nuclear proteins used for GR-1 pull down were extracted from PC-3 cell line (PC-3) and from PC-3 where CEBPβ was knocked out (PC-3_CEBPβ KO) after stimulation with Dex for 6 hours.

Lanes C-, proteins precipitated by biotinylated probe of the non-specific region; Lanes C+, proteins precipitated by biotinylated probe of the AR canonical full binding site; Lanes 1–2 proteins precipitated by biotinylated probes of the regions surrounding rs137635, Lanes 3-4 proteins precipitated by biotinylated probes as in 1-2 but with 2 nucleotides mutated to abrogate CEBPβ binding; Lanes 5-6 proteins precipitated by biotinylated probes as in 1-2 but with 1 nucleotide mutated to enhance AR binding; Lanes 7-8 proteins precipitated by biotinylated probes as in 5-6 but with 2 nucleotides mutated to abrogate CEBPβ binding.

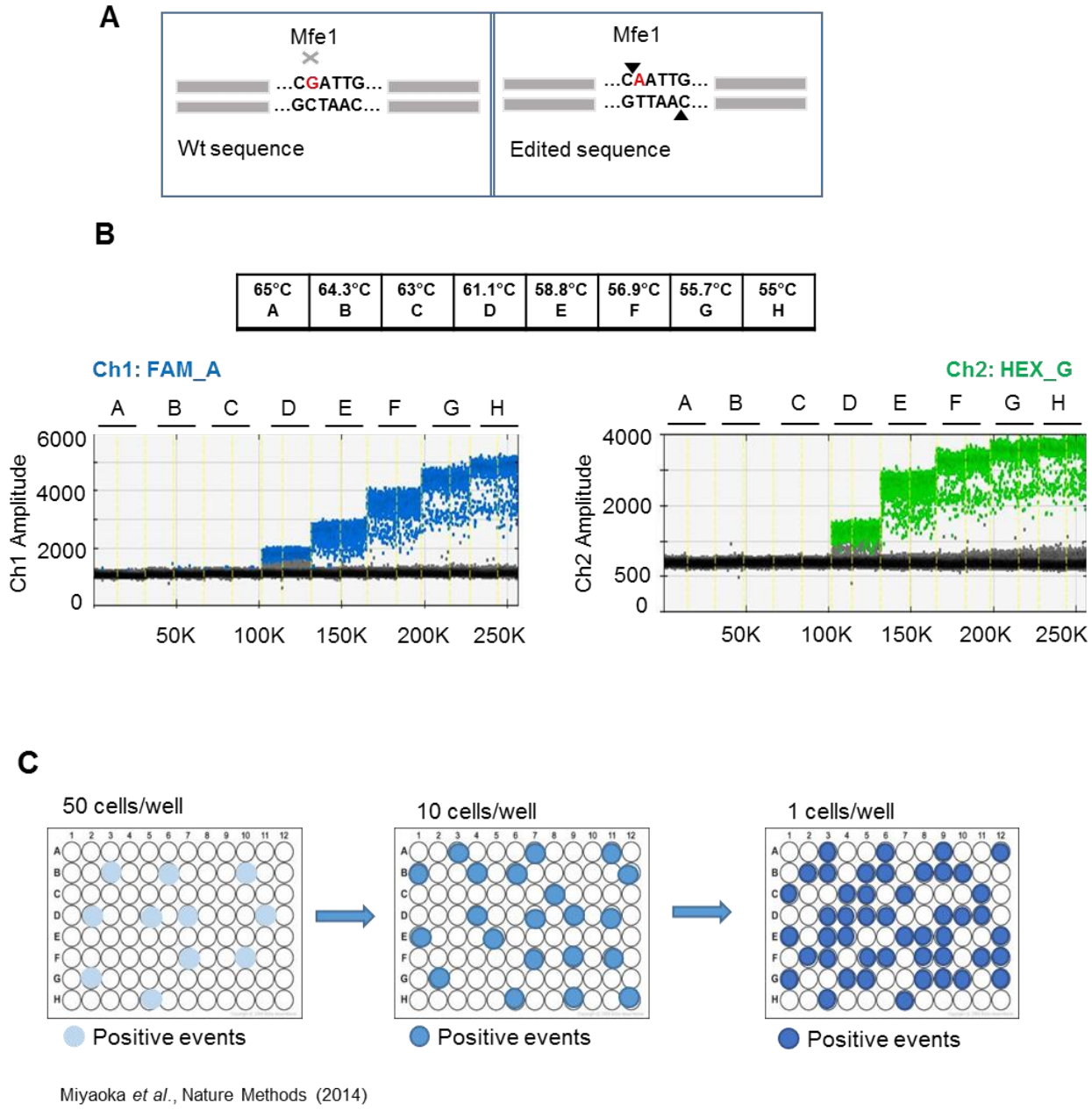
7.2 Supplementary 2

Supplementary Table 3. Description of 3 sets of experiements of editing in prostate cells.

	Macrodeletion (731 bp)	Microdeletion (50 bp)	Disruption Motifs
Vector for sgRNAs	pUC19_sgRNA (No selection)	pUC19_sgRNA/px330_sgRNA_Puro resistant	pGS_sgRNA_Neo resistant/px330_sgRNA_Puro resistant
sgRNAs	pUC19_sgRNA (up_2; down_2; up_3; down_3)	pUC19_sgRNA_E/px330_sgRNA_A	pGS_sgRNA_B and C/ px330_sgRNA_A
Vector for Cas9	pSPCas9 (1.1)_Puro resistant	pSPCas9 (1.1)_Puro resistant	LvCas9
Antibiotic Selection	Puromycin	Puromycin	Neomycin and Puromycin

Supplementary Figure 2. Three sets of experiments of editing in prostate cells. The table shows three sets of editing performed with CRISPR/Cas9 system: macrodeletion of 731 bp, microdeletion of 50 bp and disruption motifs. In the two first rows vectors used for the cloning of sgRNAs are shown, third and fourth row lists the type of Cas9 endonuclease and the type of antibiotic selection applied. Experiments for single nucleotide editing performed with RNPs system are not represented in this table

