**VILNIUS UNIVERSITY**

Life Sciences Center

Institute of Biotechnology

European Molecular Biology Laboratory

**LINA KRIKŠČIKAITĖ**

# Profiling the Specificity of the TnpB - Transposon Associated Nuclease

Master Thesis

Genetics Study Programme

Work was done at VU,

VU LSC-EMBL Partnership Institute, Jones Lab

Supervisor:

Prof. Dr. Stephen Knox Jones

Vilnius, 2023

**VILNIAUS UNIVERSITETAS**

Gyvybės mokslų centras

Biotechnologijų institutas

EMBL partnerystės institutas

**LINA KRIKŠČIKAITĖ**

# Su transpozonais susijusios nukleazės TnpB specifiškumo nustatymas

Genetikos magistro studijų baigiamasis darbas

Darbas atliktas VU,

Biotechnologijų institute, EMBL, Jones Lab

Darbo vadovas:

Prof. Dr. Stephen Knox Jones

Vilnius, 2023

Contents

**Abbreviations**

AAV – adeno associated virus

CAS - Chemical Abstracts Service (identifies a chemical substance or molecular structure)

Cas – CRISPR associated

CRISPR – clustered regularly interspaced short palindromic repeats

DBD – DNA-binding domain

EMBL – European Molecular Biology Laboratory

DSB – double-stranded breaks

HbF - foetal hemoglobin

HSPCs - hematopoietic stem and progenitor cells

gRNA – guide RNA

HBG - HbF repressor γ-globin

HDR – homology-directed repair

IVTT – *in vitro* transcription translation

GMO – genetically modified organism

LE – left end palindromic element

LSC – Life Sciences Center

NGS – next-generation sequencing

NHEJ – nonhomologous end joining

nt – nucleotide

rDNA – recombinant DNA

RE – right end palindromic element

RNP – ribonucleoprotein

RPA - recombinase polymerase amplification

RT-RPA - reverse transcriptase RPA

RT – reverse transcriptase

SCD – sickle cell disease

VU – Vilnius University

TALEN - transcription activator-like effector nuclease

UN number - United Nations Number (a four-digit number that identifies hazardous materials)

ZFN – zinc finger nuclease

ZnF – zinc finger

## Introduction

Gene editing is a powerful scientific approach for making precise changes to an organism's DNA. It has gained significant importance due to its potential to revolutionize various fields and address numerous challenges. Genetic engineering addresses such problems as treating inherited diseases, eliminating pests and increasing harvest in agriculture, providing more food supplies to maintain the growing human population, and even executing rapid molecular diagnostics (Ganbaatar & Liu, 2021). Nevertheless, gene editing comes with several drawbacks of which successful delivery of a gene editor to a target cell is one of the most common burdens. Nowadays, clustered regularly interspaced short palindromic repeats (CRISPR) associated proteins (Cas) are proven to outperform other editors by being faster, cheaper, more accurate, and more efficient (Xu & Li, 2020). One of the main benefits of CRISPR Cas is that they are easily reprogrammable due to their dependency on a single targeting molecule – guide RNA (gRNA) for target sequence recognition. The main drawback of Cas proteins is their size. Packaging SpCas9 (*Streptococcus pyogenes* Cas9) and a gRNA together (~4.2 kb) into an adeno-associated viral vector (AAV) is challenging due to AAV's packaging capacity (~4.7 kb). Due to this problem, scientists are looking for new nucleases that would be smaller and therefore easier to deliver. This study focuses on TnpB – a transposon-associated nuclease that is two times smaller than Cas9. TnpB is believed to be a progenitor for Cas nucleases and exhibits similar biochemical patterns (Sasnauskas et al., 2023). It has been already claimed that TnpB is capable of editing genomes by introducing indel (insertion and deletion) mutations to human genes (Karvelis et al., 2021).

Since TnpB is a recently discovered protein, it is not exactly known how it interacts with DNA targets. This motivates **the main goal of this project:** determine TnpB's target specificity profile with next-generation biochemistry. One of the main methods used in this project – NucleaSeq – was developed by Stephen Knox Jones (Jones et al., 2021). NucleaSeq is a platform that combines DNA digestion by a nuclease with deep sequencing. NucleaSeq reveals the cleavage kinetics of a nuclease and benchmarks its fidelity by quantifying the rates of cleavage across DNA targets, including thousands of off-targets that do not fully match the nuclease's RNA guide. It also reveals where does the nuclease cleave within a target sequence and what DNA ends (sticky or blunt) does it generate. By profiling TnpB with NucleaSeq, the specificity (the ratio of on-target cleavage to that of off-targets - across the complete library) and cleavage rates together with information about the DNA ends after the cleavage can be revealed. Data obtained with NucleaSeq will also help to understand how does TnpB and its

features compare to those of well-studied nucleases like Cas9 and Cas12a. The main goal is approached with the **following tasks:**

1. Determine the activity of purified TnpB RNP (ribonucleoprotein complex)
2. Prepare a target library containing sequences that perfectly match TnpB‘s gRNA, and thousands of off-target sequences that mispair (insertions, deletions, mismatches) with it.
3. Determine reaction conditions for performing NucleaSeq with TnpB RNP *in vitro*.
4. With NucleaSeq, determine the following TnpB properties:
    a) Cleavage rates for TnpB RNP across DNA target library.
    b) Benchmark the specificity of TnpB RNP class-leading gene editing tools like CRISPR-Cas9.
    c) Identify the cleavage sites and DNA ends (sticky or blunt) that TnpB RNP generates when cleaving DNA.

# 1. Literature review

## 1.1 Introduction to gene editing: definition and a brief history

Gene editing is a process that involves manipulating and altering the DNA of an organism to give edited DNA regions (genes) a modified function. Manipulations include removing existing DNA (deletion), introducing new DNA (insertion), and exchanging an existing nucleotide to a different nucleotide (substitution). These processes are achievable with molecular technologies (s*ubsection 1.2*) that rely on DNA nucleases. Typically, the mechanism of gene editing is based on the recognition of the DNA target through the nuclease's DNA-binding domain (DBD) formation of a nuclease-induced double-stranded break (DSB), and the introduction of a wanted edit or mutation during repair of the DSB by homology-directed recombination (HDR) or non-homologous end joining (NHEJ) (M. Liu et al., 2019).

Several events outlined a pathway for the beginning of the gene editing era. Determining DNA structure was pioneered in 1953 by James Watson and Francis Crick and just before their discovery Rosalinda's Franklin X-ray diffraction imaging of DNA gave a hint of a 'double helix' (Portin, 2014). Just in the same year, 1953, Arthur Kornberg isolated DNA polymerase from the bacteria and synthesized DNA *in vitro* (McHenry & Kornberg, 1977). The discovery of DNA ligases – essential enzymes that join DNA strands together through phosphodiester bonds, took place in 1967 and was shared between Gellert, Lehman, Richardson, and Hurwitz laboratories (Gu et al., 2012). Simultaneously, the idea of the existence of a restriction-modification system started as a hypothesis of bacterial immunity against viruses by Werner Aber (Di Felice et al., 2019).

Finally, around the 1970s, the first attempts to edit DNA started to take off little by little. Hamilton Smith, a molecular biologist at Johns Hopkins University School of Medicine, purified the first restriction enzyme in history – type II site-specific HindII from *Haemophilus influenzae* (Loenen et al., 2014). Another scientist, Daniel Nathans tested this enzyme on SV40 viral genome and found that it digests the viral genomic DNA into 11 fragments (Danna & Nathans, 1971). Nathan also demonstrated that digestion products of viral DNA can be separated and ranged by their size using gel electrophoresis and that outlined future pathways for mapping out the genome of many different organisms. Those discoveries lead to the first creation of the recombinant DNA (rDNA). By 1973, scientists Herbert Boyer and Stanley Cohen accomplished transgenesis by inserting recombinant DNA into the bacterial genome and making the bacteria resistant to kanamycin (Cohen et al., 1973). In less than 10 years after the creation of rDNA, the first transgenic animal was created. Thomas Wagner's team at Ohio

University in 1981 transferred the gene of a rabbit into the mouse genome through DNA microinjection. In 1982, Genentech commercialized the first synthetic insulin that was prepared by David Goeddel and his colleagues (Goldner, 1972). The discovery of the PCR technique in 1983 by Kary Mullis simplified the process of amplifying DNA and making many of copies of it for further experiments (Mullis, 1990).

After establishing the most useful methods for gene editing and adopting restriction enzymes for manipulations with DNA, the scientific community started developing more tools and techniques to improve the editing processes. 1985 marked the discovery of zinc finger nuclease (ZFN) – the first gene editor that could be reprogrammed to target desired DNA sequences (Klug, 2010). 1987 marked the discovery of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated (Cas) systems (Ishino et al., 2018). The first use of CRISPR-Cas systems to edit bacterial genes was reported in 2012 by Dr. Jennifer Doudna and her team at the University of California, Berkeley, in collaboration with Dr. Emmanuelle (Jinek et al., 2012). The discovery of transcription activator-like effector nucleases (TALENs) took place in 2011 (Jinek et al., 2012). Over time, CRISPR techniques started to gain popularity over other gene editors due to being fast, cheap, accurate and easy to reprogram and brought more applications (*subchapters 1.2, 1.4)*. The development of a method for genome editing based on CRISPR–Cas9 technology was awarded The Nobel Prize in Chemistry in 2020 to Emmanuelle Charpentier and Jennifer Doudna. CRISPR is the current focal center of gene editing approaches, though use of ZFNs and TALENs continues (Kim & Lee, 2022).

## 1.2 Types of gene editing tools

There are many ways to approach gene editing, yet scientists must balance the specificity (how often a specific gene editing tool makes intended modifications vs unintended) of the editing tool, the time it takes to make the edit, and the cost of the entire process. At the moment, there are 7 well-established strategies for editing (*1.2 table*). In this subsection, different strategies are briefly reviewed and compared to each other by their specificity, programmability, and applicability.

**1.2 table. Different gene editing strategies compared by their development time, the source they came from, principle of genetic editing, application and programmability.**

| Gene editing tool | Year of development | Source and Principle | Application | Programmability |
|---|---|---|---|---|
| Restriction enzymes | 1970 | Found in bacteria and archaea. Works as antiviral immunity mechanism.<br><br>Cuts DNA into fragments at or near specific recognition sites within DNA sequence (restriction sites). | Molecular cloning, DNA mapping, epigenome mapping, site directed mutagenesis, constructing DNA libraries. | Not programmable |
| Homing nucleases | 1978 | Encoded in introns or inteins<br><br>By cleaving DNA at specific sites, homing endonucleases facilitate the movement of these genetic elements to new locations. | Creating DSB in target places for further molecular biology purposes. | Programmable (reengineering the protein) |
| Zinc finger nucleases | 1985 | Zinc finger-binding domains are adopted from transcription factors, while nuclease domain is adopted from FokI.<br><br>Amino acids in each zinc finger recognize and bind DNA while nuclease domain induce DSB. | Widely used in biotechnology, gene therapy and creating GMOs. | Programmable (reengineering the protein) |
| Transcription activator-like effector nucleases (TALENs) | 2011 | TALE-DNA binding domain is adopted from *Xanthomonas*, while nuclease domain is adopted from FokI.<br><br>$12^{th}$ and $13^{th}$ amino acids in TALE-DNA binding domain recognise specific DNA while nuclease domain induce DSB. | Widely used in biotechnology, gene therapy and creating GMOs. | Programmable (reengineering the protein) |
| CRISPRs | 2012 (first gene editing approach) | Found in bacteria and archaea. Works as antiviral immunity mechanism. | Widely used in biotechnology, gene therapy, molecular diagnostics and creating GMOs. | Easily programmable |

| | | | | |
|---|---|---|---|---|
| | | Cas nuclease recognize PAM that flanks the target and after gRNA:DNA target hybridization it induces DSB. | Allows to create DSB for further manipulations with DNA or carries whole manipulation within simultaneous procedure (base editing and prime editing). | |

### 1.2.1 Restriction endonucleases

The first molecular mechanisms used for gene editing purposes were restriction nucleases. Discovered in the 1970s and first successfully adopted by Stanley Cohen and Herbert Boyer to make recombinant DNA, these restriction enzymes made gene editing a reality. As reported, Cohen and Boyer used a type II restriction enzyme EcoRI to create recombinant plasmid containing a gene from different bacterial species (Cohen et al., 1973). Nowadays, we know that restriction enzymes are categorized into 4 types (I-IV) which differ in structure, target recognition, and cleavage pattern and cofactor requirements. However, all of these enzymes recognize specific short DNA sequences, participate in the endonucleolytic cleavage of DNA, and leave specific fragments with terminal 5'-phosphates. The most popular 'type II' restriction endonucleases are structured as homodimers with the capability of recognizing small (4-8 bp), usually palindromic (5' to 3' of the forward strand is identical to its complementary, 5' to 3' reverse strand) DNA sequences (Pingoud & Jeltsch, 2001). Type II subtypes include the popular Type IIP that cleave within recognition sites, and Type IIS that cleave outside recognition sites. Overall, type II restriction endonucleases are easy to adapt for every day molecular biology applications – they cleave DNA at fixed positions, create reproducible cleavage products, and mainly depend on Mg(2+) as cofactor.

In an evolutionary context, restriction enzymes are found in bacteria and archaea genomes, where they work as antiviral immunity mechanisms (Kruger & Bickle, 1983). They participate in a process termed "restriction digestion" when they cut foreign DNA. The prokaryotic genomic DNA is modified by methyltransferase and such methyl groups protect host DNA by blocking accessibility for the restriction enzyme. Therefore, restriction digestion and methylation create the 'restriction-modification system'. In gene engineering processes, restriction enzymes cleave DNA and open up a new space for modifying the DNA. However, they are limited by the nucleotide patterns that they recognize, limiting their utility when

compared to programmable gene editors. Restriction enzymes remain inevitable tools for molecular cloning and they play key roles in DNA mapping, epigenome mapping, and constructing DNA libraries (Di Felice et al., 2019).

### 1.2.2 Homing endonucleases

Independent of the discovery of restriction enzymes and the first gene editing approaches, discovery and characterization of homing nucleases (alternatively called meganucleases) also took part in 1970s. In 1978, an intervening sequence within yeast mitochondrial ribosomal DNA (rDNA) was visualized using electron microscopy by researchers from Institut Pasteur in Paris (Rédei, 2008). They found that the sequence encodes a nuclease, capable of creating DSB in the genome. The discovery thus defined homing endonucleases: they are a group of nucleases that are either a part of a gene intron or an intein. Intein is a segment of a protein that is able to excise itself and join the remaining portions (exteins) with a peptide bond during protein splicing (Perler, 1998). Homing nucleases are divided into five families based on sequence and structure motifs, but best explored is the LAGLIDADG family (Marcaida et al., 2010). This family contains more than 200 members that share one or two copies of the consensus LAGLIDADG sequence (Silva & Belfort, 2004). Generally, LAGLIDADG proteins consist of two similar α/β domains (αββαββα) assembled either as homodimers or monomers. Domains are responsible for binding DNA and they are related by pseudo-two-fold symmetry; the center of this symmetry is the closely packed LAGLIDADG two-helix bundle. LAGLIDADG proteins target long DNA sequences. These proteins display the ability to recognize and cleave long DNA targets (15–40 bp) in a homologous allele that lacks the intron or intein sequence, and therefore, they promote the lateral transfer of their encoding intron or intein to these sites by a targeted transposition mechanism termed “homing” (Jurica et al., 1998). However, neither LAGLIDADG nor other families of homing nucleases have achieved widespread adoption as tools for genome engineering. The main reason is their troubled programmability - the construction of sequence-specific enzymes is pricy and time consuming, and the number of naturally occurring homing nucleases is limited and does not cover all potentially interesting loci (Abdallah et al., 2015).

### 1.2.3 Zinc finger nucleases

Zinc finger nucleases (ZFNs) are artificial restriction enzymes generated by fusing a zinc finger DNA-binding domain to a DNA-cleavage domain. The discovery of ZFNs took place in 1980s and in 1990s ZFNs were demonstrated as site-specific nucleases for cutting DNA at

strictly defined sites (Klug, 2010). Naturally, zinc finger motifs occur in transcription factors found in the cell: zinc finger-binding domains bind DNA and associated transcription factor effector domains activate or repress the bound gene. Zinc finger domains were adopted to build the first highly specific gene editing tool by fusing the restriction enzyme FokI to a zinc finger-binding domain to create "genomic scissors" that cleave DNA at a specified location. Each zinc finger domain contains a zinc ion coordinated by specific amino acids. These amino acid residues form a structure that interacts with DNA bases in the target sequence, allowing for sequence-specific recognition. Usually, each zinc finger recognise a 3 bp sequence and whole ZFN can be engineered to bind a specific ~20bp DNA sequence in the genome by combining 6 to 8 zinc finger-binding domains (with characterized recognition sites) with a FokI restriction nuclease domain (M. Li & Wang, 2017). ZFNs can be programmed by combining zinc finger-binding domains with different nucleotide preferences (adopted from different transcription factors), but as gene editing tools, they have some disadvantages. Domain construction of ZFN is a costly, complex and time-consuming process (Roshanravan et al., 2022). Further, off-target editing (nonspecific or unintended genome modifications) might occur due to inappropriate interactions between the fused domains of ZFN (Abdallah et al., 2015).

### 1.2.4 Transcription Activator-Like Effector Nucleases

Like ZFNs, Transcription activator-like effector nucleases (TALENs) are programmable restriction enzymes made by fusing the transcription activator-like effector (TALE) DNA binding domain of *Xanthomonas* bacteria with a FokI DNA cleavage domain (Chandrasegaran & Carroll, 2016). Just like transcription factors of eukaryotic species, TALEs regulate expression of target genes in bacteria. In general, TALENs operate under the same principle as ZFNs. Combining the catalytic domain of FokI with suitable sequence-specific TALE domains creates artificial nucleases that could target any nucleotide sequence in a genome. A single TALE domain consists of 33-34 amino acid repeating motifs with two variable positions (12th and 13th) that have a strong recognition for specific DNA sequences and can be reprogrammed to cut DNA at the specific locations (Chandrasegaran & Carroll, 2016). Importantly, the binding specificity of TALE domains requires a thymine at the 5' end of the target DNA molecule, which affects binding efficiency (Joung & Sander, 2013).

