ETH zürich

Master's Thesis

Towards a novel approach to multimodal epigenomic profiling

Maria-Andreea Dimitriu

Master of Science in Biology/Neuroscience

ETH Zürich

September 2019 – March 2020

Thesis supervisors:
Prof. Dr. Isabelle Mansuy
Martin Roszkowski, MSc.

Thesis co-supervisor:
Prof. Dr. Johannes Bohacek

Table of Contents

Acknowledgements

This entire project has truly been a team effort and I would like to thank, by this means, the people who relentlessly supported me throughout my work in the past six months.

First, I would like to thank Isabelle Mansuy, who offered me the great opportunity to conduct research in her group, provided us with all the resources needed and allowed us to significantly alter the focus of our work, which made this entire project possible. I would also like to thank Johannes Bohacek very much for agreeing to be the co-supervisor of my thesis, for his valuable scientific input and for his genuine interest in our project.

I would then like to give my deepest thanks to Martin Roszkowski, who has been, by far and from every possible angle, the best supervisor I ever had. As a great scientist with a special inclination for methodological advances, this entire project was his idea and he continued, throughout our work, to come up with small but brilliant ideas that improved and gave our project purpose. Martin provided me with guidance at every step but, at the same time, trusted me and gave me liberty to explore ideas and strategies on my own, and to conduct the entirety of experiments myself, all of which, I think, have helped me grow as a scientist in the past half a year.

Then, I would like to thank Irina Lazar for her support throughout the project, for teaching me how to do ATAC-seq and for being the ATAC-seq expert in our office, who we could always annoy with questions. I would also like to thank Kristina for showing me the cell culture techniques used in the lab.

I would like to thank Silvia Schelbert and my friend, Margaux Quiniou, for teaching me the strategies and hurdles of molecular cloning, a skill I think I will continue to use for the rest of my professional life.

I would like to give a special thanks to Pierre-Luc Germain for his valuable input about experimental design and his expertise in bioinformatic analysis and, in general, for his keen interest and faith in our project.

Very importantly, I would like to thank Rodrigo Villaseñor for giving the presentation that sparked the entire project, for brainstorming with us on conceptual ideas in its infancy and for culturing and providing us with cells for our experiments. I would also like to thank Tuncay Baubec for agreeing to collaborate with us.

I would also like to thank Steven Henikoff for his valuable input on our use of the single-cell CUT&Tag protocol, on which our strategy is based.

Last but not least, I would like to thank the entire Mansuy group for welcoming me in their midst, for teaching me about science and life, and for becoming friends.

Abstract

While the DNA sequence of every cell in an organism is highly similar, cells display vastly different phenotypes and function due to regulation of gene expression by epigenetic mechanisms. This occurs through dynamic chromatin alterations, which allow organisms to develop complex tissues and provide a mechanism for the genome to integrate environmental cues. In this way, rapid adaptation can occur in response to changing conditions. Epigenetic factors and mechanisms, such as DNA methylation, chromatin accessibility and histone modifications interact to form a functional epigenetic network. Current methods to investigate the epigenome typically only resolve one epigenetic mark and often rely on antibodies. For example, chromatin immunoprecipitation sequencing (ChIP-Seq) is a widely used method to localize histone marks within the genome, but the binding loci of only one specific target can be probed at a time, providing little information about functional interplay with other epigenetic features. In addition, such methods are limited by antibody specificity and require the integration of data from different types of experiments to infer interactions between distinct epigenetic factors.

We propose NanoTag, a novel approach for multimodal epigenomic profiling that does not rely on antibodies or chemical fixation. NanoTag is an *in situ* approach that can provide information about DNA methylation, chromatin accessibility and chromatin-associated targets, such as histone modifications or transcription factors, simultaneously in one experiment. NanoTag relies on a fusion between the Tn5 transposase - an enzyme that binds and cleaves DNA, widely used to generate DNA sequencing libraries - and an antibody mimetic that tethers Tn5 directly to target loci in the genome. In addition, since it does not involve cross-linking of chromatin, NanoTag provides a native view of the interaction between distinct factors in the epigenomic network.

As proof of principle, we produced a fusion between Tn5 and a GFP-binding nanobody (GBN-Tn5) and verified its functionality in transposing DNA. We used GBN-Tn5 to generate sequencing libraries in a stem cell model expressing various engineered chromatin readers. Preliminary results suggest that NanoTag is a promising strategy for targeted epigenomic profiling with many potential applications in epigenetics.