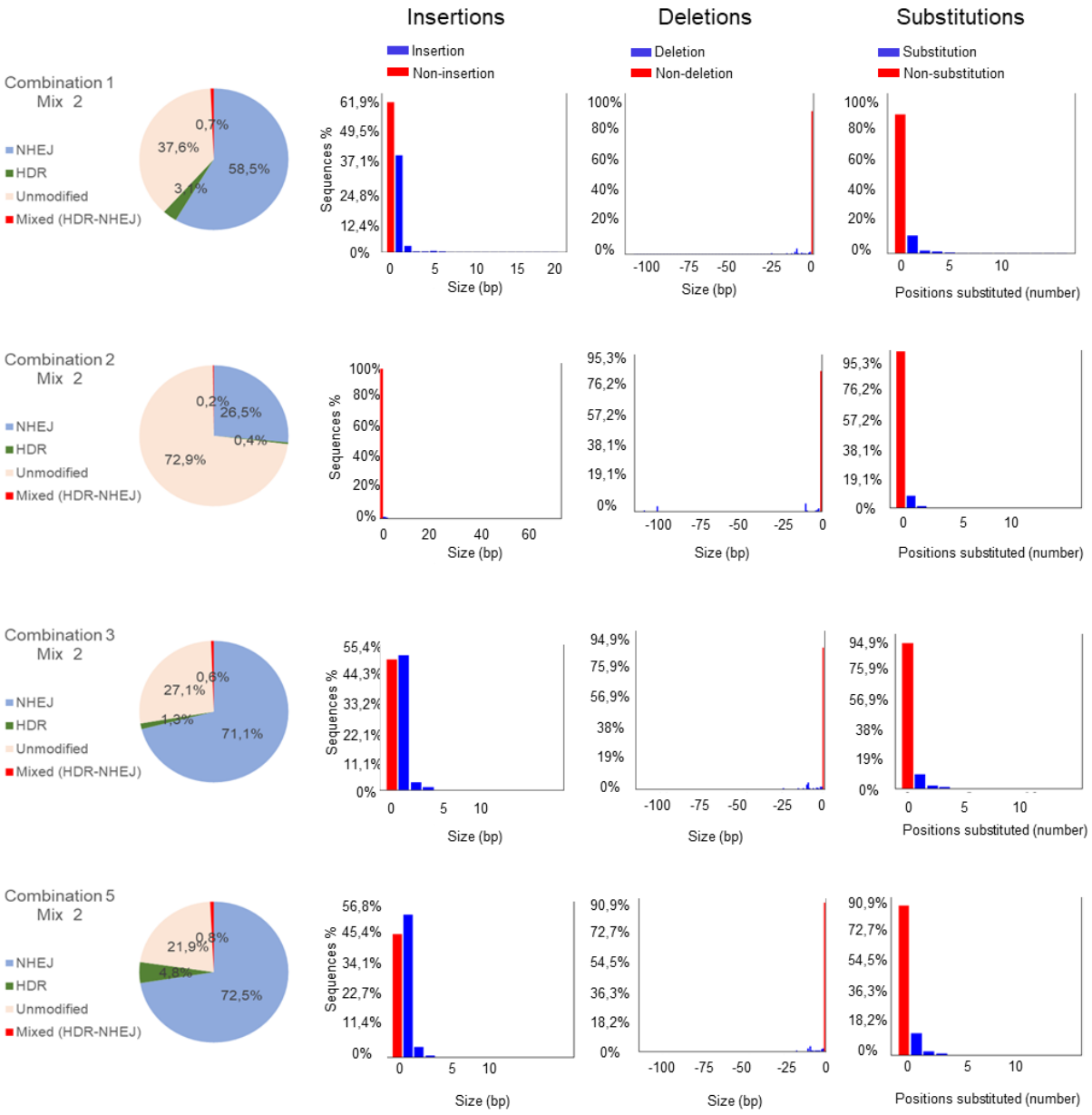
7.3 Supplementary 3



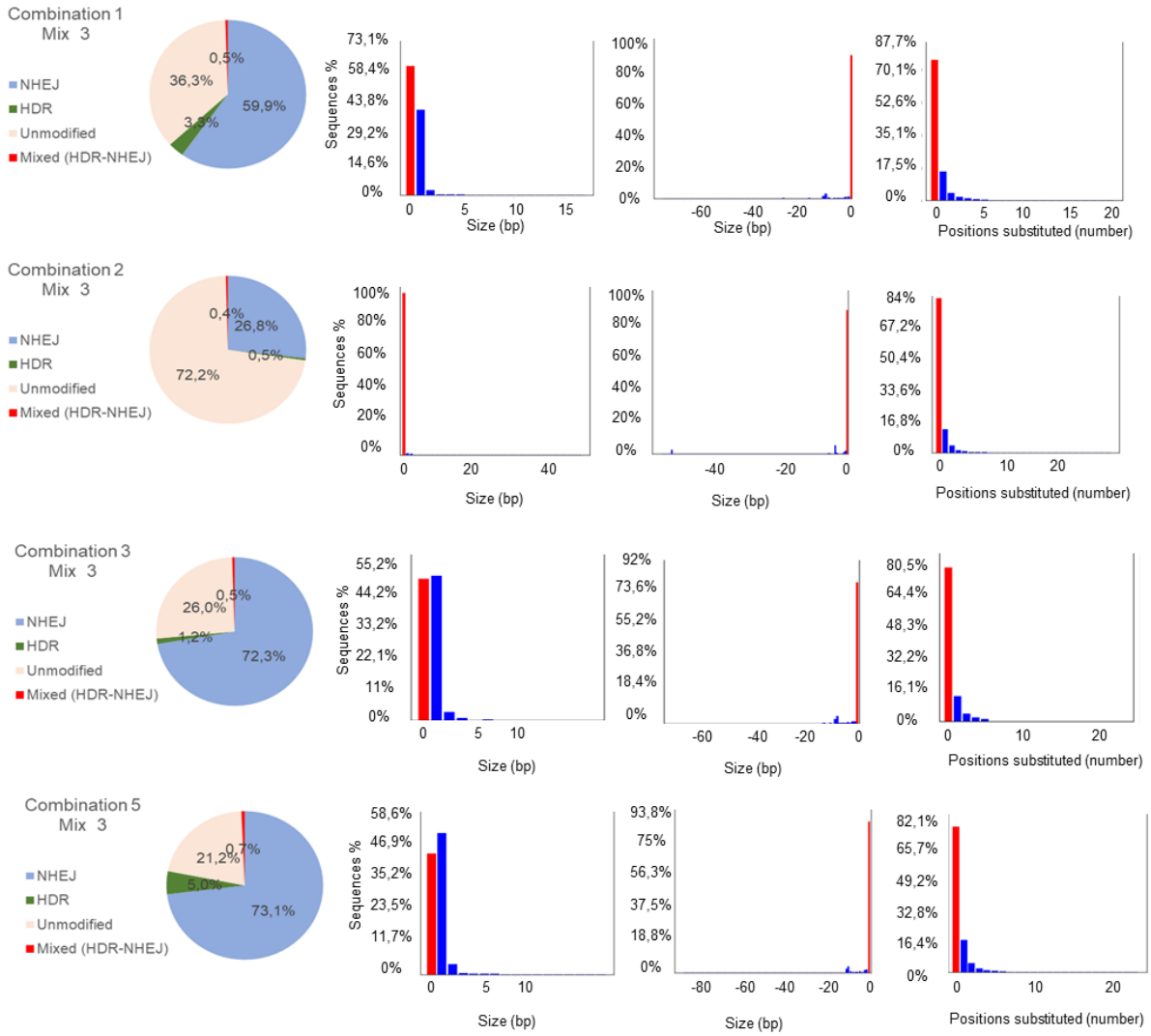
Supplementary Figure 3. Methods of screening for HDR events. A) Graphic representation of MfeI digestion cutting site. If allele G is substituted with allele A via HDR, a MfeI digestion site emerges in the locus. B) DNA extracted from Hs 578Bst (ATCC® HTB-125™) was used to set up conditions of amplification and fluorescence detection from two probes differing for one allele. Sample annealing temperature optimization demonstrate separation of FAM_A (blue dots) and HEX_G (green dots) starting from 58.8°C. C) Sib-selection method of screening. The first 96 well plate is used for the first step of screening in which wells with higher signal will be seed in a new 96 well plate with a lower distribution (10 cells/ well). Therefore, positive wells from this second plate will be selected to be distributed in a new 96 well plate with a density of 1 cell/well. (Miyaoka Y et al., Nat Methods 2014). Colored dots represent positive events for each plate.