Overall, TALENs were found to have less off-target effects than ZFNs, and are less cytotoxic to the host cell (Mussolino et al., 2014). Yet, TALENs exhibit a few obstacles: they are significantly larger than ZFNs, and therefore can be harder to deliver and express *in vivo*

(Gupta & Musunuru, 2014). They also need a thymine before the 5'- end of the target sequence, making TALEN programmability more limited.

### 1.2.5 CRISPR-Cas

As already mentioned, CRISPR systems were discovered by Japanese molecular biologist Yoshizumi Ishino from Osaka university by the year 1987. At that time, scientists were not aware of CRISPR origins or their importance to prokaryotic cells. By 1995, Francisco Mojica of the University of Alicante (Spain) came across similar structures in the archaeal genome of *Haloferax mediterranei* (Mojica & Rodriguez-Valera, 2016). According to Mojita, presence of those structures in two evolutionarily remote domains of life suggested their great functional significance and he was one of the first scientists to hypothesize that these unusual loci include fragments of foreign DNA and are, in fact, part of the immune system of bacteria and archaea (Ishino et al., 2018). Currently, CRISPR repeats have been found in most archaeal genomes and nearly half of the studied bacterial ones, but they have not been found in eukaryotic sequences.

Speaking of gene editing capacity, CRISPR-Cas systems typically outperform the gene editing tools listed above and are a popular choice for gene editing. Naturally found in bacteria and archaea, CRISPR-Cas systems defend against invading viruses (*fig. 1.2.5)*. One of their most important features is that they rely on a single targeting molecule, termed a guide RNA (gRNA) and a protospacer adjacent motif (PAM) for sequence-specific target recognition (Corsi et al., 2022). This gRNA's sequence is obtained from invading viruses and directs the activity of the Cas (CRISPR associated) proteins, which are encoded in the bacteria or archaea's genome. gRNA is recognized by specific domain of Cas protein, for example, REC1 domain of Cas9 is responsible for gRNA and Cas protein hybridization and after the ribonucleoprotein (RNP) is made by combining Cas9 with the gRNA, the RNP can target DNA sequences that complement its gRNA (Palermo et al., 2018). PAM is also important for RNP to bind DNA - PAM flanks the target sequence either at its 5' or 3' end and is recognized by Cas nuclease and helps to distinguish the viral genome from the host genome. Cas9 from *Corynebacterium diphtheriae* recognizes PAM by a specific set of amino-acid residues in the PAM-interacting (PI) domain (Hirano et al., 2019). For Cas9 nuclease the PAM sequence is typically found 2-6 nucleotides downstream of the DNA target and the Cas cuts 3-4 nucleotides upstream of it (Mojica & Rodriguez-Valera, 2016). Cas9 of *S. pyogenes* recognizes a 5′-NGG-3′ PAM ("N" stands for any nucleotide base). However, the spacers in its CRISPR array are coded by 5'-GTT-3', so the Cas9 cannot cut the bacteria's own genome. Anyway, the

importance of gRNA and PAM is to successfully set up the RNP to protect the prokaryotic cell from the invading viruses.

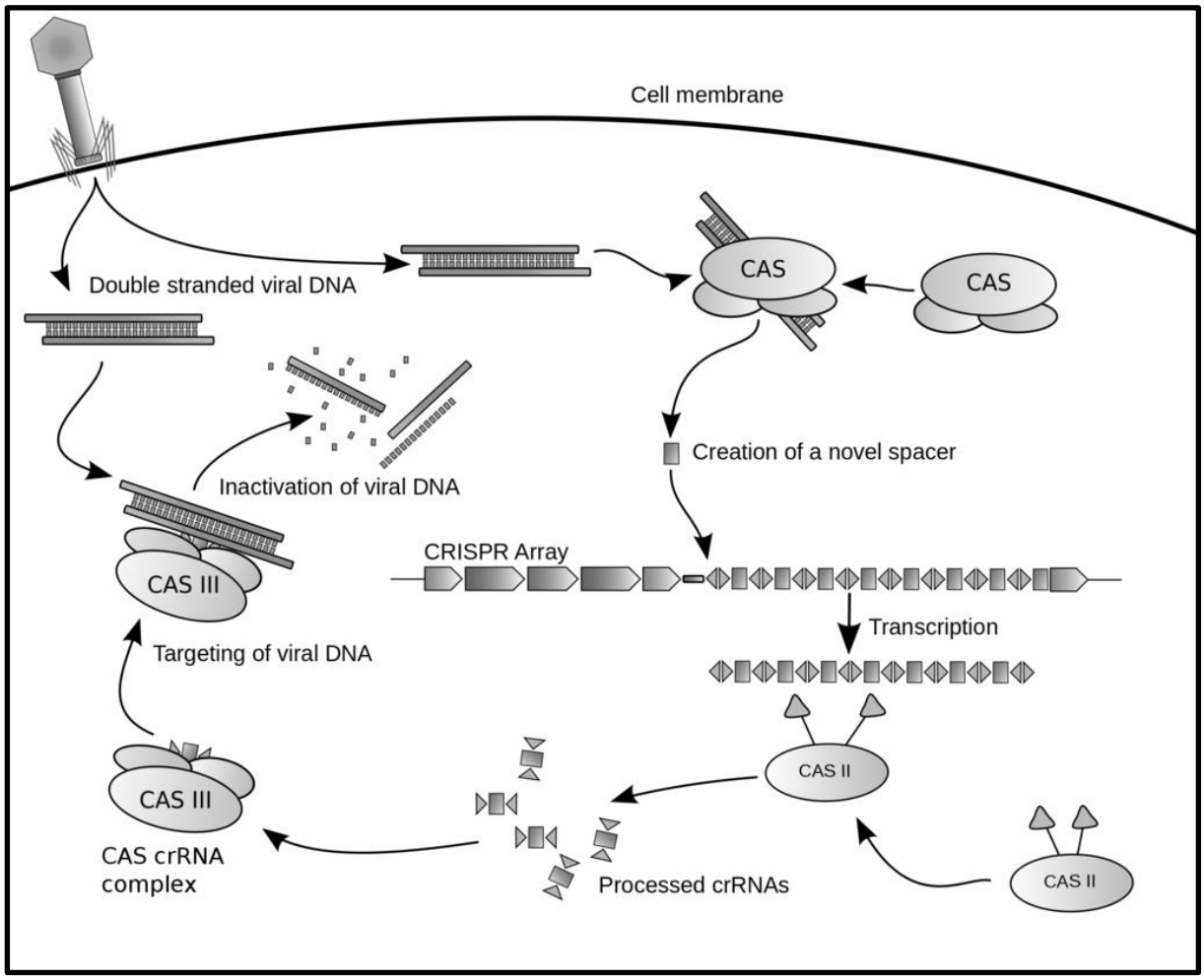



**Figure 1.2.5 CRISPR/Cas as prokariotic immunity mechanism against invading viruses (Horvath & Barrangou, 2010).**
CRISPR-Cas genetic structures are grouped into "spacers" (17-84 bp long) that represents a DNA obtained from invading viruses and short palindromic repeats (23-50 bp long) that separate the spacers. When a microbe is invaded by a bacteriophage, the first stage of the immune response is to capture phage DNA and insert it into a CRISPR locus in the form of a spacer. It is suggested that proteins Cas I and Cas II are required for insertion of new spacer. After transcription of a novel spacer, Cas II generates crRNAS which then are paired with other Cas proteins and used for degradation of crRNA complementary viral DNA.

It was already mentioned that Cas9 contains REC1 one domain. This domain is a part of REC lobe which is crucial not only for binding gRNA, but complementary DNA as well (Palermo et al., 2018). In general, REC lobe is divided to three regions (REC1-REC2-REC3) that all together mediates binding of nucleic acids. Cas9 also consist of nuclease (NUC) lobe that is responsible for cleavage of target DNA and consists of endonuclease domain (HNH, H – histidine, N - asparagine), RuvC and PI domains (Nishimasu et al., 2014). HNH is responsible for cleaving the DNA strand complementary to gRNA. Oppositely, RuvC initiates cleavage of

the DNA strand that isn't complementary to gRNA. This domain was named according protein found in *E. coli* that is involved in DNA repair (Iwasaki et al., 1991). Unlike Cas9, the Cas12a nuclease does not possess an HNH domain in its NUC lobe, but rather RuvC, PI, WED (wedge domain) and the bridge helix (BH) (Paul & Montoya, 2020). Just like REC lobe of Cas12a that consists of REC1 and REC2 domains, WED is also responsible for recognizing nucleic acids and it also contains of RNase site which is responsible for processing gRNA. The structural differences between Cas proteins highlight their different biochemical and catalytic features: Cas9 does not process its RNAs but requires both trans-activating RNA (tracrRNA) and CRISPR RNA (crRNA) for target recognition, and target cleavage generates blunt DNA ends (Dooley et al., 2021). Cas12a processes its own RNAs and only requires crRNA for target cleavage and generates staggered DNA ends (Paul & Montoya, 2020).

Overall, CRISPR-Cas based gene editing mechanisms are very popular nowadays and they show off some advantages over other kinds of gene editing tools mentioned in this subsection. First of all, they are easy to reprogram. *In vitro* RNA synthesis is a standard molecular biology method, and CRISPR-Cas dependency on gRNA makes these nucleases easily programmable (Barrangou, 2013). Another reason that CRISPR technologies are so popular is their variety: basic CRISPR-Cas mediated DSB can be followed by further manipulations, while CRISPR-based tools, such as base editors and prime editors, manipulate DNA without inducing DSBs. Base editing is a form of CRISPR technology that mediates a conversion of one DNA base into another by using a modified version of the Cas9 protein with removed nuclease activity and a fused domain that can perform chemical modifications on DNA bases while prime editors fuses a modified Cas9 protein that exhibits nickase activity with a reverse transcriptase (RT) domain, which allows to introduce new DNA sequences into the genome at a target location (Kantor et al., 2020).

While CRISPR technologies stand out with benefits, just like ZFNs and TALENs they are related to off-target effects and cytotoxicity. Another drawback of CRISPR-Cas tools is their large size: packaging SpCas9 (*Streptococcus pyogenes* Cas9) and a gRNA together (~4.2 kb) (Y. Liu et al., 2018) into common delivery vectors like an AAV (Adeno-associated virus) is challenging due to AAV packaging capacity limits (~4.7 kb) (Dong et al., 2010). Scientists are investigating smaller Cas nucleases, like type V-F CRISPR nucleases (Bigelyte et al., 2021) or systems highly similar to CRISPR-Cas like TnpB (Karvelis et al., 2021) that will be discussed in this thesis (*subsection 1.5*).

Each gene editing tool comes with advantages and disadvantages, and for different goals, one might outperform another. Generally, CRISPR-based tools lead in gene editing due to their

greater gene-editing efficiency and fewer off-target effects (Yang & Yang, 2022). Additionally, CRISPR-Cas systems are easily reprogrammable and capable of cutting DNA strands, so they do not need to be paired with separate cleaving enzymes. Also, when paired with other enzyme like deaminase or RT, CRISPRs can alter DNA without introducing a DSB. Although highly beneficial, CRISPR did not start gene editing era, and without investigation of restriction enzymes and other pioneering gene editors, the successful pathway to DNA modifications may have never been discovered.

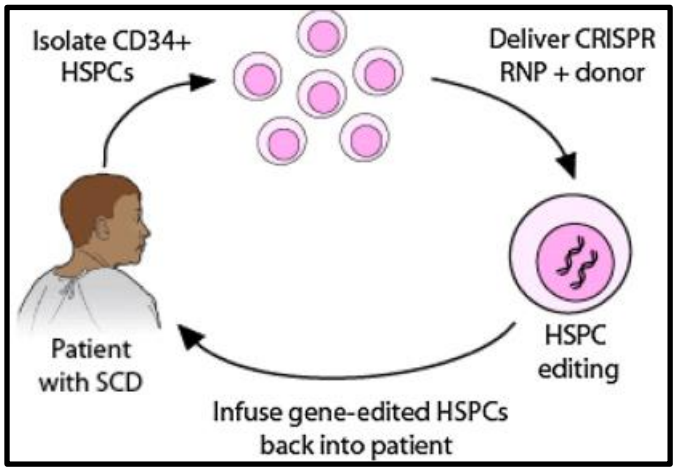
## 1.3 Importance and main drawbacks of gene editing

### 1.3.1 Achievements through CRISPR-Cas

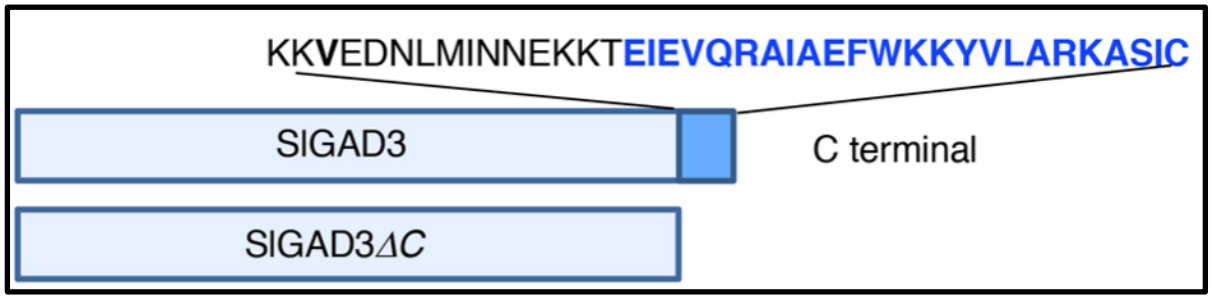
The importance of genetic engineering is clear: these approaches can solve problems like treating inherited diseases, eliminating pests and increasing harvest in the agriculture, providing more food supplies to maintain growing human population and even executing rapid molecular diagnostics. Here, several key gene editing achievements made with CRISPR-Cas will be reviewed, since it is one of the most popular gene editors and is also closely related to TnpB (Sasnauskas et al., 2023).

One of the greatest achievements of CRISPR-Cas9 is a permanent cure for sickle cell disease (SCD) that results from abnormal helmoglobin. SCD causes red blood cells to appear like a C-shaped farm tool called a sickle (Park & Bao, 2021). The CRISPR-based therapy for this disease restores the expression of foetal haemoglobin (HbF) to alleviate the symptoms of SCD. CD34+ hematopoietic stem and progenitor cells (HSPCs) can be isolated from several sources, such as cord blood, bone marrow, and peripheral blood. Isolated cells then are cultured, gene edited to knock-out HbF repressor γ-globin (HBG) and then are returned to the patient so they can generate new, healthy red blood cells potentially curing the sickle cell disease (*fig. 1.3.1 a*).

**a)**




**b)**




**c)**

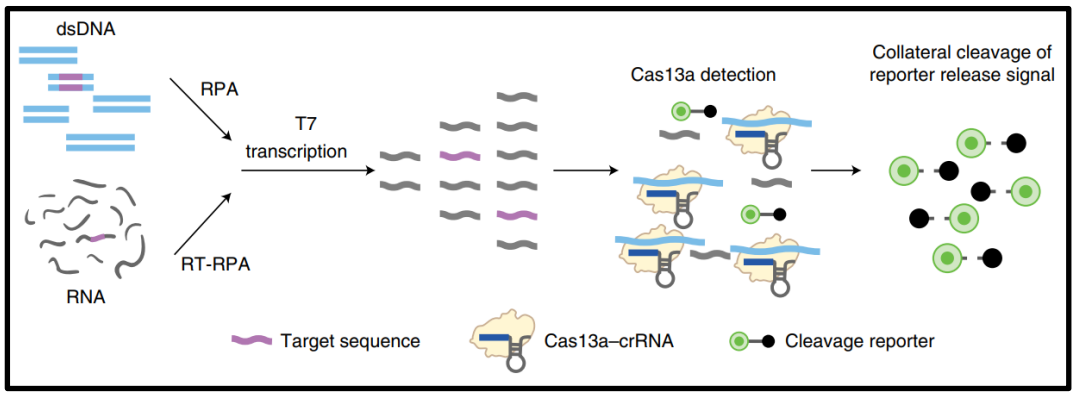



**Figure 1.3.1. Examples of CRISPR-Cas achievements.**

**a)** Genome editing based strategy for treating sickle cell disease (Park & Bao, 2021). At the beginning, CD34+ HSPCs cells are isolated from a SCD patient. The RNP complex of Cas9 and gRNA and DNA donor template are delivered into the nuclei of HSPCs via electroporation. After gene editing, HSPCs are infused back into the patient to reverse the disease phenotype.

**b)** CRISPR-Cas9 induced site-specific targeted mutagenesis of SIGAD3 gene (Lee et al., 2018). This approach causes deletion of the very end of *SIGAD3* gene, that encodes a fragment for C-terminal autoinhibitory domain. Deletion of this domain results in higher activity of the encoded protein and it increases the amounts of GABA γ-aminobutyric acid.

**c)** SHERLOCK viral detection assay (Kellner et al., 2019). Starting with pre-amplification of either a DNA or RNA target input, SHERLOCK uses either recombinase polymerase amplification (RPA) or RT-RPA reverse transcriptase RPA (RT-RPA) and can amplify either RNA or DNA. When amplified targets interact with Cas13−crRNA complexes, they become activated and then cleaves non target RNAs that are attached to fluorescent or colorimetric reporters. The released signal from the reporter molecules can be easily detected using fluorescence-based methods, such as a fluorescence reader or a handheld device equipped with a fluorescent detection module.

More examples of simple, but beneficial, CRISPR victories are engineered foods packed with nutrients. Since September of 2021, Sicilian Rouge tomatoes, which are CRISPR-edited to contain high amounts of GABA γ-aminobutyric acid (Dolgin, 2022), have been sold in the supermarkets of Japan by Tokyo-based Sanatech Seed. Such tomatoes might be beneficial part of a daily diet since GABA γ-aminobutyric acid functions to reduce neuronal excitability by inhibiting nerve transmission and the lack of it is associated with poor quality of sleep and increased anxiety (Yamatsu et al., 2016). Before Sicilian Rouge tomatoes have been launched, an attempt to CRISPR-edit them was already demonstrated by various groups of scientists. For example, Jeongeun Lee and colleagues published their research in 2018 where they study genes related to the synthesis and degradation of GABA γ-aminobutyric acid (Lee et al., 2018). They show that CRISPR-Cas9 induced site-specific targeted mutagenesis of *SIGAD3* gene which is involved in regulation of the amount of GABA γ-aminobutyric acid in the tomato, results in higher amounts of this compound (*fig. 1.3.1 b*). Another CRISPR-editing approach to enrich everyday food products was done by Jie Li and colleagues as published in 2022 (J. Li et al., 2022): they leveraged CRISPR-Cas9 technology to increase vitamin D3 levels by turning off the Sl7-DR2 enzyme, which normally metabolizes 7-DHC. This edit increased levels of 7-DHC in the tomato plants, which is converted to vitamin D3 through exposure to UVB light.