1. Introduction

1.1. The epigenome and epigenomics

1.1.1. General overview of epigenetic marks

While all cell types in the mammalian body share a unique DNA sequence, they all exert very different functions and are characterized by different molecular environments. This cellular identity is established as a result of differentiation events across development and is coupled to a unique epigenetic signature[1]. Epigenetic mechanisms modulate gene expression patterns without altering the DNA sequence, through dynamic chromatin modifications. Epigenetic factors and mechanisms include DNA methylation, histone variants, histone modifications, RNA modifications and noncoding RNAs (Figure 1).

Epigenetic changes provide a mechanism through which the environment can influence the organism via changes in gene expression[2]. Environmental factors that can modulate the epigenome include lifestyle, diet, physical activity, stress, social factors and toxin exposure. Such environmentally induced dynamic modifications can be mitotically and meiotically heritable, providing a means for phenotypic plasticity, both on an adaptive and on an evolutionary scale[3].

The epigenome regulates numerous cellular processes, ranging from cell-type identity and cell reprogramming to metabolism, physiology, and even complex behaviors, such as social plasticity[4], learning and memory formation[5]. During development, two stages of epigenetic reprogramming occur, which reset the epigenetic landscape, setting the ground for progressive lineage-dependent differentiation[6]. In homeostasis, epigenetic control is also involved in the innate immune response against infection[7]. Epigenetic dysregulation is associated with a broad range of pathologies, from cancer and autoimmune diseases to metabolic and neurodegenerative disorders[8,9], and alterations within the epigenomic network are also involved in aging and age-associated disorders[10].

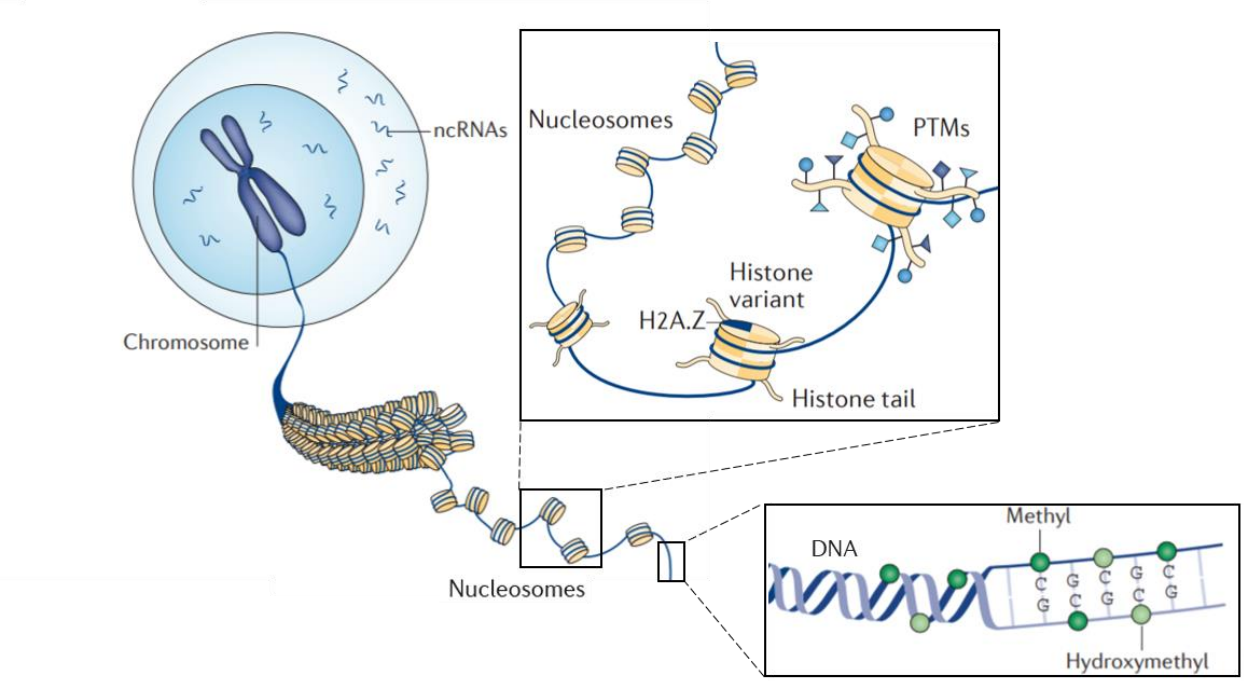


Figure 1 | **Overview of the major epigenetic marks in mammalian cells**. Histones present several isoforms (histone variants) and can carry post-translational modifications (PTMs), modulating chromatin dynamics and gene expression. DNA is wrapped around eight histones, forming nucleosomes. DNA methylation occurs at around 75% of CpG dinucleotides in the genome. Adapted from Bohacek & Mansuy (2015)[11].