7.4 Supplementary 4

A

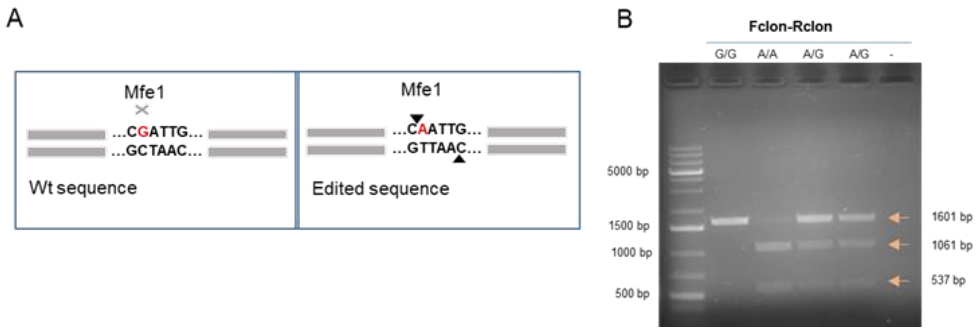


B



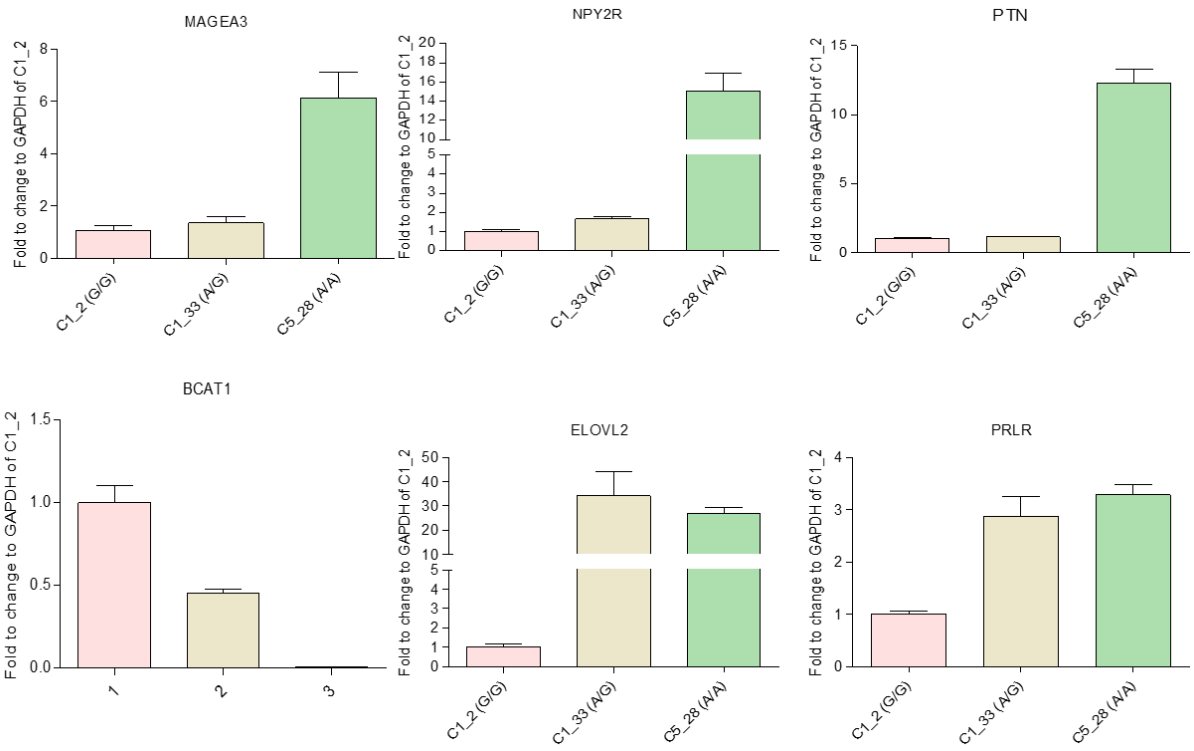
Supplementary Figure 4. Targeted DNA sequencing for Mix 2 and 3. Analysis with CRISPResso in sample 1,2, 3 and 5 for A) Mix 2 (FBS038-RBS039) and B) Mix 3 (FBS040-RBS041). In the left pie charter calculating events of HDR and not HDR while in the right graph shown insertion, deletion and substitution events after the editing for each sample.

7.5 Supplementary 5



Supplementary Figure 5. A) Graphic representation of MfeI digestion cutting site in our locus. B) Validation of editing of single nucleotide for clones positive in ddPCR. PCR product of 1603 bp correspond to not digested sequence while two bands of 1061 bp and 537 bp are the result of digestion in case allele A is present. Clones used for this test are: C1_2 (G/G); C5_28 (A/A); C1_33 (A/G).

7.6 Supplementary 6



Supplementary Figure 6. Biological replication of RT-qPCR validation of RNA-seq analysis in single nucleotide editing clones. Validation of six genes in RT-qPCR. The RNA was extracted from the same samples but in different days. Samples were analyzed in triplicates as described in Methods. Fold changes is calculated based on GAPDH gene expression in C1_2 (G/G) cells.

7.7 Supplementary 7

Supplementary Table 4A. Summary of samples characterized with RNA-seq

PC-3 Cells	Single Clone	Single Clone	Single Clone	Pool	Pool
Type of Editing	Macro Deletion	Disruption Motifs	Single Nucleotide	Macro Deletion	Micro Deletion
N Controls	1	2	1	1	-
N Samples	3	8	2	1	1
Notes	C15, C16, C32	Organized in 4 groups: AR(-/-)CEBPB(-/-), AR(+/+)CEBPB(-/-), AR(+/-)CEBPB(-/-), AR(+/+)CEBPB(+/-)	C1_2 (G/G), C1_33 (A/G), C5_28 (A/A)		ctrl as in pool macro
Name	Macro	DisMotif_ARneg_CEBPBneg, DisMotif_ARpos_CEBPBneg, DisMotif_ARhet_CEBPBneg, DisMotif_ARpos_CEBPBhet	SNP_GG, SNP_AG, SNP_AA	MacroPool	MicroPool
Cas9 Transfection	Transient	Stable	RNPs	Transient	Transient
Antibiotics Selection	Puromycin	Puromycin, Neomycin	--	Puromycin	Puromycin

Supplementary Table 4B. Summary of samples characterized with RNA-seq upon AR overexpression

Name	SingleNucl_GG_AR_EtOH/DHT, SingleNucl_AA_AR_EtOH/DHT	MacroPool_AR_EtOH/DHT
AR Transfection	Stable (LvPAIP_AR)	Transient (pCMV-AR24Q)

Supplementary Figure 7. Summary of samples characterized with RNA-seq. A) List of samples used for transcriptome analysis listed by type of experiment; macrodel; microdeletion; disruption motifs and single nucleotide editing; pool or single clones analyzed; type of editing and antibiotic selection. B) Transcriptome of clones from SNP experiment, C1_2 (G/G) and C5_28 (A/A) stable overexpressing AR upon EtOH or DHT treatment was analyzed, result reported in **section 3.4** and **Figure 18**.

7.8 Supplementary Tables (5-14)

Supplementary Table 5. Oligos sequence corresponding to sgRNAs

Locus	Name	Forward (5'----3')	Reverse (5'----3')
7p14.3	sgRNA_A	GCGATTGTGGCAAGGTCTTT	AAAGACCTTGCCACAATcGC
	sgRNA_B	gAAGTGAACTGAGAGCGATTG	CAATCGCTCTCAGTTCACTTc
	sgRNA_C	GAACTGAGAGCGATTGTGGCA	TGCCACAATCGCTCTCAGTTC
	sgRNA_E	CATTTAACACAACTGGCTTC	GAAGCCAGTTGTGTTAAATG
	sgRNA_ up(2)	GAGGCAGTCAGAAAATCTAG	CTAGATTTTCTGACTGCCTC
	sgRNA_ down(2)	GAATGCTGAAAGATAGGATG	CATCCTATCTTTCAGCATTC
	sgRNA_ up(3)	GAGAGGCGGTTTGAAGAACC	GGTTCTTCAAACCGCCTCTC
	sgRNA_ down(3)	GACTCTGAATAGCTGGCACA	TGTGCCAGCTATTCAGAGTC
CEBPβ	sgRNACEBPβ_1	GCTCTTCTCCGACGACTACGG	CCGTAGTCGTCGGAGAAGAGC
	sgRNACEBPβ_3	GTGCTGCGTACGGCGGCAAGG	CCTTGCCGCCGTACGCAGCAC
	sgRNACEBPβ_8	GTGAAGTCGATGGCGCGCTCG	CGAGCGCGCCATCGACTTCAC
	sg_Non targeting	gCACTACCAGAGCTAACTCA	TGAGTTAGCTCTGGTAGTGc

Supplementary Table 6. Number of Off Targets caluclated with Cas-OFFinder for the final set of guides

sgRNA sequence (5'---3')	Name	Position	Mismatches	
			≤2	≤5
GAGCGATTGTGGCAAGGTCTTT	sgRNA_A	chr:7 (33692747)	1	928
GAAGTGAACTGAGAGCGATTG	sgRNA_B	chr:7 (33692734)	1	1108
AACTGAGAGCGATTGTGGCA	sgRNA_C(20bp)	chr:7 (33692739)	1	1007
GAACTGAGAGCGATTGTGGCA	sgRNA_C(21bp)	chr:7 (33692738)	1	346
CATTTAACACAACTGGCTTC	sgRNA_E	chr:7 (33692700)	1	1162
CTAGATTTTCTGACTGCCTC	sgRNA_up(2)	chr:16 (56245419)	176	1176
TGTGCCAGCTATTCAGAGTC	sgRNA_down(3)		0	1134

Supplementary Table 7. Oligos used for macrodeletion

Locus	Name	(5'----3')
7p14.3	Fclon	CCAAGTGGCCCAGTATAATCC
	Rclon	AAACGTGCTTCTTCCATGGG
	FBS001	TTTCAGATGTTTGTGCCCCG
	FBS004	TTTCAGGCTGGTGTTACTTT
	FBS005	GATTTTCTGACTGCCTCTCT
	RBS004	AAAGTAACACCAGCCTGAAA
	RBS005	ACATGGAATCAAGTTGTGGA
	RBS007	AAGAGCCTTGCTCATGCTGA