One of the latest applications of CRISPR-Cas is virus detection. A method called SHERLOCK (Specific High-sensitivity Enzymatic Reporter unLOCKing) uses the Cas13 enzyme, a gRNA that targets a sequence of viral RNA, and a fluorescent or colometric reporter (Kellner et al., 2019). To enable detection of viral RNA, SHERLOCK incorporates reporter molecules that are linked to the non-target RNA molecules in the sample. When Cas13 encounters the targeted viral RNA, it starts exhibiting trans-cleavage activity and then it cleaves nearby single-stranded RNA molecules, resulting in a nonspecific collateral cleavage of a fluorescent or colorimetric reporter molecule. The resulting signal can then be detected and quantified, providing a positive result for the presence of the targeted viral RNA (*fig. 1.3.1 c*).

SHERLOCK has been successfully used to detect RNA from a variety of viruses, including Zika virus, Dengue virus, and SARS-CoV-2, the virus that causes COVID-19. The technique is highly sensitive, with the ability to detect very low levels of viral RNA, and is also relatively quick, with results available in less than an hour (Zahra et al., 2022).

Overall, the importance of CRISPR based technologies is indisputable. A wide range of adaptation of CRISPR systems empowers new approaches in many fields and only a few of

them are outlined in this subsection. Nevertheless, it must be remembered, that gene editing comes with several issues and risks that have to addressed in order to avoid unintended damage.

### 1.3.2 Main issues and risks

Gene editing comes with several drawbacks, so scientists are still working on determining the risks and effectiveness of using gene editors on people. The most common burdens of gene editing are safety, delivery of the gene editing tool and successful application. In this subsection, each of these burdens will be reviewed.

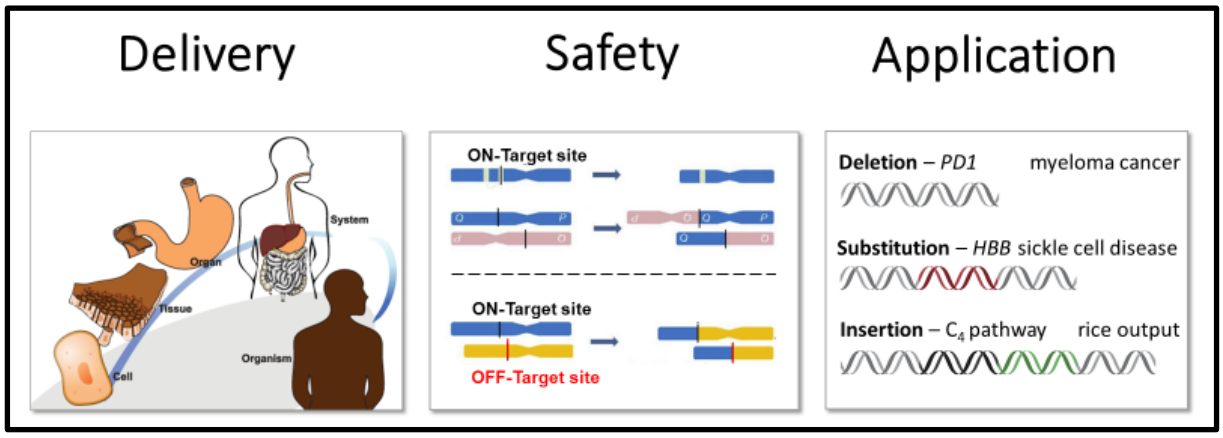



**Figure 1.3.2. The most common burdens for gene editing (Turchiano et al., 2021).**
Burdens are related to establishing efficient delivery to the target cells or tissues, avoiding non-targeted genes or regions, and applying specific gene edits.

Starting with delivery problem, for gene editing tools to effectively function in therapeutics, they must be transported directly to the cells in need of editing. Currently, gene editors are usually delivered via two methods – viral and non-viral vectors (Caffery et al., 2019). The three key viral vector strategies are based on adenoviruses, adeno-associated viruses, and lentiviruses, and they have shown high efficiency in gene delivery and expression (Bulcha et al., 2021). However, the main concern with them is their potential for causing oncogenesis due to targeting the wrong locus in the genome of the cell, *i.e.* off-target effects. This concern also relates to the safety issue. Another concerns are cell's immunogenicity against viral particles and the limited capacity of packaging as mentioned before (*subsection 1.2.5*). Non-viral delivery methods have several advantages over viral methods such as lower immunogenicity and they can deliver either a plasmid or RNP. The examples of non-viral delivery vectors can be loosely grouped as plasmid DNA, lipoplex nanoparticles (liposome-DNA complexes), and polyplex nanoparticles (polymer-DNA complexes) (Du et al., 2014). Despite the advantages, it usually results in low delivery efficiency and that is a barrier for

these approaches (Barrangou, 2013). In addition, non-viral vectors do not promise no off-targets. Since, the transduction of the target cell with viral vector is more effective than transfection of the target cell with non-viral vectors, researchers are looking for the ways to overcome the issue of viral-vector packaging capacity.

As already mentioned, some gene editing strategies might result in off-target effects and become a safety issue. In other words, off-targeting means unintended DNA modifications that arise when a gene editing tool recognizes and modifies DNA sequences that are similar, but not identical, to the intended target. Speaking of CRISPR-Cas, off-targets are sequences that do not fully match gRNA sequence: they might contain an additional nucleotide or a few (insertion relative to the gRNA sequence), might lack a nucleotide or a few (deletion relative to the gRNA sequence) and might contain a different nucleotide or a few (mismatch relative to the gRNA sequence). Interestingly, some off-targets are more tolerated, meaning that the cleavage proceeds. For example, Cas9 tolerates mismatches better than insertions or deletions (Stephen K. Jones et al., 2021). Mutations introduced at non-target genes can alter their function or disrupt their regulation (Höijer et al., 2022). It is crucial to minimize off-target effects to ensure the safety and precision of gene editing applications. It is also known that off-targeting depends on several experimental factors such as the concentration of the gene editor, delivery method, or specific properties of the cells being involved in the experiment (Höijer et al., 2022). It is also important to proceed a computational analysis to find more about the similarity of the sequences of the target and genomic DNA. Overall, CRISPR systems are highly specific, but can still exhibit off-target effects, even when used with certain variations or modifications (Guo et al., 2023). Also, computational analysis can help to predict and avoid off-target effects by helping users select efficient gRNAs with fewer predicted off-target effects (Alkan et al., 2018).

While further discussing problems that come within gene editors, researchers are looking for ways to make them more applicable. Difficulties relating applicability might lay in such factors as gene copy numbers, editing of essential genes and DNA accessibility *in vivo*.

The number of gene copies within a cell vary due to the organism's ploidy or copy number variation (CNV) and that is very common in plants (Zmieńko et al., 2014). If an effect of a particular gene has to be removed, it is best that all of the copies should be knocked out, but it might be hard to achieve if the real CNV is unknown (Gil-Humanes et al., 2017). If knock-out does not affect all the copies of a gene, the unaltered copies could still be expressed.

Another difficulty arises with essential genes. Cells rely on essential genes in order to survive, and in different cell types, the essentiality of a particular gene varies. Disruption of

such genes can lead to severe consequences, such as developmental defects or death (Okita et al., 2004). If an essential gene is edited, it could lead to developmental defects, impaired physiological function, or even death of the organism. Editing of essential genes needs to be done carefully as only very specific editing outcomes are acceptable.

DNA accessibility determines how easily editing tools can access and manipulate specific regions of DNA (Handelmann et al., 2023). If the DNA is tightly packed or inaccessible, the gene editing tool might not be able to access the target site, preventing editing. Notably, epigenetic modifications not only lead to dense chromatin formation, but DNA methylation can reduce access of CRISPR-Cas nucleases to the target site (Wu et al., 2014). This problem can have simple solutions – in some *in vitro* applications, researchers can subject DNA to techniques that increase accessibility. This include chemical treatments, digestion with restriction enzymes, or using DNA helicases to unwind and loosen the DNA structure, making it easier for the gene editing tool to access the target site.

Overall, gene editing comes with both advantages and disadvantages. Even so, gene editing is still one of the best (and in some cases, the only) way to deal with such things as genetically inherited diseases, diagnostics, and world hunger problems, so, we should investigate solutions that overcome these burdens.

## 1.4 Classification and origins of CRISPR-Cas

### 1.4.1 Classification

Bases on their effector molecules, CRISPRs are divided into 2 classes. Class 1 CRISPR-Cas systems are the most widespread and diverse, and use an RNA-guided multiprotein complex for recognizing DNA target and destroying it (T. Y. Liu & Doudna, 2020). Class 1 is divided into types I, III, and IV and 12 subtypes. Most common types are type I and type III (Zmieńko et al., 2014). Both contain proteins responsible for pre-crRNA processing, crRNA and target binding, target cleavage, spacer insertion, and regulation (Taylor et al., 2021). The different subtypes of type I have effector molecules made up of different components and they only target DNA. The signature gene of type I is *cas3*, which encodes Cas3 for unwinding double-strand DNA and RNA-DNA complexes, cutting and degrading the targets (Zheng et al., 2020). The signature gene of type III systems is *cas10*, which encodes Cas10 with a Palm domain for RNA recognition, and a cyclase domain responsible for cutting the target (Kolesnik et al., 2021). Little is known about Class 1 type IV. Type IV systems encode Cas proteins that combine with small RNAs to form multi-subunit ribonucleoprotein complexes, but overall lack

nucleases, integrases, and other genetic features observed in other CRISPR systems (Taylor et al., 2021).

CRISPR-Cas Class 2 systems are more common in gene editing technologies due to their single effector molecule that makes them easier to harness for gene editing purposes. Despite this, they only represent 10% of discovered CRISPR/Cas systems (Shmakov et al., 2017). There are 3 types and 9 subtypes in Class 2. Type II system are the most common, and are well characterized by Cas9 nuclease and the Cas1and Cas2 proteins required for spacer acquisition (Mosterd & Moineau, 2020). Type V is characterized by Cas12 nuclease (Tong et al., 2021). While type II and type V differ by structure and gRNA features, both types target DNA for cleavage. Meanwhile, type VI systems are characterized by a Cas13 protein that targets RNA (Kellner et al., 2019).

### 1.4.2 Hypothesis of the origins

The concerns of CRISPR-based gene editing has lessened with positive clinical trial results, and currently CRISPR systems have proven to be beneficial and applicable. We know their molecular structure, biochemical features, *in vitro* performance, and even their original function *in vivo* in bacterial and archaeal cells. A little bit less explored area CRISPR's origins, though there are some hypotheses regarding the evolution of different CRISPR types: theories of origins involve casposons, introns and transposons.

A popular hypothesis explaining the origins of CRISPR systems, and especially the spacer acquisition, involves casposons (Krupovic et al., 2017). Casposons are self-replicating mobile genetic elements (MGEs) of bacterial and archaeal genomes. They depend on enzymes called casposases that share homology with CRISPR/Cas1 proteins. Casposases show transposase activity and preference for integration sites in the genome (Wright et al., 2019). They also show similar target specificities as CRISPR/Cas spacer acquisition but do not require Cas2. Interestingly, type V CRISPR systems have 2 subtypes (V-C and V-D) that lack the *cas2* gene. It is assumed that Cas1 protein of CRISPR type V-C and type V-D systems functions on its own as a DNA integrase.

Another theory suggests intron-related origins. CRISPR type III targets both DNA and RNA, suggesting that they recruit a reverse transcriptase (RT) for spacer acquisition from RNA and that RT might be adopted from a group II intron (Koonin & Makarova, 2017). In type III CRISPR systems, RT usually forms a fusion protein with Cas1 or is encoded by a separate gene adjacent to *cas1*. Phylogenetic analysis indicates that most of the CRISPR-

associated RTs form a monophyletic group (a group that includes an ancestral taxon and all of its descendants) that is affiliated with the RTs of group II introns (Silas et al., 2017). Theoretically, the RT-Cas1 fusion protein emerged at a single point in evolution due to random insertion of a group II intron into a type III CRISPR-cas locus.

Another popular theory involves transposon-encoded nucleases. Even though the RuvC domain is present in both type II and V effector proteins, the sequence similarity of RuvC domains in Cas9 and Cas12 is pretty low. Furthermore, the HNH domain of Cas9 cleaves the target DNA strand (the strand complementary to the gRNA) whereas the RuvC domain of Cas9 cleaves the nontarget strand (Gasiunas et al., 2012). The NUC lobe of Cas12 family members does not contain HNH, but contains RuvC that is capable of cleaving both target and nontarget strands (Swarts et al., 2017). Interestingly, the RuvC domain of type V CRISPR nucleases shows high sequence similarity to TnpB proteins of the IS200/IS605 and IS607 families of transposons (Sasnauskas et al., 2023). In addition, Cas12 nucleases of the subtype V-F are similar in size or only slightly larger than typical, transposon-encoded TnpBs and have transposon related genes next to their CRISPR array.

The hypotheses described in this subsection reveal contributions of different MGEs to the origins of CRISPR systems and involves two unrelated families of DNA transposons (casposons and TnpB-encoding transposons) and retrotransposons (group II introns). Overall, it is possible that all defense systems found in prokaryotic genomes and capable of doing manipulations with DNA (or RNA, respectively) are evolutionary linked to MGEs.

## 1.5 Transposon-associated nuclease

### 1.5.1 Definition and function

One of the main theories explaining the origins of CRISPRs is based on TnpB. The scientific research and its results outlined in this thesis are also based on TnpB, therefore, this last subsection of the literature review is dedicated to this transposon-associated nuclease.

TnpB is the smallest RNA-guided DNA endonuclease known so far (Karvelis et al., 2021). Families of MGEs that TnpB belongs to (IS200/IS605 and IS607) consist of subterminal left end (LE) and right end (RE) palindromic elements and various configurations of transposition-essential genes (*fig. 1.5.1*). The well-characterized IS200/IS605 family of *Deinococcus radiodurans* ISDra2 consists of partially overlapping *tnpA* and *tnpB* genes flanked by LE and RE elements.

The *tnpA* gene encodes a small TnpA transposase which belongs to the HUH family (family of endonucleases so named for a conserved histidine–hydrophobic residue–histidine motif) and consists of 140 amino acids (Hickman et al., 2010). TnpA excises a DNA strand near a 5'-TTGAT target sequence to form a single-stranded transposon circle that is then integrated to the 3'-TTGAT target in a new location. This completes the transposition cycle without duplicating the target site. This kind of transposition is usually described as a 'peel and paste' mechanism. The function of the *tnpb* gene remained unclear for a long period after discovering ISDRa2 transposons, but recent cell-based experiments with TnpB showed that TnpB is capable of cleaving the donor joint that would be generated after transposon excision by TnpA (Karvelis et al., 2021). TnpB-mediated DSB potentially triggers HDR to reintegrate an excised transposon back into its original site. So, TnpA is responsible for excision of the transposon and insertion of it into a new site, while TnpB is responsible for cutting DNA in a transposon lacking allele to trigger a repair mechanism that copies the transposon back into the original position. This behaviour of TnpB is similar to that 'homing' processes of LAGLIDADG family of homing nucleases (*subsection 1.2.2*) and provides a retention mechanism for the MGE, if the integration step of the excised transposon fails. This helps ensure that both daughter cells will have identical DNA copies.

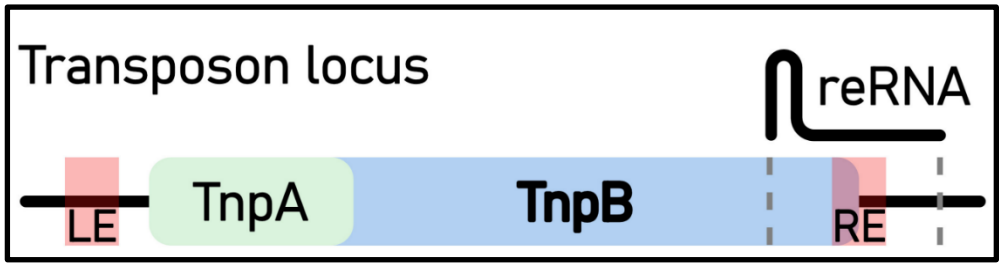


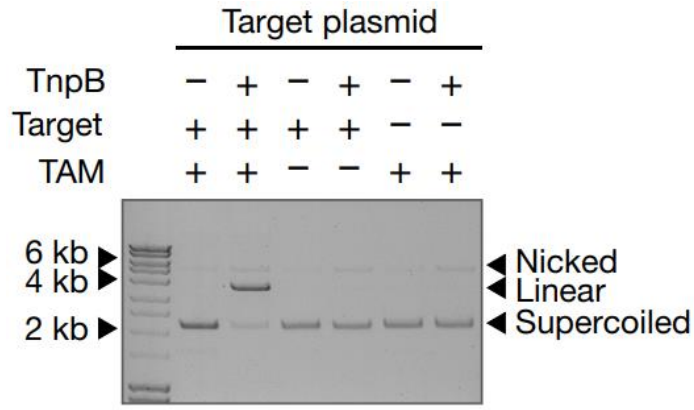

**Figure 1.5.1 Structure of Deinococcus radiodurans ISDRA2 transposon (Jones, unpublished).**
Deinococcus radiodurans ISDra2 transposon encoding sequence consists of partially overlapping *tnpA* and *tnpB* genes that are framed by LE and RE elements. The sequence coding the reRNA overlaps with both TnpB and the RE.