1.1.2. Higher-order structural chromatin organization

The three-dimensional genomic architecture, including the long-range interactions between distant chromatin regions through chromatin looping, and the global structural organization of chromatin within the nucleus play important roles in gene regulation[12], RNA splicing, DNA replication and repair[13]. Chromatin loops, which range in size from one kilobase to hundreds of kilobases, facilitate structural and functional interaction between far upstream elements, promoters and the gene body, promoting the formation of “transcriptional hubs” in which the relevant transcriptional factors cluster[12]. The most well-known feature of large-scale genome organization is the spatial segregation of transcriptionally active euchromatin from inactive heterochromatin. This segregation is facilitated by the anchoring of heterochromatin to the nuclear lamina (through lamina-associated domains, LADs) and to the nucleolus (through nucleolus-associated domains, NADs)[14]. Within euchromatin and heterochromatic regions, chromatin is further organized into megabase-scale topologically-associated domains (TADs), in which the intra-domain

contact frequency is much higher than that of interactions with DNA sequences from outside of the domain[15]. Chromatin interactions are investigated experimentally using chromosome conformation capture (3C) technology, 3C-based methods (4C, 5C, Hi-C) and other techniques, such as ChIP-loop, ChIA-PET and DamID[16]. While 3C enables the study of interactions between candidate loci in the genome, 4C facilitates a genome-wide screen for loci of interaction with a locus of interest and 5C provides a parallel analysis of chromatin contacts within a given region[17,18]. The newest among the 3C-based techniques is Hi-C, which allows the comprehensive study of genome-wide chromatin interactions[19].

1.1.3. Nucleosome occupancy and chromatin accessibility

The ability of transcription factors to bind to a specific DNA sequence within chromatin is regulated by the presence or absence of nucleosomes at that specific locus[20]. DNA accessibility and nucleosome composition are dynamically regulated by ATP-dependent chromatin remodeling complexes[21]. Methods used to study chromatin accessibility include assays in which DNA is enzymatically digested (DNase-seq, MACC, ATAC-seq) or physically disrupted (FAIRE-seq, Sono-seq)[16,22]. ATAC-seq is an increasingly popular method for genome-wide profiling of chromatin accessibility, using the Tn5 transposase, which cleaves DNA within accessible chromatin regions and ligates sequencing adapters to the ends of the DNA fragments[23]. A newly improved ATAC-seq derivative, Omni-ATAC-seq provides signal-to-background ratios and facilitates the epigenomic investigation of cells from frozen tissues[24].

1.1.4. The “histone code”

Covalent histone posttranslational modifications (HPTMs) include, among others, methylation, acetylation, phosphorylation, ubiquitylation and sumoylation, and result from the combined activity of reader, writer and eraser enzymes. HPTMs are involved in the regulation of gene expression by influencing chromatin structure and transcription factor recruitment[25] – a concept termed the “histone code”[26]. Generally, acetylation and phosphorylation of histone residues are associated with transcriptional activation, while histone lysine methylation is found in both actively transcribed and inactive genes[27].

H3K4me3 is an epigenetic mark reversibly established by lysine methyltransferase enzymes (KMTs) and removed by demethylases (KDMs) that is associated with active promoter elements and found at active transcription start sites (TSS)[28]. Within the coding region of a gene, the level of this epigenetic mark is increased, and its distribution shifted toward the 3' end with increased transcriptional activity[29]. Because of the strong association between this epigenetic mark and transcriptional activity, H3K4me3 has long been considered to have an instructive role in gene transcription. However, more recent evidence suggests this mark may actually be a result of transcription[30]. In addition, H3K4me3 may influence pre-mRNA splicing via the spliceosome-associated protein CHD1, which binds H3K4me3 sites[30]. More studies are, thus, required to elucidate the roles of this epigenetic mark and its interactions with other epigenetic mechanisms.