Supplementary Table 8. Oligos used for TOPO-TA cloning and screening

Locus	Name	(5'----3')
7p14.3	FBS005	GATTTTCTGACTGCCTCTCT
	RBS005	ACATGGAATCAAGTTGTGGA
	Fscreening	TCACCACTCAGATCACCACT

Supplementary Table 9. Oligos used for ddPCR

Locus	Name	(5'----3')
7p14.3	FBS009	CTTCAGGTGTAATGTTTCAAG
	RBS009	AGACAGACTTTCATTCCTTC
	FAM_A	TGAGAGCAATTGTGGCA
	HEX_G	TGAGAGCGATTGTGGCA

Supplementary Table 10. Targeted DNA sequencing oligos

Locus	Name	(5'----3')
7p14.3	Comon Forward	TCGTCGGCAGCGTCAGATGTG
	overhang	TATAAGAGACAG
	Comon Reverse	GTCTCGTGGGCTCGGAGATGT
	overhang	GTATAAGAGACAG
	FBS0036	CATCTGCAGCTGCTCATACA
	RBS0037	CTTGCCCTCTCTGTTATTGC
	FBS0038	GTTCTTCAAACCGCCTCTCAC
	RBS0039	TTGCTCATGCTGAAAGACAGAC
	FBS0040	GGAAAAATGTCAGAGGAGGAAA
	RBS0041	TGGCATTCCATGCCTTGTAG

Supplementary Table 11. Oligos used for validation in RT-qPCR of selected genes from RNA-seq analysis

Locus	Name	Forward (5'----3')	Reverse (5'----3')
chr:7_PTN	PTN_chr:7_RT_qPCR	CCTGGGGAGAATGTGACCTG	TGAGGTTTGGGCTTGGTCA
chr:6_ELOVL2	ELOVL2_chr:6_RT-qPCR	ACAGCCGCTGCGGATCAT	CCCAGCCATATTGAGAGCAGA
chr:5_PRLR	PRLR_chr:5_RT_qPCR	AGACCATGGATACTGGAGTA	GGAAAGATGCAGGTCACCAT
chr:12_BCAT1	BCAT1_chr:12_RT_qPCR	CCAAAGCCCTGCTCTTTGTA	TGGAGGAGTTGCCAGTTCTT
chr:4_NPY2R	NPY2R_chr:4_RT_qPCR	GGAAACGATTGCCAACTATACGA	GGCCCACTGAGTGTTGAGGA
chr:x_MAGEA3	MAGEA3_chr:x_RT_qPCR	GTGAGGAGGCAAGGTTCTGA	GGGCAATGGAGACCCACT
chr:12_GAPDH	GAPDH_RT-qPCR	TCCAAAATCAAGTGGGGCGA	AGTAGAGGCAGGGATGATGT

Supplementary Table 12. Oligos used for Pull down experiement

Name	Forward (5'----3')	Reverse (5'----3')
AR_canonical full site	AGTCCCAGAACATGATGTTCTGGATGGT	ACCATCCAGAACATCATGTTCTGGGACT
7p14.3_G	AGTGAACTGAGAGCGATTGTGGCAAGGT	ACCTTGCCACAATCGCTCTCAGTTCACT
7p14.3_A	AGTGAACTGAGAGCAATTGTGGCAAGGT	ACCTTGCCACAATTGCTCTCAGTTCACT
CEBPβ_mut_G	AGTGAACTGAGAGCGAT*G* GTGGC*G* AGGT	ACCTCGCCACCATCGCTCTCAGTTCACT
CEBPβ_mut_A	AGTGAACTGAGAGCAAT*G* GTGGC*G* AGGT	ACCTCGCCACCATTGCTCTCAGTTCACT
AR_half site_G	AGTGAACTGAGA*A* CGATTGTGGCAAGGT	ACCTTGCCACAATCGTTCTCAGTTCACT
AR_half site_A	AGTGAACTGAGA*A* CAATTGTGGCAAGGT	ACCTTGCCACAATTGTTCTCAGTTCACT
AR_half site_G/CEBPβ_mut	AGTGAACTGAGA*A* CGAT*G* GTGGC*G* AGGT	ACCT*C* GCCAC*C* ATCGTTCTCAGTTCACT
AR_half site_A/CEBPβ_mut	AGTGAACTGAGA*A* CAAT*G* GTGGC*G* AGGT	ACCT*C* GCCAC*C* ATTG*T* TCTCAGTTCACT

Supplementary Table 13. Oligos used for ChIP-qPCR

Name	Forward (5'----3')	Reverse (5'----3')
GAPDH_Promoter_ChIP	TACTAGCGGTTTTACGGGCG	TCGAACAGGAGGAGCAGAGAGCGA
NEUROD2_ChIP	TGCTACTCCAAGACGCAGAA	CGAGAGCGCCCAGATATAGT
FKBP5_ChIP	GGTTCCTGGGCAGGAGTAAG	AACGTGGATCCCACACTCTC
NCL_ChIP	CTACCACCCTCATCTGAATCC	TTGTCTCGCTGGGAAAGG
KLK3_enhancer_ChIP	ATACTGGGACAACTTGCAAACCT	CAGGCTTGCTTACTGTCCTAGATAA

Supplementary Table 14. Sequence of Donors

PS-ssDNA

5'AACATTAGATGTTGGACCCAAAGACCTTGCCACAATTGCTCTCAGTTCACTTGAAACATTACACCTGAAGCCAGTTG3'

ssDNA

5'AATGTCAGAGGAGGAAACCTGTGAACCCATTTAACACAACTGGCTTCAGGTGTAATGTTTCAAGTGAACTGAGAGCAATTGTGGCAAGGT
CTTTGGGTCCAACATCTAATGTTAAAAAAG3'

8 References

Abeshouse, A. *et al.* (2015) 'The Molecular Taxonomy of Primary Prostate Cancer', *Cell*, 163(4), pp. 1011–1025. doi: https://doi.org/10.1016/j.cell.2015.10.025.

Adamo, H. *et al.* (2018) 'Prostate cancer induces C/EBPβ expression in surrounding epithelial cells which relates to tumor aggressiveness and patient outcome', *The Prostate*. doi: 10.1002/pros.23749.

Adli, M. (2018) 'The CRISPR tool kit for genome editing and beyond', *Nature Communications*, 9(1), p. 1911. doi: 10.1038/s41467-018-04252-2.

Ahmadiyeh, N. *et al.* (2010) '8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC', *Proceedings of the National Academy of Sciences*, 107(21), pp. 9742 LP – 9746. doi: 10.1073/pnas.0910668107.

Amin Al Olama, A. *et al.* (2015) 'Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans', *Human Molecular Genetics*, 24(19), pp. 5589–5602. doi: 10.1093/hmg/ddv203.

Andriole, G. L. *et al.* (2012) 'Prostate Cancer Screening in the Randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: Mortality Results after 13 Years of Follow-up', *JNCI: Journal of the National Cancer Institute*, 104(2), pp. 125–132. doi: 10.1093/jnci/djr500.

Baca, S. C. *et al.* (2013) 'Punctuated Evolution of Prostate Cancer Genomes', *Cell*, 153(3), pp. 666–677. doi: https://doi.org/10.1016/j.cell.2013.03.021.

Baca, S. and Garraway, L. (2012) 'The genomic landscape of prostate cancer ', *Frontiers in Endocrinology* , p. 69. Available at: https://www.frontiersin.org/article/10.3389/fendo.2012.00069.

Bae, S., Park, J. and Kim, J.-S. (2014) 'Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases', *Bioinformatics*, 30(10), pp. 1473–1475. doi: 10.1093/bioinformatics/btu048.

Barakat, D. J. *et al.* (2015) 'CCAAT/Enhancer binding protein β controls androgen-deprivation-induced senescence in prostate cancer cells', *Oncogene*. Macmillan Publishers Limited, 34, p. 5912. Available at: https://doi.org/10.1038/onc.2015.41.

Barbieri, C. E. *et al.* (2012) 'Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer', *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 44, p. 685. Available at: https://doi.org/10.1038/ng.2279.

Barbieri, C. E. *et al.* (2013) 'The mutational landscape of prostate cancer', *European urology*. 2013/05/18, 64(4), pp. 567–576. doi: 10.1016/j.eururo.2013.05.029.