### 1.5.2 TnpB can do manipulations with DNA

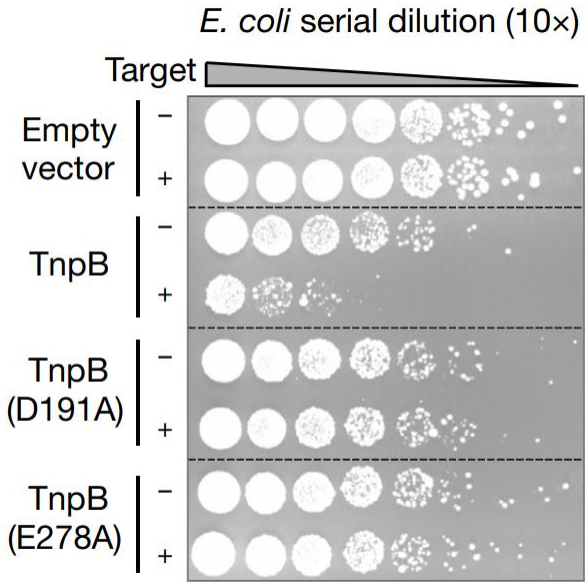
As already mentioned, TnpB can do manipulations with DNA. It can cleave DNA both *in vitro* and in *E. coli*, and it can edit the human genome (*fig. 1.5.2*). First, the experiment of DNA cleavage *in vitro* proved that TnpB, just like Cas nucleases, depends on a TAM` (PAM for TnpB) sequence. Its 5'-TTGAT-3' TAM matches the target site sequence required for TnpA-mediated transposon excision and insertion. TnpB generates 5' staggered overhangs after target cleavage. Second, researchers proved that TnpB induces DNA interference in *E. coli*. Cells were transformed with two different plasmids: TnpB RNP expression plasmid and

a separate second plasmid containing TAM-flanked target, antibiotic resistance genes (for kanamycin (Kn) and carbenicillin (Cb)). It was seen that the colonies stop growing on media with antibiotics after TnpB is expressed and cleaves the resistance-inducing plasmid. Third, sequencing results of DNA extracted from HEK293T treated with TnpB and reRNA showed insertions and deletions (indels) at the targeted sites. These results indicated that DSB repair events and genome editing occurred. In summary, TnpB is a functional nuclease that can be harnessed for genome editing purposes. In addition, TnpB only consists of 408 amino acids and therefore is easy to deliver to cells

**a)**




**b)**




**c)**

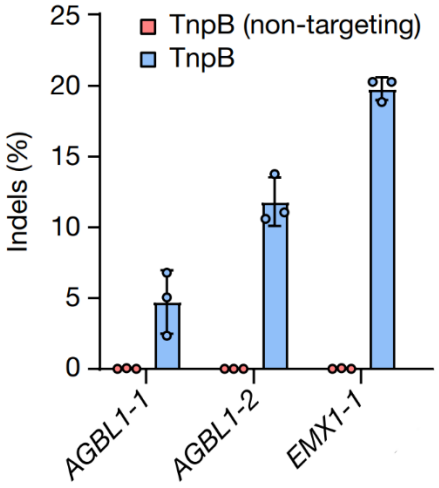



**Figure 1.5.2. Experiments with TnpB proves its ability to edit DNA (Karvelis et al., 2021).**

**a)** Target plasmid cleavage (TAM+/Target+, TAM−/Target+ and TAM+/Target−) by the TnpB RNP complex in vitro. TAM and the target complementary to the reRNA 3′-end sequence are required for plasmid DNA cleavage.

**b)** Interference in *E. coli* experiment and its results. Samples were serially diluted (10×) and the transformants were grown on the media supplemented with Cb and Kn at 25 °C for 44 hours. Results

revealed plasmid interference in the cells containing the native TnpB variant but not TnpBs with the mutated RuvC motif (D191A and E278A). Target '+' or '−' indicates the plasmids with or without the target, respectively.

**c)** Detection of indel activity in tested targets of 20 nt in length in human genomic DNA (represented as the mean of three biologically independent experiments (shown in dots) ± standard deviation). The tested regions belong to *AGBL1* and *EMX1* genes. The TnpB (non-targeting) expression plasmid used as a negative control encodes the reRNA-containing guide sequence that does not match any target in the human genomic DNA.

### 1.5.3 Structure of TnpB

Though TnpB comes from an ancient family of MGEs, its function and structure remained mysterious for a long time. Recent data confirms TnpB's structural similarity to the Cas12 nuclease family (Sasnauskas et al., 2023). TnpB is most structurally similar to Cas12f, whereas other Cas12 subtypes show more distant similarities due to the accumulated numerous insertions. The crystal structure of Isdra2 TnpB revealed that it operates as a dimer, with each monomer consisting of these domains: an N-terminal recognition (REC) lobe for DNA binding and a C-terminal nuclease (NUC) lobe. The REC lobe includes WED and REC domains. The WED domain consists of seven-stranded β-barrel motif, similar to those found in Cas12 enzymes. The REC domain consists of three α-helices and is inserted between the β1 and β2 strands of the WED domain. The NUC lobe includes RuvC domain that consists of a five-stranded mixed β-sheet flanked by two α-helices and the zinc-finger protein (ZnF) domain. The reRNA of TnpB consist of ~150 bp a single-stranded (ss) 5′ terminus followed by a 116-nt structured reRNA scaffold and a 16 nucleotides length (16-nt) 3′-terminal guide segment that changes its conformations when bound to target DNA (*fig. 1.5.3*)

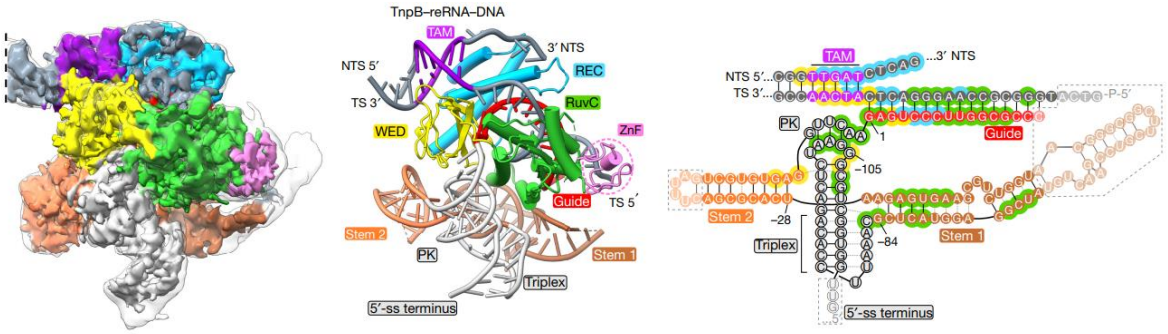



**Figure 1.5.3 Tertiary complex of TnpB-reran-DNA (Sasnauskas et al., 2023).**
TnpB (left) forms tertiary complex with reRNA and target DNA (centre). reRNA and target DNA have corresponding motives to those of TnpB (right). In left and centre parts: WED domain (yellow), Rec domain – Caribbean blue, RuvC – green, ZnF – pink. Left part: Caribbean blue contacts with REC domain, yellow – with WED domain, and green – with RuvC domain. Red marks 16-nt 3'-terminal guide part of ReRNA. Magenta marks the TAMs of the target DNA. NTS and TS stands for nontarget and target strands, respectively. When forming complex with target DNA, reRNA folds into stems 1 (brown) and 2 (orange). Dashed rectangles indicate unresolved parts of the RNA or DNA.

### 1.5.4 Motivation for further research

In conclusion, TnpB exhibits activity and structure similarities with CRISPR/Cas nucleases. Those similarities fulfil the CRISPR origins from MGEs theory. In addition, TnpB can compare to the size of CRISPR/Cas miniature nucleases of type V-F and that makes it more attractive as a gene editing tool due to simpler delivery to the target cells. Because of its small size, nuclease activity and similarities to CRISPR/Cas nucleases, TnpB becomes worth of a further research that could determine its suitability for gene editing and help answer such questions as how specific TnpB is to its targets, how fast it cleaves them ( what cleavage rates it exhibits), what cleavage site within a target it chooses, what DNA ends (sticky or blunt) it leaves after cleaving the target and what environmental conditions lead to the best results.

This study outlines some of the answers to the questions listed above and a plan for future experiments to answer the remaining ones (*subsections 3 and 4*). Here, a method called NucleaSeq is presented as the main tool for determining the specificity, cleavage rates and sites of TnpB together with the kinds of DNA ends it leaves after the cleavage (Stephen K. Jones et al., 2021). According to Jones and co-authors, NucleaSeq is 'a rapid, massively parallel, *in vitro* platform that measures the cleavage kinetics of CRISPR–Cas nucleases. By combining DNA digestion by a nuclease with NGS it captures cleavage products for large libraries of gRNA-matched and mispaired DNA sequences.' This study presents the preparation steps for NucleaSeq by outlining features of TnpB that has to be known before moving to this experiment. It also outlines future plans that include NucleaSeq together with hypotheses about what we could learn about TnpB through this method. In summary, applying NucleaSeq for study of TnpB will help to understand how safe this enzyme is for gene editing purposes and could it become one of the mainly used gene editing tools.

## 2. Methods and materials

### 2.1 Materials

#### 2.1.1 DNA sequences

DNA library D with perfect and modified targets (Purchased from GenScript; *Supplementary data, subsection 1*): standardized Cas9/12a targets D and E (Stephen K. Jones et al., 2021) were modified to have TnpB TAMs.

DNA target D (Purchased from Metabion): matches TAM + perfect target flanked by left and right elements (*subsection 2.1.1; Supplementary data, subsection 1*).

#### 2.1.2 Reagents for amplifying DNA sequences

Primers (Purchased from Integrated DNA Technologies)

10 mM dNTP (Thermo Scientific; Catalog number: R0192)

Ultrapure water

Polymerase and buffer: Thermo Scientific Phusion Plus DNA Polymerase (Catalog number: F630L)

**Table 2.1.2. Sequences of primers used for PCR amplification.**

| Primer | Sequence | Initial concentration |
|---|---|---|
| 1.52 (Forward) | 5’-AACCGCCGAATAACAGAGT-3’ | 100 μM |
| 1.53 (Reverse) | 5’-AAGAACGCCTCGCACACT-3’ | 100 μM |

#### 2.1.3 Reagents for purifying amplified DNA libraries

Thermo Scientific GeneJET PCR Purification Kit (Catalog number: K0702)

#### 2.1.4 Reagents for choosing buffers for the best performance of TnpB RNP D

RB1 (Karvelis et al., 2021): 10 mM Tris-HCl (pH 7.5; Sigma Aldrich, CAS number: 1185-53-1), 10 mM $MgCl_2$ (Fisher Scientific, CAS number: 7786-30-3), 1 mM DTT (Fisher Scientific, CAS number: 3483-12-3), 1 mM EDTA (pH 8; Carl Roth, CAS number: 60-00-4), 100 mM NaCl (Carl Roth, CAS number: 7647-14-5).

RB2 (Stephen K. Jones et al., 2021): 20mM HEPES (pH 7.5; Fisher Scientific, CAS number: 75277-39-3), 150mM KCl (Carl Roth, CAS number: 7447-40-7), 10mM $MgCl_2$, 2 mM DTT.

Target D plasmid (assembled and provided by colleagues Tautvydas Karvelis and Gytis Druteika from the Department of Protein – DNA Interactions, Biotechnology Institute of LSC): pGD41 pUC18 TnpB TAM + Target D (2759 bp) (Karvelis, Tautvydas and Druteika, Gytis, 2021).

TnpB RNP D (TnpB + reRNA D; 10 μM; purified from *Deinococcus radiodurans* and provided by colleagues Tautvydas Karvelis and Gytis Druteika from the Department of Protein – DNA Interactions, Biotechnology Institute of LSC).

### 2.1.5 Reagents for the active fraction titration of TnpB RNP D and comparing TnpB RNP's activity at 22 °C and 37 °C assays

RB1 10x, TnpB RNP D, DNA target D, STOP buffer I (components: 4 parts of 0.5 M EDTA (pH 8) and 1 part of Proteinase K; Thermo Scientific, Catalog number: EO0492).

### 2.1.6 Reagents for assay of TnpB RNP D reRNA quality

TnpB RNP D**,** STOP buffer I.

### 2.1.7 Reagents for the NucleaSeq cleavage

RB1 10x, TnpB RNP D, DNA library D, STOP buffer II (components: 4 μL 0.5 M EDTA, pH 8 and 10 μL Proteinase K).

### 2.1.8 Liquids for DNA Precipitation

Ice cold ethanol (Honeywell, CAS Number: 67-63-0): 100% and 75%.

3 M Sodium acetate solution (Sigma Aldrich, 99% powder, CAS number: 127-09-3).

### 2.1.9 Reagents for Electrophoresis

Buffers: TBE buffer (10X), TAE buffer (10X).

Reagents for making 10% polyacrylamide gels (for making 2, 1 mm width gels): 7.8 mL of ultrapure water, 3mL of acrylamide (Carl Roth, acrylamide/bis-solution, concentration: 40%, UN number: 3426), 1.2 mL of TBE 10x, 200 μL of 10% APS (Carl Roth, made from 99% powder, CAS number: 7727-54-0), 10 μL of TEMED (Carl Roth, 99%, CAS number: 110-18-9).

Reagents for making 12% urea-PAGE gels (Denaturing urea polyacrylamide gel electrophoresis; for making 2, 1 mm width gels): 3.75 mL of ultrapure water, 3.75 mL of

acrylamide, 1.25 mL of TBE 10X, 5.25 g urea (Sigma Aldrich, 98% powder, CAS number: 57-13-6), 62.5 µL of 10% APS, 12.5 µL of TEMED.

Reagents for making 2% agarose gel in TBE: 2g of agarose powder (Thermo Scientific Top Vision Agarose, Catalog number: R0492), 98 mL TBE 1X (made from TBE 10X solution).

Reagents for making 1% Agarose gel in TAE: 1g of agarose powder, 99 mL TBE 1X (made from TBE 10X solution).

Loading dyes for DNA and RNA samples: 6X Purple Gel Loading Dye (New England Biolabs, Catalog number: B7024S), 2X RNA Loading Dye (New England Biolabs, Catalog number: B0363S).

DNA and RNA ladders: 1 Kb Plus DNA Ladder (Thermo Fisher Scientific, Catalog number: 0787018); GeneRuler Ultra Low Range DNA Ladder (Thermo Fisher Scientific, Catalog number: SM1213); Low Range ssRNA Ladder (New England Biolabs, Catalog number: N0364S).

Dyes for post-staining gels after electrophoresis: SYBR Safe DNA Gel Stain (Invitrogen, Catalog number: S33102), SYBR Gold Nucleic Acid Gel Stain (Invitrogen, Catalog number: S11494).

### 2.1.10 Reagents for Bioanalyzer

Bioanalyzer High Sensitivity DNA Analysis (Agilent, Part number: 5067-4626).

## 2.2 Methods

### 2.2.1 Designing DNA sequences

DNA library D: standardized Cas9/12a targets D and E (Stephen K. Jones et al., 2021) were modified to have TnpB TAMs.

TAM + target was flanked with the left and right elements (*table 2.2.1*).

**Table 2.2.1. Different parts of DNA library D construct**

| Left primer | Left barcode | Left buffer | TAM | Target | Right buffer | Right fill | Right barcode | Right primer |
|---|---|---|---|---|---|---|---|---|

Left and right primers serve as templates for PCR amplification. They are followed by barcodes that are unique to each target and are used to distinguish each end of the cleaved target apart during the sequencing. Barcodes are followed by the buffers that serve for controlling the same length of DNA, especially in the case of insertions and deletions as they assure the overall lengths of targets are even. Right fill serves for keeping length of each sequence even. The

remaining parts – right and left adapters for time barcoding – are added after the NucleaSeq cleavage during the library preparation and provide a cleavage time resolution of the targets during sequencing.

DNA library D contains of total 6236 sequences and modifications (*subsection 3.5; Supplementary data, subsection 1*).

Linear DNA target D substrate matches TAM flanked with perfect target sequence surrounded with the same left and right elements (*Supplementary data, subsection 1*).

### 2.2.2 PCR amplification

DNA library D was amplified via PCR with Phusion polymerase by following the manufacturer's instructions. At the beginning , the number of PCR cycles and primer annealing temperature matched the manufacturer's protocol (25 cycles and 65 °C), but due to unspecific product appearing in a gel (*subsection 3.5*) several cycles were lowered to 20 and primer annealing temperature was set to 60 °C.

### 2.2.3 Purification of PCR-amplified products

The purification of PCR amplified product was done with a commercial Thermo Scientific GeneJET PCR Purification Kit (Catalog number: K0702) and all the procedure was followed according to the protocol.

### 2.2.4 Buffer assay for best performance of TnpB RNP

Separate reactions (with RB1 and with RB2) were initiated by mixing Target D plasmid (final concentration - 3 nM) with TnpB RNP D (final concentration - 100 nM) and either RB1 or RB2 (final concentration – 1X). Reaction mixes were incubated at 37 °C for 60 min and after that they were stopped by introducing STOP buffer I into samples (1 part of STOP buffer I : 4 parts of the reaction mix) and incubating them at 37 °C for 60 min. Samples were run on 1% agarose gel in 1X TAE.

### 2.2.5 Active fraction titration of TnpB RNP D

Active fraction titration assay was divided into 5 samples where Metabion target D DNA concentration was constant (final concentration in each of 5 samples – 1 nM), but concentration of TnpB RNP D was varying (*table 2.3.5*). Buffer used to carry out the assay - RB1 (final concentration – 1X).

**Table 2.3.5.** Active fraction titration assay with Metabion target D DNA and TnpB RNP D was divided into 5 samples in which DNA concentration stayed stable, but RNP concentration varied.

| | | | | | |
|---|---|---|---|---|---|
| Final target DNA concentration | 1 nM | 1 nM | 1 nM | 1 nM | 1 nM |
| Final TnpB RNP D concentration | 1 μM | 300 nM | 100 nM | 30 nM | 10 nM |
| RB1 | 1X | 1X | 1X | 1X | 1X |

All 5 samples were incubated at 37 °C for 60 min and then reactions were stopped by introducing STOP buffer I into samples (1 part STOP buffer I : 4 parts of the reaction mix) and incubating them at 37 °C for 60 min. Samples were run on native 10% PA-gel in 1X TBE.

### 2.2.6 Determining TnpB RNP D activity and cleavage rate

To estimate the activity of TnpB's RNP D, GelAnalyzer 19.1 (Istvan, n.d.) software was used to determine the DNA band intensity so the intensities between intact and cut DNA bands could be compared. The software allows to convert the DNA band intensity to the raw volume numbers that are then normalized to be in a range from 0 to 1, using the formula below:

$$Normalized\ value = \frac{\frac{Uncut_t}{(Uncut_t + Cut_t)} - \frac{Uncut_f}{(Uncut_f + Cut_f)}}{\frac{Uncut_0}{(Uncut_0 + Cut_0)} - \frac{Uncut_f}{(Uncut_f + Cut_f)}}$$

In the formula, t stands for the raw volume at a timepoint, f is the raw volume for the final timepoint and 0 is the raw volume from timepoint 0.