Traditionally, the method of choice to profile genome-wide distribution of histone modifications and of DNA-binding proteins has been ChIP-seq, but this technique presents several significant limitations. For example, ChIP-seq is critically dependent on the quality of antibodies available for the desired target, requires large amounts of input DNA[31] and suffers from low signal, high background and epitope masking due to cross-linking[32]. An alternative to ChIP-seq is CUT&Tag[32]. To target the loci of interest, CUT&Tag relies on antibodies against the protein of interest, followed by antibody recognition by protein A, which is fused to the Tn5 transposase. Once the pA-Tn5 is tethered to antibody-bound sites, it acts by cleaving the DNA and inserting adapters. DNA fragments are then amplified by PCR and sequenced (Figure 3b). This method provides much lower background, requires significantly lower input (even single cells are sufficient) and facilitates the profiling of a wide range of targets for which commercially available antibodies exist.

1.1.5. DNA methylation

The major form of DNA modification that occurs in mammals and the most studied epigenetic mark is the methylation of cytosine at position C5 within CpG dinucleotides (m^5C)[33]. In human somatic cells, 70-80% of all CpG dinucleotides in the genome are methylated, but when CpG dinucleotides occur as clusters, forming CpG islands (present within around 60% of gene promoters), they are generally unmethylated[1]. DNA

methylation (DNAme) is established and maintained by the DNA methyltransferase enzymes (DNMT1, 3A and 3B), recognized by methyl-CpG binding proteins (which can recruit other chromatin-associated factors, leading to transcriptional repression), and actively removed by methylcytosine dioxygenase (TET1-3) enzymes[34]. DNA methylation is involved in numerous biological processes, including development[35], hematopoiesis[36], genomic imprinting[37] and X chromosome inactivation[38]. Methylation of promoter associated CpG islands is an epigenetic mark that correlates with transcriptional silencing[39]. Alterations in DNA methylation patterns can lead to disease; for example, hypermethylation causing silencing of tumor suppressor genes is associated with cancer[40]. DNA methylation has also been implicated in the pathogenesis of autoimmune diseases (such as systemic lupus erythematosus)[41] and neurological disorders (such as the neurodevelopmental Rett syndrome)[42].

Methods used to investigate genome-wide DNA methylation include: whole-genome bisulfite sequencing, which relies on bisulfite treatment to convert unmethylated Cytosine to Thymidine[43], MeDIP-seq, which relies on immunoprecipitation of methylated DNA[44], MethylCap-seq, which uses the MBD of MeCP2 to capture methylated DNA[45]. In addition, NOMe-seq and methyl ATAC-seq enable the simultaneous assessment of nucleosome occupancy and DNA methylation. NOMe-seq relies on a methyltransferase that acts on GpC sites unprotected by nucleosomes[46]. Methyl-ATAC-seq uses the Tn5 transposase used in ATAC-seq, loaded with methylated oligonucleotides and includes bisulfite treatment of tagmented DNA[47].

1.1.6. Other epigenetic mechanisms

Other epigenetic factors and mechanisms include non-coding RNAs such as: miRNAs - which are important for RNA silencing[48], lncRNA - whose function is important for developmental processes including genomic imprinting and X-linked gene dosage compensation[49], and piRNAs - which are involved in transcriptional silencing of transposable elements in germline and gonadal somatic cells[50]. Finally, numerous RNA modifications play important roles in various cellular processes by modulating RNA processing, stability and decay[51].

1.1.7. Technical challenges in the study of epigenomic interactions

Despite their success in providing deep insight into epigenetic mechanisms, current methods used to study the epigenome do not allow for profiling of multiple epigenetic features simultaneously to understand their interaction in the cell. Currently, to obtain a more integrative view of the epigenomic mechanisms active in a cell, integration of data sets obtained in different experiments using different assays must be employed[16] – an approach that is computationally cumbersome and may fail in capturing the dynamic interactions native to cellular mechanisms.

Yet, the numerous DNA and chromatin modifications that dynamically occur in the cell act in a cooperative manner – creating an interactive epigenomic network that acts as a unitary molecular machine to modulate gene expression in the organism. Such crosstalk and interdependence between epigenetic mechanisms is functionally relevant for cellular processes including transcription regulation and DNA repair[52]. One example is the interaction between HPTMs and DNA methylation. H3K36me3 was found to promote *de novo* DNA methylation in mammals[53]. A further example of interactions between epigenetic mechanisms is the dependence of Dnmt3b-dependent DNA methylation on H3K9me3, which requires the ICBP90 chromatin-bound protein – a suspected oncogene that is overexpressed in breast-cancer cells[52].