Bégay, V., Smink, J. and Leutz, A. (2004) 'Essential Requirement of CCAAT/Enhancer Binding Proteins in Embryogenesis', *Molecular and Cellular Biology*, 24(22), pp. 9744 LP – 9751. doi: 10.1128/MCB.24.22.9744-9751.2004.

Bell, K. J. L. *et al.* (2015) 'Prevalence of incidental prostate cancer: A systematic review of autopsy studies', *International Journal of Cancer*. John Wiley & Sons, Ltd, 137(7), pp. 1749–1757. doi: 10.1002/ijc.29538.

Berger, M. F. *et al.* (2011) 'The genomic complexity of primary human prostate cancer', *Nature*. The Author(s), 470, p. 214. Available at: https://doi.org/10.1038/nature09744.

Blattner, M. *et al.* (2014) 'SPOP Mutations in Prostate Cancer across Demographically Diverse Patient Cohorts', *Neoplasia*, 16(1), pp. 14-W10. doi: https://doi.org/10.1593/neo.131704.

Blattner, M. *et al.* (2017) 'SPOP Mutation Drives Prostate Tumorigenesis In Vivo through Coordinate Regulation of PI3K/mTOR and AR Signaling', *Cancer Cell*, 31(3), pp. 436–451. doi: https://doi.org/10.1016/j.ccell.2017.02.004.

Bluemn, E. G. *et al.* (2017) 'Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling', *Cancer Cell*, 32(4), pp. 474-489.e6. doi: https://doi.org/10.1016/j.ccell.2017.09.003.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Boysen, G. *et al.* (2015) 'SPOP mutation leads to genomic instability in prostate cancer', pp. 1–18. doi: 10.7554/eLife.09207.

Boysen, G. *et al.* (2018) 'SPOP-Mutated/CHD1-Deleted Lethal Prostate Cancer and Abiraterone Sensitivity', *Clinical Cancer Research*, 24(22), pp. 5585 LP – 5593. doi: 10.1158/1078-0432.CCR-18-0937.

Brinkman, E. K. *et al.* (2014) 'Easy quantitative assessment of genome editing by sequence trace decomposition', *Nucleic acids research*. 2014/10/09. Oxford University Press, 42(22), pp. e168–e168. doi: 10.1093/nar/gku936.

Brinkman, E. K. *et al.* (2018) 'Easy quantification of template-directed CRISPR/Cas9 editing', *Nucleic Acids Research*, 46(10), pp. e58–e58. doi: 10.1093/nar/gky164.

Bushman, W. (2009) 'Etiology, Epidemiology, and Natural History', *Urologic Clinics of North America*, 36(4), pp. 403–415. doi: https://doi.org/10.1016/j.ucl.2009.07.003.

Byrne, S. M., Mali, P. and Church, G. M. (2014) 'Genome editing in human stem cells', *Methods in enzymology*, 546, pp. 119–138. doi: 10.1016/B978-0-12-801185-0.00006-4.

C., N. A. and T., D. E. (2013) 'Expression quantitative trait loci: present and future', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society, 368(1620), p. 20120362. doi: 10.1098/rstb.2012.0362.

Calo, E. and Wysocka, J. (2013) 'Modification of enhancer chromatin: what, how, and why?', *Molecular cell*, 49(5), pp. 825–837. doi: 10.1016/j.molcel.2013.01.038.

Carter, H. *et al.* (2017) 'Interaction landscape of inherited polymorphisms with somatic events in cancer', *Cancer Discovery*. doi: 10.1158/2159-8290.CD-16-1045.

Casini, A. *et al.* (2018) 'A highly specific SpCas9 variant is identified by in vivo screening in yeast', *Nature biotechnology*. 2018/01/29, 36(3), pp. 265–271. doi: 10.1038/nbt.4066.

Chatterjee, P., Jakimo, N. and Jacobson, J. M. (2018) 'Minimal PAM specificity of a highly similar SpCas9 ortholog', *Science Advances*, 4(10), p. eaau0766. doi: 10.1126/sciadv.aau0766.

Chen, H. *et al.* (2015) 'Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci', *The Prostate*. John Wiley & Sons, Ltd, 75(12), pp. 1264–1276. doi: 10.1002/pros.23008.

Chen, Q.-R. *et al.* (2014) 'Systematic Genetic Analysis Identifies Cis-eQTL Target Genes Associated with Glioblastoma Patient Survival', *PLOS ONE*. Public Library of Science, 9(8), p. e105393. Available at: https://doi.org/10.1371/journal.pone.0105393.

Chu, V. T. *et al.* (2015) 'Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells', *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 33, p. 543. Available at: https://doi.org/10.1038/nbt.3198.

Cirillo, L. A. *et al.* (1998) 'Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome', *The EMBO Journal*, 17(1), pp. 244 LP – 254. doi: 10.1093/emboj/17.1.244.

Cirillo, L. A. *et al.* (2002) 'Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4', *Molecular Cell*, 9(2), pp. 279–289. doi: https://doi.org/10.1016/S1097-2765(02)00459-8.

Cohen, D. M. *et al.* (2018) 'Shared nucleotide flanks confer transcriptional competency to bZip core motifs', *Nucleic Acids Research*, 46(16), pp. 8371–8384. doi: 10.1093/nar/gky681.

Cong, L. *et al.* (2013) 'Multiplex genome engineering using CRISPR/Cas systems', *Science (New York, N.Y.)*. 2013/01/03, 339(6121), pp. 819–823. doi: 10.1126/science.1231143.

Consortium, T. E. P. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*. The Author(s), 489, p. 57. Available at: https://doi.org/10.1038/nature11247.

Culig, Z. (2016) 'Androgen Receptor Coactivators in Regulation of Growth and Differentiation in Prostate Cancer', *Journal of Cellular Physiology*. John Wiley & Sons, Ltd, 231(2), pp. 270–274. doi: 10.1002/jcp.25099.

Delaneau, O. *et al.* (2019) 'Chromatin three-dimensional interactions mediate genetic effects on gene expression', *Science*, 364(6439), p. eaat8266. doi: 10.1126/science.aat8266.

Demaison, C. *et al.* (2002) 'High-Level Transduction and Gene Expression in Hematopoietic Repopulating Cells

Using a Human Imunodeficiency Virus Type 1-Based Lentiviral Vector Containing an Internal Spleen Focus Forming Virus Promoter', *Human Gene Therapy*. Mary Ann Liebert, Inc., publishers, 13(7), pp. 803–813. doi: 10.1089/10430340252898984.

Descombes, P. and Schibler, U. (1991) 'A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the sam mRNA', *Cell*, 67(3), pp. 569–579. doi: https://doi.org/10.1016/0092-8674(91)90531-3.

Di-Poï, N. *et al.* (2005) 'Transcriptional Repression of Peroxisome Proliferator-activated Receptor β/δ in Murine Keratinocytes by CCAAT/Enhancer-binding Proteins', *Journal of Biological Chemistry* , 280(46), pp. 38700–38710. Available at: http://www.jbc.org/content/280/46/38700.abstract.

Dobin, A. *et al.* (2012) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Edwards, J. and Bartlett, J. M. S. (2005) 'The androgen receptor and signal-transduction pathways in hormone-refractory prostate cancer. Part 1: modifications to the androgen receptor', *BJU International*. John Wiley & Sons, Ltd (10.1111), 95(9), pp. 1320–1326. doi: 10.1111/j.1464-410X.2005.05526.x.

Eeles, R. A. *et al.* (2013) 'Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array', *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 45, p. 385. Available at: https://doi.org/10.1038/ng.2560.

Eid, A., Alshareef, S. and Mahfouz, M. M. (2018) 'CRISPR base editors: genome editing without double-stranded breaks', *Biochemical Journal*, 475(11), pp. 1955 LP – 1964. doi: 10.1042/BCJ20170793.

Elliott, B. *et al.* (1998) 'Gene conversion tracts from double-strand break repair in mammalian cells', *Molecular and cellular biology*. American Society for Microbiology, 18(1), pp. 93–101. doi: 10.1128/mcb.18.1.93.

Ernst, J. *et al.* (2011) 'Mapping and analysis of chromatin state dynamics in nine human cell types', *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 473, p. 43. Available at: https://doi.org/10.1038/nature09906.

Ewels, P. *et al.* (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Fakhry, C. T. *et al.* (2016) 'Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks', *BMC Bioinformatics*, 17(1), p. 318. doi: 10.1186/s12859-016-1181-8.