Once the data for one cleavage assay was normalized, then it was fitted into an exponential decay function to give a cleavage rate in $s^{-1}$. Two models were used to obtain the best function fit to the dataset: $e^{-kt}$ or $e^{-kt}$ +c, where k is the cleavage rate, it is time, and c is horizontal asymptote.

The cleavage rate of TnpB RNP D was calculated using an ipynb (Interactive Python Notebook) script (Steven Knox Jones, 2023).

### 2.2.6 NucleaSeq cleavage of target library DNA

Cleavage reaction was initiated by mixing DNA library D (final concentration – 1 nM) with TnpB RNP D (final concentration - 1 μM) and 10X RB1 (final concentration – 1X). Reaction was incubated at 37 °C and a sample was taken out of it according to the time points

(*table 2.3.6*) and then stopped by introducing STOP buffer II into samples (1 part of STOP buffer II : 4 parts of the reaction mix) and incubating them at 37 °C for 60 min.

**Table 2.3.6.** Each sample was taken out of DNA library D and TnpB RNP D mix and then stopped by introducing STOP buffer II at the following time points:

| Time point (min) | 0 | 0.2 | 1 | 3 | 10 | 30 | 100 | 300 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Stopped by introducing STOP buffer II | | | | | | | | | |

### 2.2.7 DNA precipitation

For DNA precipitation, a sample was mixed and vortexed with 0.1 volumes of 3 M sodium acetate and 2.5 volumes ice-cold 100% ethanol. The mixture was precipitated at -20 °C overnight and centrifuged at 13000 rpm in a cold centrifuge (4 °C) for 30 min afterwards. The supernatant was thrown away and DNA pellet was washed twice with 500 µL ice-cold 75% ethanol while centrifuging it at 13000 rpm for 10 min (ethanol was thrown away after each centrifugation). The tube with DNA pellet was centrifuged to spin down the remaining ethanol at 13000 rpm for 1 min. The last drop of ethanol was removed with a pipette and the DNA pellet was air-dried with the tube's lid open for 5 minutes at room temperature. DNA pellet was mixed with 20 µL of ultrapure water.

### 2.2.8 Assay of the quality of TnpB RNP D reRNA

To estimate if TnpB RNP's reRNA is homogeneous or not, TnpB RNP D was incubated with STOP buffer I (1 part of STOP buffer : 4 parts of the RNP solution) at 37 °C for 60 min to degrade the protein and elute RNA. After the incubation, the sample was mixed with 2X RNA Loading Dye and further incubated at 80 °C for 1 min to denature RNA and make it single-stranded. The sample was loaded on denaturing 12% urea-PAGE gel.

### 2.2.9 Electrophoresis

Samples for electrophoresis for analysis of big DNA fragments (Target D plasmid after buffer assay for best performance of TnpB RNP) were run on 1% agarose gel in 1X TAE at 120 V for 30 min.

Samples for electrophoresis for analysis of small DNA fragments (Metabion target D after active fraction titration assay and library D after the NucleaSeq cleavage of target library

DNA) were run on native 10% polyacrylamide gel in 1X TBE at 100 V for 120 min (before loading the samples, the gel is prerun at 50 V for 20 min).

Sample for electrophoresis analysis of RNA homogeneity was run on denaturing 12% urea-PAGE gel at 120 V for 45 min (before loading the samples, the gel was prerun at 100 V for 15 min).

Each of the electrophoresis was run using the power source: PowerEase Touch 120W Power Supply (Thermo Fisher Scientific, Catalog number: PSC120MB).

After electrophoresis, each gel was stained in gel stain-electrophoresis buffer solution for 30 minutes on a plate rocker. 1% agarose gel was stained with SYBR Safe DNA Gel Stain dissolved in 1X TAE. 10% polyacrylamide gels and 12% polyacrylamide-urea gel was stained with SYBR Gold Nucleic Acid Gel Stain dissolved in 1X TBE.

### 2.2.10 Library Preparation for sequencing

DNA target library D was prepared for Illumina sequencing with the following reagents: NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Catalog number: E7770S) and NEBNext Multiplex Oligos for Illumina, Index Primers Set 1 (New England Biolabs, Catalog number: E7335S). All the procedure was followed according to the manufacturer's protocol.

### 2.2.11 Analysis of samples by Bioanalyzer

Samples were analysed by capillary electrophoresis by Agilent High Sensitivity DNA Analysis kit and Bioanalyzer software.

# 3. Results

The main goal of this study is profiling the specificity of TnpB through NucleaSeq. In order to approach this goal, the study was divided into tasks: determining the activity of the purified TnpB RNP D, determining reaction conditions for NucleaSeq cleavage, designing DNA library that includes on- and off-targets, and finally, with further parts of Nucleaseq, determine the cleavage rates and specificity of TnpB RNP D across the library. Results outlined in this section covers some of these tasks: here the activity of the RNP, needed reaction conditions for NucleaSeq cleavage and design of DNA library is described.

## 3.1 Performance of TnpB RNP D in RB1 vs RB2

Two different buffers – RB1 and RB2 - for TnpB RNP D cleavage of DNA were compared as they provide cofactors (magnesium ions), ensure neutral pH of the reaction (pH 7.5) and ionic strength resulting from the salts they include (NaCl in RB1 and KCl in RB2) (*subsection 2.1.4*). Scientific data proves their compatibility with DNA cleaving nucleases (Jones et al., 2021; Karvelis et al., 2021). To estimate which of them makes TnpB perform better, two separate reactions with Target D plasmid and TnpB RNP D, but one of two buffers were carried out at 37 °C for 60 min since such reaction conditions were showed to work for TnpB (Karvelis et al., 2021).

Gel electrophoresis results indicated that RB1 provides suitable conditions for plasmid cleavage by TnpB RNP D while RB2 does not (*fig. 3.1*).

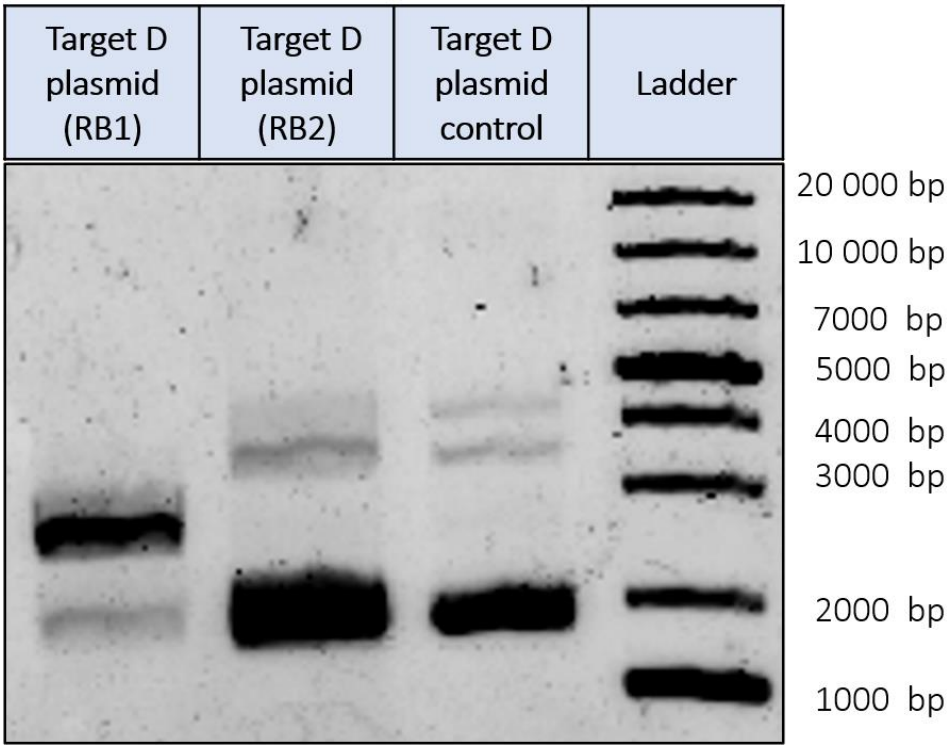



**Figure 3.1. Result of Target D plasmid cleavage by TnpB RNP D in different buffers (RB1 and RB2).** Plasmid cleaved in RB1 (first lane) appears to be linearized, while the plasmid cleaved in RB2 (second lane) has a thick fraction of supercoiled DNA (bottom band). Third lane – intact plasmid (control). Upper bands in the second and third lanes might be fractions of the nicked plasmid.
Gel features: 1% agarose gel in TAE, 1 Kb Plus DNA Ladder, SYBR Safe DNA Gel Stain, Typhoon imaging.

## 3.2 Active fraction titration of TnpB RNP

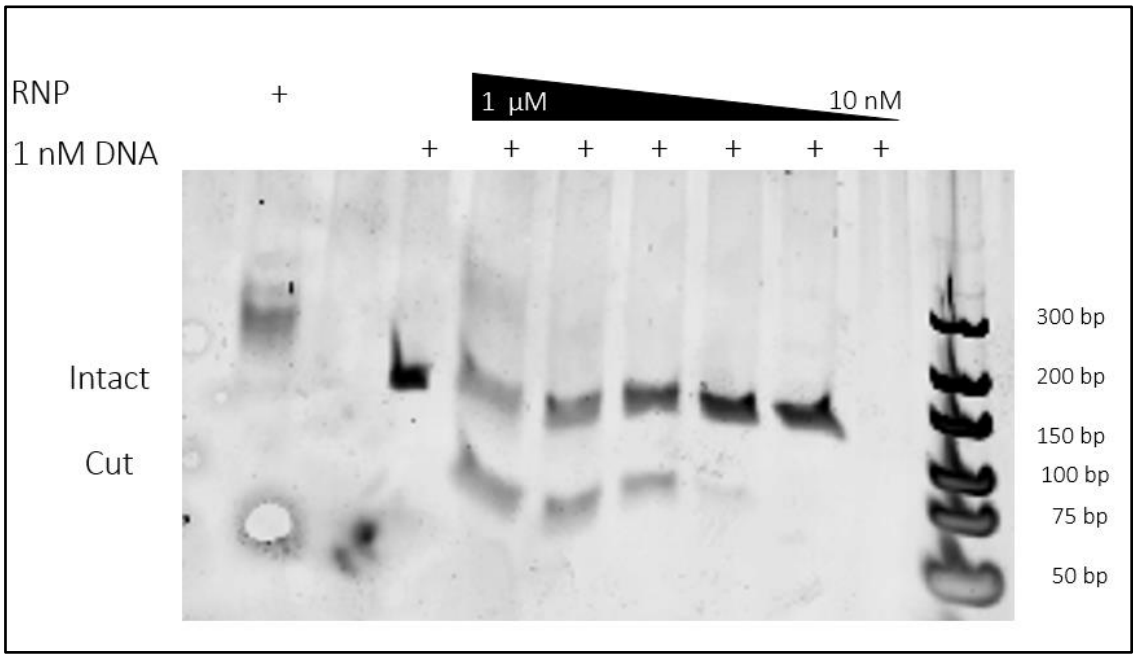
Well-controlled biochemical assays require certain conditions to be met. Since TnpB RNP D was provided by colleagues, one of the main tasks was to determine its activity so the optimal ratio of both TnpB RNP D and DNA for the cleavage reactions could be established.

To determine the fraction of active TnpB RNP D, a linear DNA target D substrate (1 nM) was digested with increasing concentrations of RNP (10 nM – 1 μM) for 60 min at 37°C. Gel electrophoresis revealed incomplete digestion of the DNA even with a 100-fold excess of RNP (*fig. 3.2 a*). It was nonetheless calculated that the fraction of active RNP was ~1% (*fig. 3.2 b*) meaning that in 1 μM of purified TnpB RNP D only about 10 nM is actually active. This activity is substantially lower than that of other purified programmable nucleases (e.g. Cas9, Cas12a) (Stephen K. Jones et al., 2021).

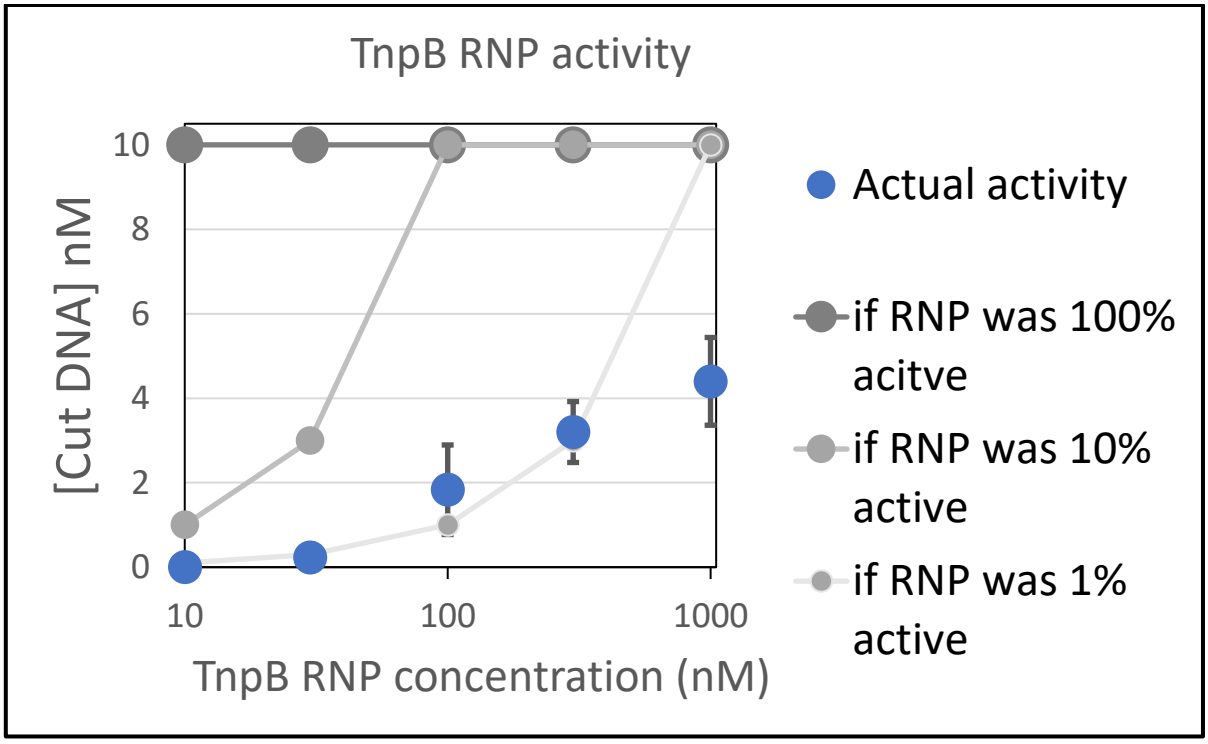
In order to investigate the low activity of TnpB RNP D, it was decided to check TnpB's reRNA homogeneity. Since TnpB is expressed in bacterial cells and purified as RNP (Karvelis et al., 2021), cellular factors like ribonucleases could impact reRNA's quality by degrading it. To determine the quality of RNA, TnpB RNP D was treated with STOP solution I that contains proteinase K which degrades the protein and elutes reRNA from RNP complex. The single-

stranded reRNA was analysed by gel electrophoresis under denaturing conditions. Instead of reRNA of homogeneous length of ~270 nt that is required for optimal TnpB RNP activity (Sasnauskas et al., 2023), RNAs of differing length were observed (*fig. 3.2 c*, *table 3.2*).

**a)**




**b)**



**c)**

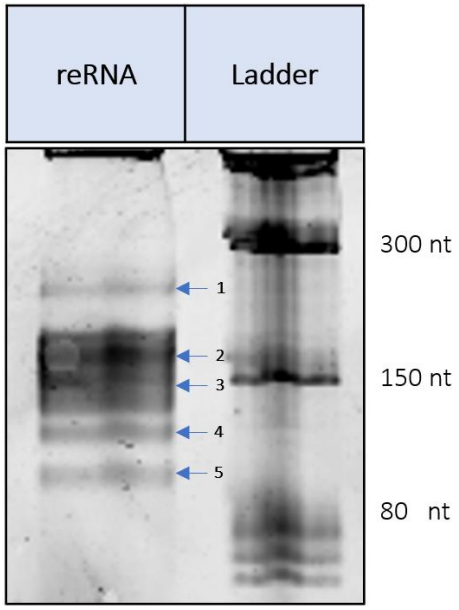



**Figure 3.2. Active fraction titration assay of TnpB RNP D (a), TnpB RNP D activity (b) and inspection of the quality of TnpB RNP D reRNA (c).**
**a)** First lane – RNP control (1 μM), second lane – empty, third lane – DNA control (1 nM), and following lanes – samples with stable DNA, but varying RNP concentrations. Last lane - Ultra Low Range DNA Ladder.
Gel features: 10% polyacrylamide gel in TBE, GeneRuler Ultra Low Range DNA Ladder, SYBR Gold Nucleic Acid Gel Stain, Typhoon imaging.
**b)** X axis corresponds to varying RNP concentrations, Y axis – to the amount of the cut DNA (nM). Blue dots mark the actual obtained activity, grey lines – theoretical activities that show how data would look if RNP was 1% active (light grey, overlaps with blue dots); 10% active (medium dark grey), and 100% active (dark grey).
**c)** Instead of obtaining a fraction of homogeneous reRNAs, there is a bunch of reRNAs that differ in length (marked with blue arrows).
Gel features: 12% urea-PAGE gel in TBE, Low Range ssRNA Ladder, SYBR Gold Nucleic Acid Gel Stain, Typhoon imaging.