Thus, a novel approach for the simultaneous investigation of multiple epigenetic features within the same cellular environment would greatly improve epigenomic studies and further insight into the role of the epigenome in biological processes and disease. To provide distinct advantages over existing epigenomics and multi-omics approaches, such a method would have to improve data quality, be easily applied to different targets, be adaptable to low input and even single cells, readily amenable to large-scale implementation.

1.2. NanoTag for targeted, multimodal epigenomic profiling

1.2.1. NanoTag – purpose and mode of action

We propose NanoTag, a novel experimental strategy for multimodal epigenomic profiling, which uses a bimodal enzyme – a fusion between the Tn5 transposase and an antibody mimetic – to facilitate targeted epigenomic investigation. NanoTag is designed for the study of interactions between epigenetic factors, enabling the assessment of chromatin accessibility, DNA methylation and multiple histone marks or transcription factors simultaneously – all without the use of antibodies, or chemical fixation techniques (i.e. crosslinking). In contrast to ChIP-seq, NanoTag is a low-input technique, which we expect will be fully scalable down to the level of single cells, based on previous successful use of Tn5[54] and pA-Tn5[32] in single-cell studies.

NanoTag is a multimodal epigenomic profiling approach derived from Omni-ATAC-seq[24] and CUT&Tag[32]. Omni-ATAC-seq provides information about chromatin accessibility and DNAme when using methylated adapters (methyl-ATAC-seq). CUT&Tag is a low-background, low-input method for mapping chromatin-associated binding sites. Building on the advancements brought by Omni-ATAC-seq and CUT&Tag in epigenomics, NanoTag capitalizes on the unique aspects of both approaches to ultimately allow the study of all three epigenetic features simultaneously. NanoTag relies on an innovative protein fusion between the Tn5 transposase and an antibody mimetic, such as a nanobody or a DARPin. The ubiquity of GFP-tagged proteins in biological studies makes GFP an ideal first candidate target for NanoTag. Therefore, a fusion between Tn5 and an anti-GFP nanobody (from here on abbreviated as GBN-Tn5) marks the starting point in our showcasing of the advantages and qualities of NanoTag in epigenomics.

1.2.2. NanoTag – molecular workflow and enzymatic prerequisites

Tn5 is a DNA-cutting transposase, which cleaves DNA within accessible chromatin regions and attaches oligonucleotides to the ends of the DNA fragments. The oligonucleotides with which Tn5 is loaded are comprised of a transposase recognition site Mosaic End (ME) and a sequencing adaptor region[55]. Initially, Tn5 was used to prepare

next-generation sequencing libraries. Recently, it has been adopted for ATAC-seq, enabling the whole-genome profiling of accessible chromatin regions[23].

Conventional antibodies (immunoglobulins, Ig) are large (~150 kDa) heterotetramers, containing two light chains and two heavy chains, each consisting of variable (V_H and V_L) and constant regions[56]. In contrast, heavy chain-only antibodies (HCAbs), found in camelids, are homodimers lacking light chains whose binding region is confined to a single domain – the variable domain of the heavy chain (VHH)[57] (Figure 2). Nanobodies are small (~15 kDa) recombinant single-domain antibodies derived from the HCAb VHH[58]. While immunoglobulins have been an invaluable biological and clinical tool in the past century, their production is very time-consuming and involves the immunization and, usually, sacrifice of animals. In addition, antibodies vary greatly in binding affinity and specificity, can present cell permeability issues due to their large size[57], and their production exhibits batch-to-batch variation[59]. In contrast, nanobodies can be expressed in bacteria and yeast, can recognize less accessible epitopes than conventional antibodies, are more resistant to degradation, are more easily cell permeable, and display low immunogenicity, making them suitable for a wide range of applications – from cell biology to human therapy[60]. These properties of nanobodies facilitated our production of the GBN-Tn5 in *E. coli* and our ability to apply it for NanoTag, which requires the transposase to easily reach nuclear chromatin. Furthermore, our production of GBN-Tn5 in *E. coli* renders the NanoTag strategy highly feasible, technically low-demand and cost-effective.