Farashi, S. *et al.* (2019) 'Post-GWAS in prostate cancer: from genetic association to biological contribution', *Nature Reviews Cancer*, 19(1), pp. 46–59. doi: 10.1038/s41568-018-0087-3.

Ferlay, J. *et al.* (2018) 'Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018', *European Journal of Cancer*, 103, pp. 356–387. doi: https://doi.org/10.1016/j.ejca.2018.07.005.

Findlay, S. D. *et al.* (2016) 'A Digital PCR-Based Method for Efficient and Highly Specific Screening of Genome Edited Cells', *PLOS ONE*. Public Library of Science, 11(4), p. e0153901. Available at: https://doi.org/10.1371/journal.pone.0153901.

Franceschini, A. *et al.* (2012) 'STRING v9.1: protein-protein interaction networks, with increased coverage and integration', *Nucleic Acids Research*, 41(D1), pp. D808–D815. doi: 10.1093/nar/gks1094.

Fraser, M. *et al.* (2017) 'Genomic hallmarks of localized, non-indolent prostate cancer', *Nature*. Macmillan Publishers Limited, part of Springer Nature. All rights reserved., 541, p. 359. Available at: https://doi.org/10.1038/nature20788.

Gao, P. *et al.* (2018) 'Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus', *Cell*, 174(3), pp. 576-589.e18. doi: https://doi.org/10.1016/j.cell.2018.06.003.

Gaudelli, N. M. *et al.* (2017) 'Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage', *Nature*. Macmillan Publishers Limited, part of Springer Nature. All rights reserved., 551, p. 464.

Geng, C. *et al.* (2013) 'Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover', *Proceedings of the National Academy of Sciences*, 110(17), pp. 6997 LP – 7002. doi: 10.1073/pnas.1304502110.

Gerald, A. *et al.* (2004) 'Dihydrotestosterone and the Prostate: The Scientific Rationale for 5α-Reductase Inhibitors in the treatment of Benign Prostatic Hyperplasia', *Journal of Urology*. AUA Elsevier (Legacy Content), 172(4 Part 1), pp. 1399–1403. doi: 10.1097/01.ju.0000139539.94828.29.

Gnanapragasam, V. J. *et al.* (2001) 'Expression of RAC 3, a steroid hormone receptor co-activator in prostate cancer', *British journal of cancer*. Nature Publishing Group, 85(12), pp. 1928–1936. doi: 10.1054/bjoc.2001.2179.

Goldgar, D. E. *et al.* (1994) 'Systematic Population-Based Assessment of Cancer Risk in First-Degree Relatives of Cancer Probands', *JNCI: Journal of the National Cancer Institute*, 86(21), pp. 1600–1608. doi: 10.1093/jnci/86.21.1600.

Grasso, C. S. *et al.* (2012) 'The mutational landscape of lethal castration-resistant prostate cancer', *Nature*, 487(7406), pp. 239–243. doi: 10.1038/nature11125.

Grünewald, J. *et al.* (2019) 'Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors', *Nature*, 569(7756), pp. 433–437. doi: 10.1038/s41586-019-1161-z.

Guo, S. *et al.* (2001) 'Insulin Suppresses Transactivation by CAAT/Enhancer-binding Proteins β (C/EBPβ): SIGNALING TO p300/CREB-BINDING PROTEIN BY PROTEIN KINASE B DISRUPTS INTERACTION WITH THE MAJOR ACTIVATION DOMAIN OF C/EBPβ ', *Journal of Biological Chemistry* , 276(11), pp. 8516–8523. Available at: http://www.jbc.org/content/276/11/8516.abstract.

Heintzman, N. D. *et al.* (2007) 'Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome', *Nature Genetics*. Nature Publishing Group, 39(3), pp. 311–318. doi: 10.1038/ng1966.

Henttu, P., Liao, S. S. and Vihko, P. (1992) 'Androgens up-regulate the human prostate-specific antigen messenger ribonucleic acid (mRNA), but down-regulate the prostatic acid phosphatase mRNA in the LNCaP cell line.', *Endocrinology*, 130(2), pp. 766–772. doi: 10.1210/endo.130.2.1370795.

Hjelmborg, J. B. *et al.* (2014) 'The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer', *Cancer Epidemiology Biomarkers & Prevention*, 23(11), pp. 2303 LP – 2310. doi: 10.1158/1055-9965.EPI-13-0568.

Hornstein, B. D. *et al.* (2016) 'Effects of Circular DNA Length on Transfection Efficiency by Electroporation into HeLa Cells', *PLOS ONE*. Public Library of Science, 11(12), p. e0167537. Available at: https://doi.org/10.1371/journal.pone.0167537.

Hsieh, C.-L. *et al.* (2014) 'Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation', *Proceedings of the National Academy of Sciences*, 111(20), pp. 7319 LP – 7324. doi: 10.1073/pnas.1324151111.

Hua, J. T. *et al.* (2018) 'Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19', *Cell*, 174(3), pp. 564-575.e18. doi: https://doi.org/10.1016/j.cell.2018.06.014.

Huang, K. *et al.* (2018) 'Pathogenic Germline Variants in 10,389 Adult Cancers', *Cell*, 173(2), pp. 355-370.e14. doi: https://doi.org/10.1016/j.cell.2018.03.039.

Hughes, J. R. *et al.* (2014) 'Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment', *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 46, p. 205. Available at: https://doi.org/10.1038/ng.2871.

Jansen, R. *et al.* (2017) 'Conditional eQTL analysis reveals allelic heterogeneity of gene expression', *Human Molecular Genetics*, 26(8), pp. 1444–1451. doi: 10.1093/hmg/ddx043.

Jordan, M. and Wurm, F. (2004) 'Transfection of adherent and suspended cells by calcium phosphate', *Methods*, 33(2), pp. 136–143. doi: https://doi.org/10.1016/j.ymeth.2003.11.011.

Kanchi, K. L. *et al.* (2014) 'Integrated analysis of germline and somatic variants in ovarian cancer', *Nature Communications*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 5, p. 3156. Available at: https://doi.org/10.1038/ncomms4156.

Ki, S. H. *et al.* (2005) 'Glucocorticoid Receptor (GR)-Associated SMRT Binding to C/EBPβ TAD and Nrf2 Neh4/5: Role of SMRT Recruited to GR in GSTA2 Gene Repression', *Molecular and Cellular Biology*, 25(10), pp. 4150 LP – 4165. doi: 10.1128/MCB.25.10.4150-4165.2005.

Kim, M. H. and Field, J. (2008) 'Translationally regulated C/EBPβ isoform expression upregulates metastatic genes in hormone-independent prostate cancer cells', *The Prostate*. John Wiley & Sons, Ltd, 68(12), pp. 1362–1371. doi: 10.1002/pros.20801.

Kim, M. H., Minton, A. Z. and Agrawal, V. (2009) 'C/EBPβ regulates metastatic gene expression and confers TNF-α resistance to prostate cancer cells', *The Prostate*. John Wiley & Sons, Ltd, 69(13), pp. 1435–1447. doi: 10.1002/pros.20993.

Komor, A. C. *et al.* (2016) 'Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage', *Nature*, 533, p. 420. Available at: https://doi.org/10.1038/nature17946.

Kovács, K. A. *et al.* (2003) 'CCAAT/Enhancer-binding Protein Family Members Recruit the Coactivator CREB-binding Protein and Trigger Its Phosphorylation', *Journal of Biological Chemistry* , 278(38), pp. 36959–36965. Available at: http://www.jbc.org/content/278/38/36959.abstract.

de Laat, W. and Dekker, J. (2012) '3C-based technologies to study the shape of the genome', *Methods*, 58(3), pp. 189–191. doi: https://doi.org/10.1016/j.ymeth.2012.11.005.

Lappalainen, T. *et al.* (2013) 'Transcriptome and genome sequencing uncovers functional variation in humans', *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 501, p. 506. Available at: https://doi.org/10.1038/nature12531.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, K. *et al.* (2014) 'Optimization of Genome Engineering Approaches with the CRISPR/Cas9 System', *PLOS ONE*. Public Library of Science, 9(8), p. e105779. Available at: https://doi.org/10.1371/journal.pone.0105779.

Li, Q. *et al.* (2013) 'Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci', *Cell*, 152(3), pp. 633–641. doi: https://doi.org/10.1016/j.cell.2012.12.034.