**3.2 Table. reRNA fractions differing in length and their ratios over general amount of reRNA.**

| reRNA fraction | Predicted length (nt) | Ratio over total amount of reRNA (%) |
|---|---|---|
| 1 | 270 | 12 |
| 2 | 170 | 15 |
| 3 | 150 | 13 |
| 4 | 115 | 26 |
| 5 | 100 | 34 |

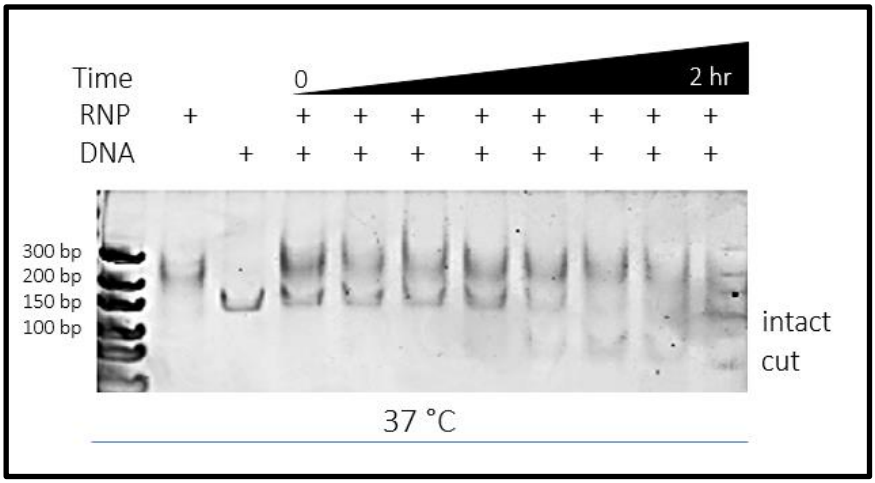
### 3.3 Determining temperature for optimal TnpB cleavage of target DNA

In order to identify an optimal temperature for capturing fast and slow cleavage events with TnpB, a linear DNA target D substrate was exposed to TnpB RNP D over a two hour time-course at 22°C and 37 °C. For this experiment, RNP and DNA concentrations were stable (1 μM of purified TnpB RNP D and 1 nM of DNA) and TnpB RNP D was used in 100-fold excess due to its 1% activity (*fig. 3.2 b*). Analysis of gel electrophoresis revealed that TnpB RNP D has a higher on-target cleavage rate when exposed to 37 °C (*fig. 3.3 a*) than 22 °C (*fig. 3.3 b*).

Cleavage of targets by TnpB follows an exponential decay function (f(t) = e^(-k* t) where k is the cleavage rate, and t is time in seconds (*Supplementary data fig. 2S*). The cleavage rate of the reaction carried out at 37 °C was 0.69 $seconds^{-1}$ and the cleavage rate of the reaction carried out at 22 °C was 0.05 $seconds^{-1}$. That means that the cleavage rate is higher when TnpB

RNP D cleaves at 37 °C, but still within the range of detection. Cleavage of off-target DNAs would likely be too slow at 22°C for upcoming experiments.

**a)**




**b)**

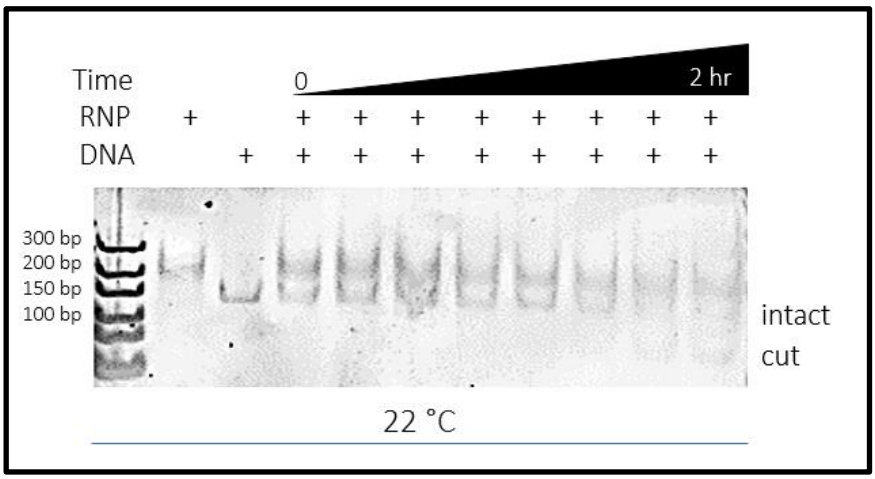



**Figure 3.3. TnpB RNP D mediated cleavage of linear target D DNA at 37 °C (a) and 22 °C (b).**
RNP and DNA concentrations were stable (1 µM of RNP and 1 nM of DNA) but each reaction was divided into varying time points: 0-2 hrs (starting from a third lane: 0, 10, 20, 40, 60, 80, 100, 120 min). Label 'intact' marks the fraction of uncut DNA while label 'cut' marks the fraction of the cleavage product. The results of both reactions show that the fraction of 'cut' DNA starts appearing at sooner time point (60 min) in the reaction carried out at 37 °C (a) while the fraction of 'cut' DNA starts appearing at later time point (100 min) in the reaction carried out at 22 °C (b).
Gel features: 10% polyacrylamide gel in TBE, GeneRuler Ultra Low Range DNA Ladder, SYBR Gold Nucleic Acid Gel Stain, Typhoon imaging.

## 3.4 Designing DNA library

In order to proceed to NucleaSeq experiment and obtain such data as the specificity of TnpB RNP D, cleavage rates across perfect and modified targets, where does it cleave within a target sequence and what DNA ends does it generate (sticky or blunt) a proper DNA library is required. DNA library D was designed using a Python script that generates on- and off-target sequences with TAM- and reRNA-relative alterations (mismatches, insertions or deletions). The library contained the 'perfect' on-target sequence, the set of off-target sequences with alterations at one or two positions within the target's 23 nucleotides, and sequences where altered TAMs replaced TnpB's canonical 5'-TTGAT-3' TAM (*table 3.4*).

**Table 3.4. Examples of different kinds of sequences in Library D and total amount of them.** The library contains a 'perfect' on-target sequence assembled from a canonical TnpB TAM and reRNA-matched target, various sets of 'perfect' TAMs + off-targets, interesting TAMs + on-targets, and positive and negative controls.

| TAM + target combination | TAM | Target | Amount |
|---|---|---|---|
| Perfect TAM + Perfect target sequence | 5'-TTGAT | GTGATAAGTGGAATGCCATGTGG-3' | 173 |
| Perfect TAM + Single deletion target | 5'-TTGAT | _TGATAAGTGGAATGCCATGTGG-3' | 20 |
| Perfect TAM + Double deletion target | 5'-TTGAT | _TGATAAGTGGAATGCCATGTG_-3' | 191 |
| Perfect TAM + Single insertion target | 5'-TTGAT | TGTGATAAGTGGAATGCCATGTGG-3' | 76 |
| Perfect TAM + Double insertion target | 5'-TTGAT | TGTGATAAGTGGAATGCCATGTGGG-3' | 2205 |
| Perfect TAM + Single mismatch target | 5'-TTGAT | CTGATAAGTGGAATGCCATGTGG-3' | 75 |
| Perfect TAM + Double mismatch target | 5'-TTGAT | CTGATAAGTGGAATGCCATGTGT-3' | 2700 |
| Perfect TAM + Complementary stretched mismatch target | 5'-TTGAT | CACATAAGTGGAATGCCATGTGG-3' | 300 |
| Interesting TAM + Perfect target sequence | 5'-TTGAT | GTGATAAGTGGAATGCCATGTGG-3' | 64 |
| Negative control | 5'-CCTGC | AGACCGCTAATGGTGAAACCTGT-3' | 432 |
| Total | | | 6236 |

## 3.5 Amplification of DNA library

In order to have enough material for the NucleaSeq cleavage experiments, amplification of the DNA library was carried out through PCR with Phusion polymerase. Above the expected DNA library band (130 bp), a bigger, unspecific band appeared. In an attempt to reduce such unspecific products, it was decided to reduce PCR amplification cycles (20) and set a higher primer annealing temperature (60 °C), but these changes did not impact the problem (*fig. 3.5*).

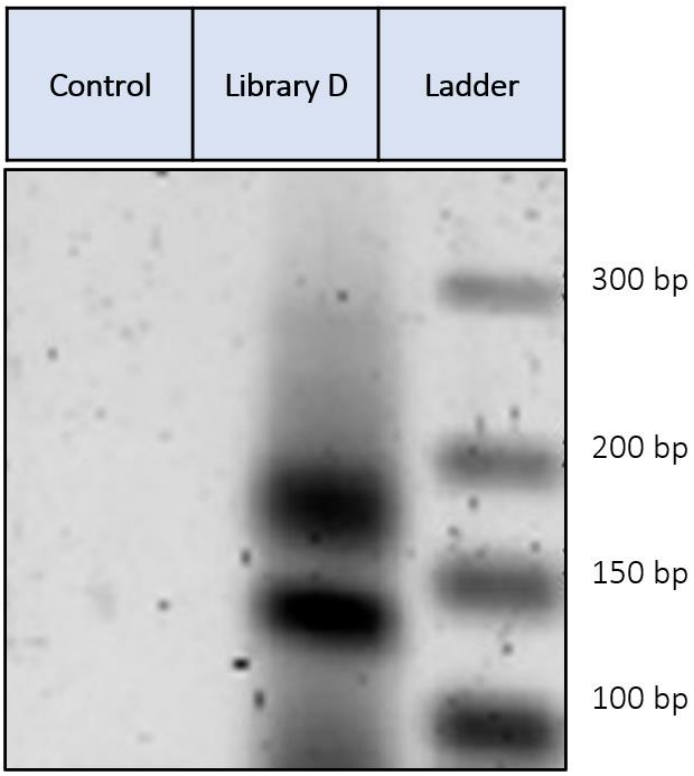



**Figure 3.5. Amplification of DNA library D.** DNA library was amplified through PCR with Phusion polymerase. Modifications of the manufacturer's protocol did not help to avoid the unspecific band (between 150 and 200 bp) appearing above the DNA library band of 130 bp.
Gel features: 2% agarose gel in TBE, GeneRuler Ultra Low Range DNA Ladder, SYBR Gold Nucleic Acid Gel Stain Typhoon imaging.

## 3.6 NucleaSeq cleavage of DNA library D and preparation for sequencing

NucleaSeq cleavage is the first part of whole NucleaSeq experiment that combines obtained cleavage products with NGS and provides data that finally answers the questions outlined in this study. Within a target there is a seed region for TnpB RNP D that contains crucial sequence motifs for reRNA binding. RNP cleaves within a seed region at canonical cut sites (*table 3.6*).

Target D seed:

(14 bp): TTGATGTGATAAGT

**Table 3.6. Canonical cut sites within a target and corresponding lengths of the products**. Seed regions are underlined.

| Canonical cut site within a target | Lengths of products |
|---|---|
| 5'-AACCGCCGAATAACAGAGTATCCGCGCTGGTTGTATATCAGATTGATGTGATAAGTGGAATGC x CATGTGGTCGAGTCGATAGTACTCTGCACATCTCGTAAGGTGGTTCCGGAGTGTGCGAGGCGTTCTT-3' (cuts between 21 nt and 22 nt within a target) | 63 nt + 67 nt |
| 5-'AACCGCCGAATAACAGAGTATCCGCGCTGGTTGTATATCAGATTGATGTGATAAGTGGAATGCCATGT x GGTCGAGTCGATAGTACTCTGCACATCTCGTAAGGTGGTTCCGGAGTGTGCGAGGCGTTCTT-3' (cuts between 26 nt and 27 nt within the target) | 68 nt + 62 nt |

To assure that TnpB RNP D-treated DNA library D was properly prepared for the next-generation sequencing (NGS), samples were loaded on a capillary electrophoresis chip from Agilent Bioanalyzer High Sensitivity DNA Analysis kit and analysed via Bioanalyzer software (*fig. 3.6*).

After library preparation which involves adaptor ligation and PCR, each of the samples are lengthened by additional 140 bp. That means that intact DNAs should be ~ 270 bp (130 bp + 140 bp) and cut DNA products should vary between ~ 202 – 208 bp (63 bp + 140 bp; 67 bp + 140 bp; 68 bp + 140 bp; 62 bp + 140 bp). DNA bands obtained at 140 bp and below could be traces of adaptors or their dimers, DNA bands obtained between above 140 but below 202 could be result of RNP induced trimming and bands obtained above 270 bp could result from improper sample flow on the chip.

Capillary electrophoresis results obtained by Bioanalyzer software (*fig. 3.7*) bring out these observations:

- some of the DNA was lost in a process of doing library preparation
- gel does not seem to be traced with adapter dimers, but one of the lanes (TnpB RNP D – 0.2) contains DNA bands that are too large to be intact DNAs (~270 bp) and that could results from improper sample flow on the chip.
- both bands observed after PCR amplification (*fig. 3.6*) are present in the gel. Without sequencing data, it is impossible to estimate whether the unspecific DNA product gets digested by TnpB RNP D or not

- lanes ‘TnpB D - 3’, ‘TnpB D - 100’ and ‘TnpB D - 1000’ seems to be too faint and lack DNA in general, while ‘TnpB D - 30’ and ‘TnpB D - 300’ lacks lower bands (150 – 170 bp) that could be cleavage products and that are present in ‘TnpB D - 10’. On the other hand, it is possible, that ‘TnpB D – 1000’ does not contain intact DNA fraction because it could get completely cut since it is the longest time point.

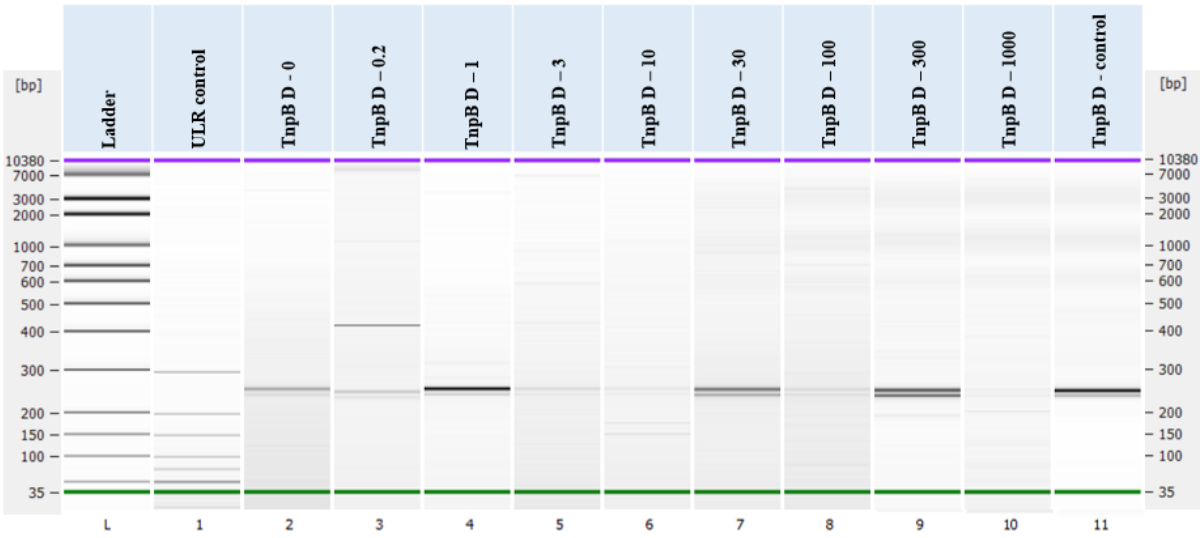



**Figure 3.6. Capillary electrophoresis results by Bioanalyzer.** Lane labeled ‘L’ contains ladder (length of the green band – 35 bp, length of the purple band – 10380 bp). Lanes labeled ‘1’ and ‘11’ contain controls: ‘1’ – ULR*, ‘11’ – intact DNA library. Lanes labeled from ‘2’ to ‘10’ contains samples that correspond to each of the time points.
*It was decided to include ULR control to make sure that all the samples including Bioanalyzer ladder spreads evenly, according to their lengths.

Overall, results outlined in this section cover some of the tasks assigned for the study. The activity of TnpB RNP D was determined to be ~1%. It was also determined that TnpB RNP D cleaves DNA faster at 37 °C than at 22 °C and reaction carried out at 37 °C helps to capture cleavage products and their quantity at different TnpB RNP D concentrations. This section also describes DNA target library D that contains perfect targets matching TnpB‘s reRNA flanked either by perfect or interesting TAM at 5’ position, and thousands of off-target sequences that mispair reRNA (insertions, deletions, mismatches). The design of the library should help to unravel TnpB’s specificity and cleavage rates across on- and off-targets, but for that to happen, NGS and interpretation of its results should be processed.

## 4. Discussion

### 4.1 TnpB RNP performs better in RB1 than in RB2

RB1 is suitable for carrying plasmid cleavage by TnpB RNP D while RB2 is not (*fig. 3.1*). Since components for making both buffers differ (*subsection 2.1.4*) they might have a different impact on enzyme's activity. Both buffers contain magnesium ions that serve as cofactors for enzymatic reactions, but one of the main differences is that RB1 has lower saltiness level than RB2: RB1 contains 100 mM of NaCl, while RB2 contains 150 mM of KCl. There is some scientific data that suggests that higher concentrations of salts might be inhibitory to some enzymes (Weimberg, 1967). Some researchers also claim, that certain concentrations of DTT might have an inhibitory effect to an enzyme (Yi et al., 1999) and RB2 has a higher concentration of DTT (2mM) than RB1 (1 mM).

### 4.2 TnpB RNP activity is ~1%

The active fraction of TnpB RNP D is close to 1% (*fig. 3.2 b*), meaning that within a sample of TnpB RNP D with a concentration of 10000 nM (10 μM), the concentration of an active fraction is only about ~100 nM. There are a few possible reasons for such low activity. As already mentioned, one of them could be a presence of at least 5 reRNA fractions that vary in length (*fig. 3.2 c; table 3.2*). TnpB forms an active RNP with an reRNA of ~270 nt length (Sasnauskas et al., 2023) and reRNA of improper length could be occupying the REC domain of the enzyme and therefore troubling RNP:DNA interactions. This hypothesis is also supported by the fact that the amount of cut DNA gets plateaued while concentration of TnpB RNP D increases. That could mean that some RNP molecules loaded by reRNA of improper length sits on DNA due to binding activity but does not cut it and therefore prevents cutting of DNA by RNPs loaded with reRNA of proper length. Since TnpB RNPs are purified from the cells, separate fractions of reRNAs with different lengths could be result of degradation by cellular RNases. In this case such approach like *in vitro* transcription and translation (IVTT) could serve as solution since it avoids cellular compounds including RNases.

Another possible reason for such low activity of TnpB RNP D could be improper composition of reaction buffer RB1. Despite the fact that RNP exhibits some levels of activity when combined with this buffer, it is possible, that some changes like decreasing saltiness and regulating levels of cofactor ions could result in higher activity. A new composition of reaction buffer should be tested in future experiments.

Low active fraction concentration might seem as an issue that needs to be addressed. Nevertheless, results shown in this thesis prove that when using TnpB RNP in a great excess and low amounts of DNA, cleavage products are generated, but such low input of DNA (1 nM) is hard to recover and precipitate after the cleavage, therefore library preparation becomes complicated.