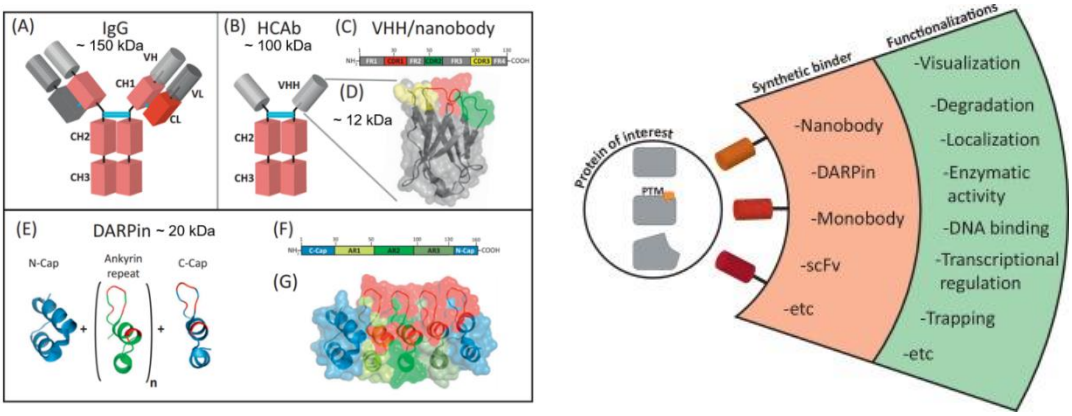


Figure 2 | **Synthetic protein binders and their potential applications**. Single-domain antibodies, VHHs (nanobodies) and designed ankyrin repeat proteins (DARPins) are engineered small antibody mimetics (left), which can be functionalized by fusion to other proteins or enzymes (right). Adapted from Bieli et al. (2016)[58].

In ATAC-seq[23], the Tn5 transposase randomly cleaves DNA within accessible chromatin regions (Figure 3a). In contrast, in CUT&Tag[32] and in NanoTag, Tn5 is directed to specific loci in the genome at which a protein of interest is bound, enabling a targeted epigenomic profiling strategy. In CUT&Tag (Figure 3b), cells are treated with antibodies against the protein of interest, and the protein A-Tn5 fusion is then directed via the IgG-protein A interaction to the specific loci within accessible chromatin at which the target protein is located. Thus, while CUT&Tag is a flexible approach for targeted profiling of chromatin-associated proteins, it depends on the availability of high-specificity antibodies and suffers from loss of material during repeated antibody incubation and washing steps. The lengthy incubations and washes also decrease the possibility that RNA integrity is preserved during the process, rendering CUT&Tag unlikely to be amenable to concomitant transcriptomic analysis.

In contrast, NanoTag targets the loci at which the GFP-tagged protein of interest is bound via the interaction between GFP and the anti-GFP nanobody (Figure 3c, d). Thus, in NanoTag, Tn5 can be directed to specific sites in the genome without the use of antibodies. This makes NanoTag faster, more efficient and amenable to the study of targets for which proper antibodies are not available. While this particular flavor of NanoTag requires the use of transgenic models, fusing Tn5 to other antibody mimics can overcome this (see section 4.2.1). Finally, NanoTag involves a shorter protocol, which does not include lengthy incubations and washes typically required for antibody-based approaches such as ChIP-seq or CUT&Tag. This makes the preservation of RNA integrity during processing more likely, rendering NanoTag more likely to be amenable to simultaneous or parallel transcriptome sequencing.