Lichtenstein, P. *et al.* (2000) 'Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland', *New England Journal of Medicine*. Massachusetts Medical Society, 343(2), pp. 78–85. doi: 10.1056/NEJM200007133430201.

Lin, C.-Y. *et al.* (2018) 'Elevation of androgen receptor promotes prostate cancer metastasis by induction of epithelial-mesenchymal transition and reduction of KAT5', *Cancer Science*. John Wiley & Sons, Ltd (10.1111), 109(11), pp. 3564–3574. doi: 10.1111/cas.13776.

Lin, S. *et al.* (2014) 'Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery', *eLife*. Edited by D. Weigel. eLife Sciences Publications, Ltd, 3, p. e04766. doi: 10.7554/eLife.04766.

Liu, D. *et al.* (2018) 'C/EBPβ enhances platinum resistance of ovarian cancer cells by reprogramming H3K79 methylation', *Nature Communications*. doi: 10.1038/s41467-018-03590-5.

Lorenzin, F. and Demichelis, F. (2019) 'Evolution of the prostate cancer genome towards resistance', pp. 1–12.

Lu, C. *et al.* (2015) 'Patterns and functional implications of rare germline variants across 12 cancer types', *Nature Communications*. The Author(s), 6, p. 10086. Available at: https://doi.org/10.1038/ncomms10086.

Luedde, T. *et al.* (2004) 'C/EBP β isoforms LIP and LAP modulate progression of the cell cycle in the regenerating mouse liver', *Hepatology*. John Wiley & Sons, Ltd, 40(2), pp. 356–365. doi: 10.1002/hep.20333.

Mandelker, D. *et al.* (2017) 'Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline TestingMutation Detection in Patients With Advanced CancerMutation Detection in Patients With Advanced Cancer', *JAMA*, 318(9), pp. 825–835. doi: 10.1001/jama.2017.11137.

McCarthy, D. J., Chen, Y. and Smyth, G. K. (2012) 'Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation', *Nucleic Acids Research*, 40(10), pp. 4288–4297. doi: 10.1093/nar/gks042.

McNeal, J. E. (1988) 'Normal Histology of the Prostate', *The American Journal of Surgical Pathology*, 12(8). Available at: https://journals.lww.com/ajsp/Fulltext/1988/08000/Normal_Histology_of_the_Prostate.3.aspx.

Miah, S. and Catto, J. (2019) ‘BPH and prostate cancer risk’, (2), pp. 32–34.

Mink, S. *et al.* (1996) ‘A novel function for Myc: inhibition of C/EBP-dependent gene activation’, *Proceedings of the National Academy of Sciences of the United States of America*, 93(13), pp. 6635–6640. doi: 10.1073/pnas.93.13.6635.

Miyaoka, Y. *et al.* (2014) ‘Isolation of single-base genome-edited human iPS cells without antibiotic selection’, 11(3), pp. 291–293. doi: 10.1038/nmeth.2840.Isolation.

Nelson, P. S. *et al.* (2002) ‘The program of androgen-responsive genes in neoplastic prostate epithelium’, *Proceedings of the National Academy of Sciences*, 99(18), pp. 11890 LP – 11895. doi: 10.1073/pnas.182376299.

Osada, S. *et al.* (1996) ‘DNA Binding Specificity of the CCAAT/Enhancer-binding Protein Transcription Factor Family’, *Journal of Biological Chemistry* , 271(7), pp. 3891–3896. Available at: http://www.jbc.org/content/271/7/3891.abstract.

Pernar, C. H. *et al.* (2018) ‘The Epidemiology of Prostate Cancer’, *Cold Spring Harbor Perspectives in Medicine* , 8(12). Available at: http://perspectivesinmedicine.cshlp.org/content/8/12/a030361.abstract.

Pinello, L. *et al.* (2016) ‘Analyzing CRISPR genome-editing experiments with CRISPResso’, *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 34, p. 695. Available at: https://doi.org/10.1038/nbt.3583.

Pritchard, C. C. *et al.* (2016) ‘Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer’, *New England Journal of Medicine*. Massachusetts Medical Society, 375(5), pp. 443–453. doi: 10.1056/NEJMoa1603144.

Rachakonda, P. S. *et al.* (2013) ‘TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism’, *Proceedings of the National Academy of Sciences*, 110(43), pp. 17426 LP – 17431. doi: 10.1073/pnas.1310522110.

RAMJI, D. P. and FOKA, P. (2002) ‘CCAAT/enhancer-binding proteins: structure, function and regulation’, *Biochemical Journal*, 365(3), pp. 561 LP – 575. doi: 10.1042/bj20020508.

Ran, F. A. *et al.* (2013) ‘Genome engineering using the CRISPR-Cas9 system’, *Nature Protocols*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 8, p. 2281.

Rees, H. A. and Liu, D. R. (2018) ‘Base editing: precision chemistry on the genome and transcriptome of living cells’, *Nature Reviews Genetics*, 19(12), pp. 770–788. doi: 10.1038/s41576-018-0059-1.

Renaud, J.-B. *et al.* (2016) ‘Improved Genome Editing Efficiency and Flexibility Using Modified Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases’, *Cell Reports*, 14(9), pp. 2263–2272. doi: https://doi.org/10.1016/j.celrep.2016.02.018.

Richardson, C. D. *et al.* (2016) ‘Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA’, *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 34, p. 339. Available at: https://doi.org/10.1038/nbt.3481.

Rickman, D. S. *et al.* (2012) ‘Oncogene-mediated alterations in chromatin conformation’, *Proceedings of the National Academy of Sciences of the United States of America*. 2012/05/21. National Academy of Sciences, 109(23), pp. 9083–9088. doi: 10.1073/pnas.1112570109.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) ‘edgeR: a Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Romanel, A. *et al.* (2017) ‘Inherited determinants of early recurrent somatic mutations in prostate cancer’, *Nature Communications*, 8(1). doi: 10.1038/s41467-017-00046-0.

Rubin, M. A. and Demichelis, F. (2018) ‘The Genomics of Prostate Cancer : A Historic Perspective’. doi: 10.1101/cshperspect.a034942.

Sadar, K. L. M. and M. D. (2003) ‘ANDROGENS AND ANDROGEN RECEPTOR IN PROSTATE AND OVARIAN MALIGNANCIES’, *Frontiers in Bioscience*, 8, pp. 780–800.

Sahu, B. *et al.* (2011) ‘Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer’, *The EMBO Journal*, 30(19), pp. 3962 LP – 3976. doi: 10.1038/emboj.2011.328.

Salami, S. S. *et al.* (2018) 'Transcriptomic heterogeneity in multifocal prostate cancer', *JCI Insight*. The American Society for Clinical Investigation, 3(21). doi: 10.1172/jci.insight.123468.

Schröder, F. H. *et al.* (2012) 'Prostate-Cancer Mortality at 11 Years of Follow-up', *New England Journal of Medicine*. Massachusetts Medical Society, 366(11), pp. 981–990. doi: 10.1056/NEJMoa1113135.

Schug, J. (2008) 'Using TESS to Predict Transcription Factor Binding Sites in DNA Sequence', *Current Protocols in Bioinformatics*. John Wiley & Sons, Ltd, 21(1), pp. 2.6.1-2.6.15. doi: 10.1002/0471250953.bi0206s21.

Shaffer, P. L. *et al.* (2004) 'Structural basis of androgen receptor binding to selective androgen response elements', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0401123101.

Shand, R. and Gelmann, E. (2019) 'Molecular biology of prostate-cancer pathogenesis', pp. 2–3. doi: 10.1097/01.mou.0000193384.39351.64.

Shang, Y., Myers, M. and Brown, M. (2002) 'Formation of the androgen receptor transcription complex', *Molecular Cell*. doi: 10.1016/S1097-2765(02)00471-9.

Sharon, E. *et al.* (2018) 'Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing', *Cell*, 175(2), pp. 544-557.e16. doi: https://doi.org/10.1016/j.cell.2018.08.057.

Shimatani, Z. *et al.* (2017) 'Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion', *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 35, p. 441.

Shukla, A. and Huangfu, D. (2018) 'Decoding the noncoding genome via large-scale CRISPR screens', *Current Opinion in Genetics & Development*, 52, pp. 70–76. doi: https://doi.org/10.1016/j.gde.2018.06.001.