### 4.3 TnpB RNP cleaves DNA faster at 37 °C than at 22 °C

After comparing how fast TnpB RNP D cleaves DNA at 22 °C vs 37 °C it was obtained that cleavage products start appearing at sooner time points when cleavage reaction is carried out at 37 °C (*fig. 3.3 a*) than at 22 °C (*fig. 3.3 b*) and reaction carried out at 37 °C exhibits higher cleavage rates (*subsection 3.3*). Comparison of results of 8 time points between both reactions shows that cleavage product appears at the fifth time point (60 min) when the reaction is carried out 37 °C and at the seventh time point when the reaction is carried out 22 °C. Since TnpB RNPs are assembled within a *Deinococcus radiodurans* cells and purified from them, the enzyme's obtained higher activity at 37 °C could be related to natural or preferred living conditions of the microorganism. It's already known that the optimal temperature for *Deinococcus radiodurans* cultivation is 37 °C (He, 2009) and there is data that claims that *Deinococcus radiodurans* prefers warm habitats like animal guts, hot springs and deserts (Slade & Radman, 2011).

This data outlines the decision to perform further experiments at 37 °C. It was also decided to not go higher that 37 °C in order to control the speed of the reaction and capture cleavage products and their differing quantities and different concentrations of TnpB RNP D or varying time points.

### 4.4 Unspecific PCR product

The amplification of DNA library D turned out to be successful, but gel data shows that there is a thick fraction of unspecific product above our amplified DNA library D band (*fig. 3.5*). The product is absent in the negative control that does not include any DNA input material so that leads to the assumption that it originates when the DNA template is present. One of the possible reasons for the unspecific product could be library synthesis errors.

Unfortunately, modification of PCR conditions did not help to solve the problem. It was tried to adopt gel purification strategy to get rid of the unspecific product while only keeping the wanted amplified DNA, but both bands are too close to each other and attempt to excise

DNA library D fragment from the gel failed. Until the sequencing is not completed it is hard to tell if unspecific product gets cut during the NucleaSeq cleavage. In case if it does not, it might be possible to computationally sort out sequencing data and remove unspecific DNA from further analysis.

### 4.5 Results of the NucleaSeq cleavage of DNA libraries and plans for the future experiments

Bioanalyzer results prove that a lot of DNA was lost in a library preparation process, but sample 'TnpB D – 10' might be indicating some cleavage products (*fig. 3.6*). This experiment should be repeated after achieving some improvements that are outlined in this subsection.

Overall, the main goal of this study is determining TnpB's target specificity profile with next-generation biochemistry. Some of the tasks that are meant to approach this goal were completed and some of them were not. Despite the fact that the activity of TnpB RNP D was successfully estimated, turns out that it is too low to smoothly move to the NucleaSeq. Some of the things need to be redone to avoid losing so much DNA. First of all, one of the future plans is improving the quality of the RNP. As already mentioned, one of the strategies that could help to improve RNP's activity could be to produce it by IVTT to avoid cellular RNases and ensure reRNA's homogeneity. If that would result in the increased activity of TnpB RNP D, then higher inputs of DNAs would become possible in order to obtain cleavage products and therefore such inputs would be easier to precipitate after the cleavage reaction and prepare for sequencing without losing a lot of significant data.

Speaking of other tasks outlined in this study, the determination of reaction conditions for the NucleaSeq cleavage was successful. Obtained data proves that reaction temperature set to 37 °C allows capturing of the cleavage products and their differing quantities at different concentrations of TnpB RNP D or varying time points. It is also expected that DNA library D design was successful since it was designed by the instructions of the previous study (Stephen K. Jones et al., 2021). Nevertheless, results of PCR show that there is an unspecific product appearing after amplification and without sequencing data it is impossible to tell if it affects the NucleaSeq experiment.

In case the activity of TnpB RNP D becomes improved and higher inputs of DNA gets to be successfully prepared for NGS, the idea of benchmarking TnpB's specificity finally becomes possible. It would be very interesting to see how it compares to well-studied Cas nucleases like Cas9 and Cas12a. If TnpB turns out to exhibit high specificity (high cleavage rates over on-targets and low cleavage rates over off-targets) it could become one of the gene

editing tools applied for various biochemical, biomedical and agricultural approaches. In such case it is also important to know where exactly TnpB cuts within a target sequence and what DNA ends (sticky or blunt) does it generate – this data should also be obtained from NGS results. In addition, if TnpB turns out to be highly specific, it should be tested *in vivo* to determine which repair mechanism (NHEJ or HDR) is more common after the cleavage.

Overall, this study supports the idea that TnpB could be one of the potential gene editors, but some improvements regarding experiments need to be achieved in order to benchmark important TnpB's features and understand how well it compares to other gene editors.

## Conclusions

1. The low activity of the studied TnpB RNP D might be related to improper length of reRNA (<270 nt).
2. Reaction temperature set to 37 ℃ allows capturing of the cleavage products and their differing quantities at different concentrations of TnpB RNP D or varying time points and makes TnpB perform faster than reaction temperature set to 22 ℃.
3. Due to low activity of TnpB RNP D it had to be used in a 100-fold excess making DNA input low and hard to recover.
4. The activity of TnpB RNP D has to be improved in future experiments.
5. The specificity of TnpB and its cleavage rates across on- and off-targets still needs to be determined from data that will be obtained from NGS.
6. Results obtained from NGS could help to decide how does TnpB compare to class leading gene editing tools and if it could be one of them.

**Summary**

Gene editing gained significant importance due to its potential to revolutionize fields of biomedicine, biochemistry, agriculture and address many problems. By harnessing gene editing technologies, scientists can quickly and easily alter an organism's DNA. Despite its indisputable importance, gene editing comes with several issues including delivery of the gene editor to the cell, safety and applicability. To address these problems, scientists work on improving current gene editors or looking for new ones. This study focuses on TnpB – a transposon associated nuclease that is two times smaller than class leading gene editor Cas9 and therefore easier to deliver to the target cells. Though TnpB has been known for a while, its potential as a gene editor was only recently unravelled. That means that some of its features including specificity is unknown.

Here, the main research goal to determine TnpB's specificity profile with next-generation biochemistry is outlined and divided into smaller steps. Overall, the main method used in this study is NucleaSeq - a platform that combines DNA digestion by a nuclease with deep sequencing. NucleaSeq reveals the cleavage kinetics of a nuclease and benchmarks its specificity by quantifying cleavage rates across DNA target sequences including thousands of off-targets that do not fully match the nuclease's RNA guide. It also reveals cleavage sites within target sequences and what kind of ends (sticky or blunt) does nuclease generate after the cleavage. In conclusion, by adopting NucleaSeq strategy the main features of TnpB and how does it compare to other well-studied gene editing tools should be determined. Obtained data could help to establish new gene editing strategies based on TnpB.

**Santrauka**

Genų redagavimo reikšmė nenuginčijama - jis taikomas biomedicinos, biochemijos ir agrikultūros srityse ir padeda išspręsti daugybę problemų. Naudodami genų redagavimo technologijas, mokslininkai gali greitai ir lengvai pakeisti organizmo DNR. Nepaisant didelės svarbos, genų redagavimas turi keletą trūkumų, įskaitant genų redagavimo įrankio pateikimą į ląstelę, saugumą ir pritaikomumą. Siekdami išspręsti šias problemas, mokslininkai tobulina esamus genų redagavimo įrankius arba ieško naujų. Šiame tyrime daugiausia dėmesio skiriama TnpB – su transpozonais susijusiai nukleazei, kuri yra du kartus mažesnė nei žymus genų redagavimo įrankis Cas9, todėl ją lengviau pateikti į ląsteles. Nors apie TnpB egzistavimą buvo žinoma jau kurį laiką, jos, kaip genų redagavimo įrankio, potencialas buvo atskleistas tik neseniai. Dėl šios priežasties kai kurios TnpB savybės dar neatskleistos, pavyzdžiui, nėra duomenų apie jos specifiškumą.

Šiame darbe keliamas pagrindinis tikslas yra nustatyti TnpB specifiškumo profilį naudojant naujos kartos biochemijos metodus. Pagrindinis šiame tyrime naudojamas metodas yra NucleaSeq – tai platforma, kuri apjungia DNR karpymą nukleaze su naujos kartos sekoskaita. NucleaSeq atskleidžia nukleazės karpymo kinetiką ir specifiškumą, kiekybiškai įvertinant karpymo greitį DNR taikinių sekose, įskaitant tūkstančius pašalinių taikinių, kurie nevisiškai atitinka nukleazės gidinę RNR. Metodas taip pat atskleidžia kuriose taikinių vietose nukleazė sukuria trūkį ir kokius galus (lipnius ar bukus) nukleazė palieka po karpymo. NucleaSeq metodas gali padėti nustatyti pagrindines TnpB savybes ir palyginti ją su kitais gerai ištirtais genų redagavimo įrankiais. Gauti duomenys gali padėti sukurti naujas genų redagavimo strategijas, pagrįstas TnpB.

**Acknowledgements**

First of all, I would like to thank my mentor and supervisor Prof. Dr. Stephen Knox Jones for a great leadership in science and the ideas for my Master thesis project and all the Jones!Lab students and employes for their companionship in laboratory. I would like to thank my colleagues Dr. Tautvydas Karvelis and PhD Gytis Druteika for their shared knowledge and materials that made this project come to reality. I would also like to thank Dr. Mindaugas Zaremba and his laboratory members for sharing their inventory (Bioanalyzer). Last but not least I would like to thank Prof. Dr. Virginijus Šikšnys for his warm welcoming of our Jones!Lab laboratory team and shared laboratory space – without that the science we do would not be possible.

## Supplementary data

### 1. DNA sequences

Linear DNA target D (PAM is marked in green, target sequence – in orange):
AACCGCCGAATAACAGAGTATCCGCGCTGGTTGTATATCAGATTGATGTGATAA
GTGGAATGCCATGTGGTCGAGTCGATAGTACTCTGCACATCTCGTAAGGTGGTTC
CGGAGTGTGCGAGGCGTTCTT

DNA library D:

(Steven Knox Jones, 2023c)

### 2. Cleavage rates

**a)**

**b)**

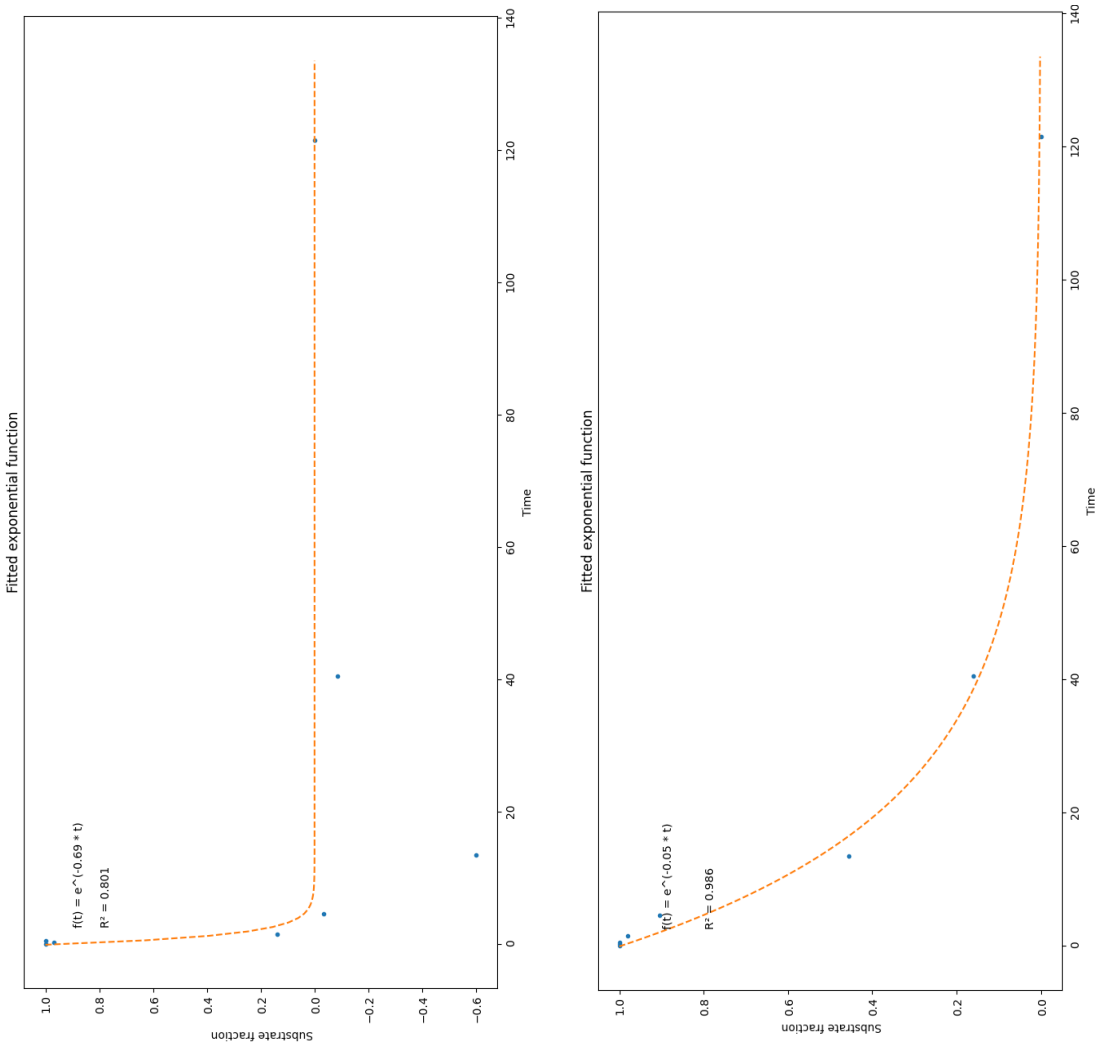



**Figure 2S. Cleavage rates obtained during the reactions set to different temperatures (*Subsection 3.3*).** Cleavage of targets by TnpB follows an exponential decay function (f(t) = e^(-k* t) where k is the cleavage rate, and t is time in seconds. X axis – time in minutes, Y axis – substrate fraction $R^2$.

a) Cleavage rate obtained during the reaction set to 37 °C.
b) Cleavage rate obtained during the reaction set to 22 °C.

**Literature sources**


1. Abdallah, N. A., Prakash, C. S., & McHughen, A. G. (2015). Genome editing for crop improvement: Challenges and opportunities. *GM Crops & Food*, *6*(4), 183–205. https://doi.org/10.1080/21645698.2015.1129937
2. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H., & Gorodkin, J. (2018). CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology*, *19*(1), 1–13. https://doi.org/10.1186/s13059-018-1534-x
3. Barrangou, R. (2013). CRISPR-Cas systems and RNA-guided interference. *Wiley Interdisciplinary Reviews: RNA*, *4*(3), 267–278. https://doi.org/10.1002/wrna.1159
4. Bigelyte, G., Young, J. K., Karvelis, T., Budre, K., Zedaveinyte, R., Djukanovic, V., Van Ginkel, E., Paulraj, S., Gasior, S., Jones, S., Feigenbutz, L., Clair, G. S., Barone, P., Bohn, J., Acharya, A., Zastrow-Hayes, G., Henkel-Heinecke, S., Silanskas, A., Seidel, R., & Siksnys, V. (2021). Miniature type V-F CRISPR-Cas nucleases enable targeted DNA modification in cells. *Nature Communications*, *12*(1), 1–8. https://doi.org/10.1038/s41467-021-26469-4
5. Bulcha, J. T., Wang, Y., Ma, H., Tai, P. W. L., & Gao, G. (2021). Viral vector platforms within the gene therapy landscape. *Signal Transduction and Targeted Therapy*, *6*(1). https://doi.org/10.1038/s41392-021-00487-6
6. Caffery, B., Lee, J. S., & Alexander-Bryant, A. A. (2019). Vectors for glioblastoma gene therapy: Viral & non-viral delivery strategies. *Nanomaterials*, *9*(1). https://doi.org/10.3390/nano9010105
7. Chandrasegaran, S., & Carroll, D. (2016). Origins of Programmable Nucleases for Genome Engineering. *Journal of Molecular Biology*, *428*(5), 963–989. https://doi.org/10.1016/j.jmb.2015.10.014
8. Cohen, S. N., Chang, A. C. Y., Boyer, H. W., & Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(11), 3240–3244. https://doi.org/10.1073/pnas.70.11.3240
9. Corsi, G. I., Qu, K., Alkan, F., Pan, X., Luo, Y., & Gorodkin, J. (2022). CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context. *Nature Communications*, *13*(1), 1–14. https://doi.org/10.1038/s41467-022-30515-0
10. Danna, K., & Nathans, D. (1971). Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(12), 2913–2917. https://doi.org/10.1073/pnas.68.12.2913
11. Di Felice, F., Micheli, G., & Camilloni, G. (2019). Restriction enzymes and their use in molecular biology: An overview. *Journal of Biosciences*, *44*(2). https://doi.org/10.1007/s12038-019-9856-8
12. Dolgin, E. (2022). T-cell vaccines could top up immunity to COVID, as variants loom large. *Nature Biotechnology*, *40*(1), 3–4. https://doi.org/10.1038/d41587-021-00025-3
13. Dong, B., Nakai, H., & Xiao, W. (2010). Characterization of genome integrity for oversized recombinant AAV vector. *Molecular Therapy*, *18*(1), 87–92. https://doi.org/10.1038/mt.2009.258
14. Dooley, S. K., Baken, E. K., Moss, W. N., Howe, A., & Young, J. K. (2021). Identification and Evolution of Cas9 tracrRNAs. *CRISPR Journal*, *4*(3), 438–447. https://doi.org/10.1089/crispr.2020.0093
15. Du, Z., Munye, M. M., Tagalakis, A. D., Manunta, M. D. I., & Hart, S. L. (2014). *The Role of the Helper Lipid on the DNA Transfection Efficiency of Lipopolyplex Formulations*. 4–9. https://doi.org/10.1038/srep07107

16. Ganbaatar, U., & Liu, C. (2021). CRISPR-Based COVID-19 Testing: Toward Next-Generation Point-of-Care Diagnostics. *Frontiers in Cellular and Infection Microbiology*, *11*(April), 1–11. https://doi.org/10.3389/fcimb.2021.663949
17. Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(39), 2579–2586. https://doi.org/10.1073/pnas.1208507109
18. Gil-Humanes, J., Wang, Y., Liang, Z., Shan, Q., Ozuna, C. V., Sánchez-León, S., Baltes, N. J., Starker, C., Barro, F., Gao, C., & Voytas, D. F. (2017). High-efficiency gene targeting in hexaploid wheat using DNA replicons and CRISPR/Cas9. *Plant Journal*, *89*(6), 1251–1262. https://doi.org/10.1111/tpj.13446
19. Goldner, M. G. (1972). History of insulin. *Annals of Internal Medicine*, *76*(2), 329–330. https://doi.org/10.7326/0003-4819-76-2-329
20. Gu, W., Wang, T., Maltais, F., Ledford, B., Kennedy, J., Wei, Y., Gross, C. H., Parsons, J., Duncan, L., Ryan Arends, S. J., Moody, C., Perola, E., Green, J., & Charifson, P. S. (2012). Design, synthesis and biological evaluation of potent NAD +-dependent DNA ligase inhibitors as potential antibacterial agents. Part I: Aminoalkoxypyrimidine carboxamides. *Bioorganic and Medicinal Chemistry Letters*, *22*(11), 3693–3698. https://doi.org/10.1016/j.bmcl.2012.04.037
21. Guo, C., Ma, X., Gao, F., & Guo, Y. (2023). Off-target effects in CRISPR/Cas9 gene editing. *Frontiers in Bioengineering and Biotechnology*, *11*(March), 1–11. https://doi.org/10.3389/fbioe.2023.1143157
22. Gupta, R. M., & Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. *Journal of Clinical Investigation*, *124*(10), 4154–4161. https://doi.org/10.1172/JCI72992
23. Handelmann, C. R., Tsompana, M., Samudrala, R., & Buck, M. J. (2023). The impact of nucleosome structure on CRISPR/Cas9 fidelity. *Nucleic Acids Research*, *51*(5), 2333–2344. https://doi.org/10.1093/nar/gkad021
24. He, Y. (2009). High cell density production of Deinococcus radiodurans under optimized conditions. *Journal of Industrial Microbiology and Biotechnology*, *36*(4), 539–546. https://doi.org/10.1007/s10295-008-0524-5
25. Hickman, A. B., James, J. A., Barabas, O., Pasternak, C., Ton-Hoang, B., Chandler, M., Sommer, S., & Dyda, F. (2010). DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. *EMBO Journal*, *29*(22), 3840–3852. https://doi.org/10.1038/emboj.2010.241
26. Hirano, S., Abudayyeh, O. O., Gootenberg, J. S., Horii, T., Ishitani, R., Hatada, I., Zhang, F., Nishimasu, H., & Nureki, O. (2019). Structural basis for the promiscuous PAM recognition by Corynebacterium diphtheriae Cas9. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-09741-6
27. Höijer, I., Emmanouilidou, A., Östlund, R., van Schendel, R., Bozorgpana, S., Tijsterman, M., Feuk, L., Gyllensten, U., den Hoed, M., & Ameur, A. (2022). CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nature Communications*, *13*(1), 1–10. https://doi.org/10.1038/s41467-022-28244-5
28. Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the immune system of Bacteria and Archaea. *Science*, *327*(5962), 167–170. https://doi.org/10.1126/science.1179555
29. Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious. *Journal of Bacteriology*, *200*(7), e00580-17.
30. Istvan, L. (n.d.). *GelAnalyzer 19.1*. http://www.gelanalyzer.com/?i=1#section-citation
31. Iwasaki, H., Takahagi, M., Shiba, T., Nakata, A., & Shinagawa, H. (1991). Escherichia

coli RuvC protein is an endonuclease that resolves the Holliday structure. *EMBO Journal*, *10*(13), 4381–4389. https://doi.org/10.1002/j.1460-2075.1991.tb05016.x
32. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). *A Programmable Dual-RNA – Guided*. *337*(August), 816–822.
33. Jones, Stephen K., Hawkins, J. A., Johnson, N. V., Jung, C., Hu, K., Rybarski, J. R., Chen, J. S., Doudna, J. A., Press, W. H., & Finkelstein, I. J. (2021). Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology*, *39*(1), 84–93. https://doi.org/10.1038/s41587-020-0646-5
34. Jones, Steven Knox. (2023a). *Decay program*. NucleaSeq/notebooks at master · skjonesjr/NucleaSeq · GitHub
35. Jones, Steven Knox. (2023b). *Locus of TnpB, Unpublished*.
36. Jones, Steven Knox. (2023c). *Target libraries for TnpB*. https://github.com/skjonesjr/NucleaSeq/blob/master/resources/exploded_TnpB_Buffered_seqs_DandE.txt
37. Joung, J. K., & Sander, J. D. (2013). *Nat Rev Mol Cell Biol. 14*. *14*(1), 49–55. https://doi.org/10.1038/nrm3486.TALENs
38. Jurica, M. S., Monnat, R. J., & Stoddard, B. L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-Crel. *Molecular Cell*, *2*(4), 469–476. https://doi.org/10.1016/S1097-2765(00)80146-X
39. Kantor, A., McClements, M. E., & Maclaren, R. E. (2020). Crispr-cas9 dna base-editing and prime-editing. *International Journal of Molecular Sciences*, *21*(17), 1–22. https://doi.org/10.3390/ijms21176240
40. Karvelis, Tautvydas; Druteika, G. (2021). *Target D Plasmid*. Benchling.Com. https://benchling.com/s/seq-L844BPZTW50nirzFJpjn/edit
41. Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, Č., & Siksnys, V. (2021). Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature*, *599*(November). https://doi.org/10.1038/s41586-021-04058-1
42. Kellner, M. J., Koob, J. G., Gootenberg, J. S., Abudayyeh, O. O., & Zhang, F. (2019a). SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature Protocols*, *14*(10), 2986–3012. https://doi.org/10.1038/s41596-019-0210-2
43. Kellner, M. J., Koob, J. G., Gootenberg, J. S., Abudayyeh, O. O., & Zhang, F. (2019b). SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature Protocols*, *14*(10), 2986–3012. https://doi.org/10.1038/s41596-019-0210-2
44. Kim, T. H., & Lee, S. W. (2022). Therapeutic Application of Genome Editing Technologies in Viral Diseases. *International Journal of Molecular Sciences*, *23*(10). https://doi.org/10.3390/ijms23105399
45. Klug, A. (2010). The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Quarterly Reviews of Biophysics*, *43*(1), 1–21. https://doi.org/10.1017/S0033583510000089
46. Kolesnik, M. V., Fedorova, I., Karneyeva, K. A., Artamonova, D. N., & Severinov, K. V. (2021). Type III CRISPR-Cas Systems: Deciphering the Most Complex Prokaryotic Immune System. *Biochemistry (Moscow)*, *86*(10), 1301–1314. https://doi.org/10.1134/S0006297921100114
47. Koonin, E. V., & Makarova, K. S. (2017). Mobile genetic elements and evolution of crispr-cas systems: All theway there and back. *Genome Biology and Evolution*, *9*(10), 2812–2825. https://doi.org/10.1093/gbe/evx192
48. Kruger, D. H., & Bickle, T. A. (1983). Bacteriophage survival: Multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiological Reviews*, *47*(3), 345–360. https://doi.org/10.1128/mmbr.47.3.345-

360.1983
49. Krupovic, M., Béguin, P., & Koonin, E. V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Current Opinion in Microbiology*, *38*, 36–43. https://doi.org/10.1016/j.mib.2017.04.004
50. Lee, J., Nonaka, S., Takayama, M., & Ezura, H. (2018). Utilization of a Genome-Edited Tomato (Solanum lycopersicum) with High Gamma Aminobutyric Acid Content in Hybrid Breeding. *Journal of Agricultural and Food Chemistry*, *66*(4), 963–971. https://doi.org/10.1021/acs.jafc.7b05171
51. Li, J., Scarano, A., Gonzalez, N. M., D'Orso, F., Yue, Y., Nemeth, K., Saalbach, G., Hill, L., de Oliveira Martins, C., Moran, R., Santino, A., & Martin, C. (2022). Biofortified tomatoes provide a new route to vitamin D sufficiency. *Nature Plants*, *8*(6), 611–616. https://doi.org/10.1038/s41477-022-01154-6
52. Li, M., & Wang, D. (2017). Gene editing nuclease and its application in tilapia. *Science Bulletin*, *62*(3), 165–173. https://doi.org/10.1016/j.scib.2017.01.003
53. Liu, M., Rehman, S., Tang, X., Gu, K., Fan, Q., Chen, D., & Ma, W. (2019). Methodologies for improving HDR efficiency. *Frontiers in Genetics*, *10*(JAN), 1–9. https://doi.org/10.3389/fgene.2018.00691
54. Liu, T. Y., & Doudna, J. A. (2020). Chemistry of Class 1 CRISPR-Cas effectors: Binding, editing, and regulation. *Journal of Biological Chemistry*, *295*(42), 14473–14487. https://doi.org/10.1074/jbc.REV120.007034
55. Liu, Y., Zhao, M., Gong, M., Xu, Y., Xie, C., Deng, H., Li, X., Wu, H., & Wang, Z. (2018). Inhibition of hepatitis B virus replication via HBV DNA cleavage by Cas9 from Staphylococcus aureus. *Antiviral Research*, *152*(February), 58–67. https://doi.org/10.1016/j.antiviral.2018.02.011
56. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G., & Murrayy, N. E. (2014). Highlights of the DNA cutters: A short history of the restriction enzymes. *Nucleic Acids Research*, *42*(1), 3–19. https://doi.org/10.1093/nar/gkt990
57. Marcaida, M. J., Muñoz, I. G., Blanco, F. J., Prieto, J., & Montoya, G. (2010). Homing endonucleases: From basics to therapeutic applications. *Cellular and Molecular Life Sciences*, *67*(5), 727–748. https://doi.org/10.1007/s00018-009-0188-y
58. McHenry, C., & Kornberg, A. (1977). DNA polymerase III holoenzyme of Escherichia coli. Purification and resolution into subunits. *Journal of Biological Chemistry*, *252*(18), 6478–6484. https://doi.org/10.1016/s0021-9258(17)39983-0
59. Mojica, F. J. M., & Rodriguez-Valera, F. (2016). The discovery of CRISPR in archaea and bacteria. *FEBS Journal*, *283*, 3162–3169. https://doi.org/10.1111/febs.13766
60. Mosterd, C., & Moineau, S. (2020). *crossm*. *5*(3), 1–13.
61. Mullis, K. B. (1990). The unusual origin of the polymerase chain reaction. *Scientific American*, *262*(4), 56–65. https://doi.org/10.1038/scientificamerican0490-56
62. Mussolino, C., Alzubi, J., Fine, E. J., Morbitzer, R., Cradick, T. J., Lahaye, T., Bao, G., & Cathomen, T. (2014). TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. *Nucleic Acids Research*, *42*(10), 6762–6773. https://doi.org/10.1093/nar/gku305
63. Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., & Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, *156*(5), 935–949. https://doi.org/10.1016/j.cell.2014.02.001
64. Okita, K., Kiyonari, H., Nobuhisa, I., Kimura, N., Aizawa, S., & Taga, T. (2004). Targeted disruption of the mouse ELYS gene results in embryonic death at peri-implantation development. *Genes to Cells*, *9*(11), 1083–1091. https://doi.org/10.1111/j.1365-2443.2004.00791.x

65. Palermo, G., Chen, J. S., Ricci, C. G., Rivalta, I., Jinek, M., Batista, V. S., Doudna, J. A., & McCammon, J. A. (2018). Key role of the REC lobe during CRISPR-Cas9 activation by "sensing", "regulating", and "locking" the catalytic HNH domain. *Quarterly Reviews of Biophysics*, *51*. https://doi.org/10.1017/S0033583518000070
66. Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science*, *60*(1), 103060. https://doi.org/10.1016/j.transci.2021.103060
67. Paul, B., & Montoya, G. (2020). CRISPR-Cas12a: Functional overview and applications. *Biomedical Journal*, *43*(1), 8–17. https://doi.org/10.1016/j.bj.2019.10.005
68. Perler, F. B. (1998). Protein splicing of inteins and hedgehog autoproteolysis: Structure, function, and evolution. *Cell*, *92*(1), 1–4. https://doi.org/10.1016/S0092-8674(00)80892-2
69. Pingoud, A., & Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. *Nucleic Acids Research*, *29*(18), 3705–3727. https://doi.org/10.1093/nar/29.18.3705
70. Portin, P. (2014). The birth and development of the DNA theory of inheritance: Sixty years since the discovery of the structure of DNA. *Journal of Genetics*, *93*(1), 293–302. https://doi.org/10.1007/s12041-014-0337-4
71. Rédei, G. P. (2008). Homing Endonucleases. *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, 895–895. https://doi.org/10.1007/978-1-4020-6754-9_7753
72. Roshanravan, N., Tutunchi, H., Najafipour, F., Dastouri, M., Ghaffari, S., & Jebeli, A. (2022). A glance at the application of CRISPR/Cas9 gene-editing technology in cardiovascular diseases. *Journal of Cardiovascular and Thoracic Research*, *14*(2), 77–83. https://doi.org/10.34172/jcvtr.2022.14
73. Sasnauskas, G., Tamulaitiene, G., Druteika, G., Carabias, A., Silanskas, A., Kazlauskas, D., Venclovas, Č., Montoya, G., Karvelis, T., & Siksnys, V. (2023). TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature*, *616*(October 2022). https://doi.org/10.1038/s41586-023-05826-x
74. Science, T. (1999). *Cause of Irreversible Inactivation of Aminoacylase in Urea Solution*. *4*(3).
75. Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O. O., Gootenberg, J. S., Makarova, K. S., Wolf, Y. I., Severinov, K., Zhang, F., & Koonin, E. V. (2017). Diversity and evolution of class 2 CRISPR-Cas systems. *Nature Reviews Microbiology*, *15*(3), 169–182. https://doi.org/10.1038/nrmicro.2016.184
76. Silas, S., Makarova, K. S., Shmakov, S., Páez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S., Krishnamurthy, S. R., Fu, B. X. H., Hansen, L. L., Wang, D., Sullivan, M. B., Millard, A., Clokie, M. R., Bhaya, D., Lambowitz, A. M., Kyrpides, N. C., Koonin, E. V., & Fire, A. Z. (2017). On the origin of reverse transcriptase- using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *MBio*, *8*(4). https://doi.org/10.1128/mBio.00897-17
77. Silva, G. H., & Belfort, M. (2004). Analysis of the LAGLIDADG interface of the monomeric homing endonuclease I-DmoI. *Nucleic Acids Research*, *32*(10), 3156–3168. https://doi.org/10.1093/nar/gkh618
78. Slade, D., & Radman, M. (2011). Oxidative Stress Resistance in Deinococcus radiodurans . In *Microbiology and Molecular Biology Reviews* (Vol. 75, Issue 1). https://doi.org/10.1128/mmbr.00015-10
79. Swarts, D. C., van der Oost, J., & Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell*, *66*(2), 221-233.e4. https://doi.org/10.1016/j.molcel.2017.03.016

80. Taylor, H. N., Laderman, E., Armbrust, M., Hallmark, T., Keiser, D., Bondy-Denomy, J., & Jackson, R. N. (2021). Positioning Diverse Type IV Structures and Functions Within Class 1 CRISPR-Cas Systems. *Frontiers in Microbiology*, *12*(May), 1–10. https://doi.org/10.3389/fmicb.2021.671522
81. Tong, B., Dong, H., Cui, Y., Jiang, P., Jin, Z., & Zhang, D. (2021). The Versatile Type V CRISPR Effectors and Their Application Prospects. *Frontiers in Cell and Developmental Biology*, *8*(February), 1–11. https://doi.org/10.3389/fcell.2020.622103
82. Turchiano, G., Andrieux, G., Klermund, J., Blattner, G., Pennucci, V., el Gaz, M., Monaco, G., Poddar, S., Mussolino, C., Cornu, T. I., Boerries, M., & Cathomen, T. (2021). Quantitative evaluation of chromosomal rearrangements in gene-edited human stem cells by CAST-Seq. *Cell Stem Cell*, *28*(6), 1136-1147.e5. https://doi.org/10.1016/j.stem.2021.02.002
83. Weimberg, R. (1967). Effect of sodium chloride on the activity of a soluble malate dehydrogenase from pea seeds. *Journal of Biological Chemistry*, *242*(12), 3000–3006. https://doi.org/10.1016/s0021-9258(18)99604-3
84. Wright, A. V., Wang, J. Y., Burstein, D., Harrington, L. B., Paez-Espino, D., Kyrpides, N. C., Iavarone, A. T., Banfield, J. F., & Doudna, J. A. (2019). A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Molecular Cell*, *73*(4), 727-737.e3. https://doi.org/10.1016/j.molcel.2018.12.015
85. Wu, X., Kriz, A. J., & Sharp, P. A. (2014). Target specificity of the CRISPR-Cas9 system. *Quantitative Biology*, *2*(2), 59–70. https://doi.org/10.1007/s40484-014-0030-x
86. Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. *Computational and Structural Biotechnology Journal*, *18*, 2401–2415. https://doi.org/10.1016/j.csbj.2020.08.031
87. Yamatsu, A., Yamashita, Y., Pandharipande, T., Maru, I., & Kim, M. (2016). Effect of oral γ-aminobutyric acid (GABA) administration on sleep and its absorption in humans. *Food Science and Biotechnology*, *25*(2), 547–551. https://doi.org/10.1007/s10068-016-0076-9
88. Yang, Y. C., & Yang, H. C. (2022). Recent progress and future prospective in HBV cure by CRISPR/cas. *Viruses*, *14*(1), 1–15. https://doi.org/10.3390/v14010004
89. Zahra, A., Shahid, A., Shamim, A., Khan, S. H., & Arshad, M. I. (2022). The SHERLOCK Platform: An Insight into Advances in Viral Disease Diagnosis. *Molecular Biotechnology*, *65*(5), 699–714. https://doi.org/10.1007/s12033-022-00625-7
90. Zheng, Y., Li, J., Wang, B., Han, J., Hao, Y., Wang, S., Ma, X., Yang, S., Ma, L., Yi, L., & Peng, W. (2020). Endogenous Type I CRISPR-Cas: From Foreign DNA Defense to Prokaryotic Engineering. *Frontiers in Bioengineering and Biotechnology*, *8*(March), 1–17. https://doi.org/10.3389/fbioe.2020.00062
91. Zmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics*, *127*(1), 1–18. https://doi.org/10.1007/s00122-013-2177-7