Siegel, R. L., Miller, K. D. and Jemal, A. (2017) 'Cancer statistics, 2017', *CA: A Cancer Journal for Clinicians*. American Cancer Society, 67(1), pp. 7–30. doi: 10.3322/caac.21387.

Slaymaker, I. M. *et al.* (2016) 'Rationally engineered Cas9 nucleases with improved specificity', *Science*, 351(6268), pp. 84 LP – 88. doi: 10.1126/science.aad5227.

Sotelo, J. *et al.* (2010) 'Long-range enhancers on 8q24 regulate c-Myc', *Proceedings of the National Academy of Sciences*, 107(7), pp. 3001 LP – 3005. doi: 10.1073/pnas.0906067107.

Steinberg, X. P. *et al.* (2012) 'Human CCAAT/Enhancer-Binding Protein β Interacts with Chromatin Remodeling Complexes of the Imitation Switch Subfamily', *Biochemistry*. American Chemical Society, 51(5), pp. 952–962. doi: 10.1021/bi201593q.

Stelloo, S. *et al.* (2018) 'Integrative epigenetic taxonomy of primary prostate cancer', *Nature communications*. doi: 10.1038/s41467-018-07270-2.

Sternberg, S. H. and Doudna, J. A. (2015) 'Expanding the Biologist's Toolkit with CRISPR-Cas9', *Molecular Cell*, 58(4), pp. 568–574. doi: https://doi.org/10.1016/j.molcel.2015.02.032.

Takayama, K. and Inoue, S. (2013) 'Transcriptional network of androgen receptor in prostate cancer progression', *International Journal of Urology*. John Wiley & Sons, Ltd (10.1111), 20(8), pp. 756–768. doi: 10.1111/iju.12146.

Tan, J. *et al.* (2019) 'Engineering of high-precision base editors for site-specific single nucleotide replacement', *Nature Communications*, 10(1), p. 439. doi: 10.1038/s41467-018-08034-8.

Tan, M. E. *et al.* (2015) 'Androgen receptor: Structure, role in prostate cancer and drug discovery', *Acta Pharmacologica Sinica*. doi: 10.1038/aps.2014.18.

Tong, Y. *et al.* (2018) 'Cumulative Evidence for Relationships Between 8q24 Variants and Prostate Cancer ', *Frontiers in Physiology* , p. 915.

Wang, Q. *et al.* (2007) 'A Hierarchical Network of Transcription Factors Governs Androgen Receptor-Dependent Prostate Cancer Growth', *Molecular Cell*, 27(3), pp. 380–392. doi: https://doi.org/10.1016/j.molcel.2007.05.041.

Wang, Q. *et al.* (2009) 'Androgen Receptor Regulates a Distinct Transcription Program in Androgen-Independent Prostate Cancer', *Cell*, 138(2), pp. 245–256. doi: https://doi.org/10.1016/j.cell.2009.04.056.

Wang, Q., Carroll, J. S. and Brown, M. (2005) 'Spatial and Temporal Recruitment of Androgen Receptor and Its

Coactivators Involves Chromosomal Looping and Polymerase Tracking', *Molecular Cell*, 19(5), pp. 631–642. doi: https://doi.org/10.1016/j.molcel.2005.07.018.

Wang, W., Bergh, A. and Damber, J.-E. (2007) 'Increased expression of CCAAT/enhancer-binding protein beta in proliferative inflammatory atrophy of the prostate: Relation with the expression of COX-2, the androgen receptor, and presence of focal chronic inflammation', *The Prostate*. John Wiley & Sons, Ltd, 67(11), pp. 1238–1246. doi: 10.1002/pros.20595.

Watson, P. A., Arora, V. K. and Sawyers, C. L. (2015) 'Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer', *Nature Reviews Cancer*. doi: 10.1038/nrc4016.

Wedge, D. C. *et al.* (2018) 'Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets', *Nature Genetics*, 50(5), pp. 682–692. doi: 10.1038/s41588-018-0086-z.

Weinhold, N. *et al.* (2014) 'Genome-wide analysis of noncoding regulatory mutations in cancer', *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 46, p. 1160. Available at: https://doi.org/10.1038/ng.3101.

White, C. W., Xie, J. H. and Ventura, S. (2013) 'Age-related changes in the innervation of the prostate gland', *Organogenesis*. Taylor & Francis, 9(3), pp. 206–215. doi: 10.4161/org.24843.

Yan, M. *et al.* (2017) 'Identification of SPOP related metabolic pathways in prostate cancer', *Oncotarget*. Impact Journals LLC, 8(61), pp. 103032–103046. doi: 10.18632/oncotarget.21460.

Young, C. Y.-F. *et al.* (1991) 'Hormonal Regulation of Prostate-specific Antigen Messenger RNA in Human Prostatic Adenocarcinoma Cell Line LNCaP', *Cancer Research*, 51(14), pp. 3748 LP – 3752. Available at: http://cancerres.aacrjournals.org/content/51/14/3748.abstract.

Yu, G. *et al.* (2012) 'clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters', *OMICS: A Journal of Integrative Biology*. Mary Ann Liebert, Inc., publishers, 16(5), pp. 284–287. doi: 10.1089/omi.2011.0118.

Zafra, M. P. *et al.* (2018) 'Optimized base editors enable efficient editing in cells, organoids and mice', *Nature biotechnology*. 2018/07/03, 36(9), pp. 888–893. doi: 10.1038/nbt.4194.

Zentner, G. E., Tesar, P. J. and Scacheri, P. C. (2011) 'Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions', *Genome research*. Cold Spring Harbor Laboratory Press, 21(8), pp. 1273–1283. doi: 10.1101/gr.122382.111.

Zhang, Q. *et al.* (2006) 'A Hedgehog-Induced BTB Protein Modulates Hedgehog Signaling by Degrading Ci/Gli Transcription Factor', *Developmental Cell*, 10(6), pp. 719–729. doi: https://doi.org/10.1016/j.devcel.2006.05.004.

Zhou, C. *et al.* (2019) 'Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis', *Nature*. doi: 10.1038/s41586-019-1314-0.

Zidek, L. M. *et al.* (2015) 'Deficiency in mTORC1-controlled C/EBPβ-mRNA translation improves metabolic health in mice', *EMBO reports*, 16(8), pp. 1022 LP – 1036. doi: 10.15252/embr.201439837.

9 Acronyms

Table of Acronyms

3C	Chromosome Conformation Capture
ABE	Adenine Base Editors
ACTH	Adrenocorticotropic Hormone
AD	Androgen Deprivation
AD	Androstenedione
APOBEC	Apolipoprotein B mRNA Editing Catalytic Polypeptide-like
BHQ	Black Hole Quencher
BPH	Benign Prostatic Hyperplasia
CBE	Cytosine Base Editor
ChIP-qPCR	Chromatin Immunoprecipitation coupled with quantitative PCR
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/ CRISPR associated (Cas) nucleases
CRPC	Castrate Resistant Prostate Cancer
CYP17A1	Cytochrome P450 family 17 subfamily A polypeptide 1
DBD	DNA Binding Domain
ddPCR	Droplet Digital PCR
DEG	Differentially Expressed Genes
DHEA	Dehydroepiandrosterone
DHEA-S	Dehydroepiandrosterone-Sulfate
DHT	Dihydrotestosterone
DSB	Double Strand Brake
ENCODE	Encyclopedia of DNA Elements
eQTLs	Quantitative Trait Loci
FSH	Follicle Stimulating Hormone
GnRH	Gonadotropin-Releasing Hormone
GWAS	Genome-Wide Association Studies
HADAC	Histone Deacetylation
HAT	Histone Acetyltransferase
HDR	Homologous Dependent Recombination
HR	Homologous Recombination
HSD17β	17-β Hydroxysteroid Dehydrogenase
ICGC	International Cancer Genome Consortium
LD	Linkage Disequilibrium
LH	Luteinizing Hormone
NHEJ	Non-homologous end Joining
PAM	Protospacer Adjacent Motifs
PCAWG	Pan Cancer Analysis of Whole Genomes
RNP	Ribonucleoprotein
SCNA	Somatic Copy Number Alteration
SHBG	Sex Hormone Binding Globulin
SNPs	Single Nucleotide Polymorphisms
SPOP	Speckle-Type POZ protein
SRC	Steroid Receptor Coactivator
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
UGI	Uracil Glycosylase Inhibitor
WGS	Whole-Genome Sequencing