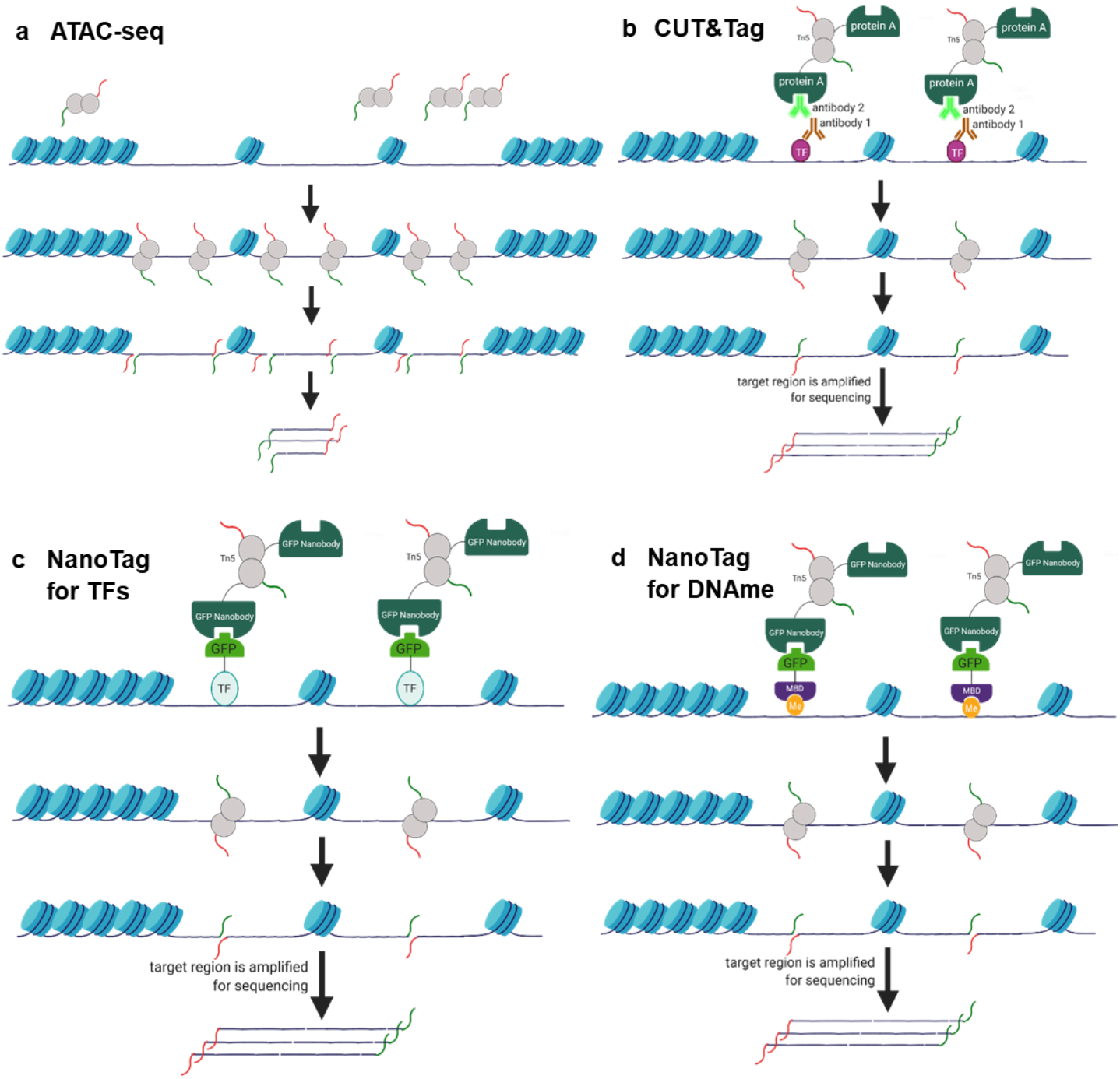


Figure 3 | **Methods for epigenetic analysis.** (**a**) ATAC-seq is used to probe chromatin accessibility within the genome using the Tn5 transposase. Tn5 cleaves the DNA and attaches oligonucleotides (red and green) to the ends of the fragments, which are then amplified and sequenced. (**b**) CUT&Tag is an enzyme-tethering strategy designed for profiling of chromatin features, using a Tn5-protein A fusion, and antibodies against the protein of interest. (**c, d**) NanoTag is a multimodal epigenomic profiling strategy that uses a fusion between Tn5 and an antibody mimetic (here, an anti-GFP nanobody). The nanobody targets the transposase to the loci in the genome at which a GFP-tagged protein of interest is located. Once targeted, GBN-Tn5 cleaves DNA and attaches sequencing adapters. Library preparation and sequencing of the DNA fragments follows.

NanoTag is a highly versatile technique, due to the possibility of exchanging the nanobody fused to Tn5 for any other antibody mimetic. In principle, NanoTag can be used to profile any DNA-bound target, even those against which no antibodies are available. In the case of using GBN-Tn5, any chromatin-associated protein can be profiled in cells, as long as that specific protein is expressed as a fusion to GFP. In cases where transgenics is not possible or not desired, a fusion between Tn5 and a custom engineered antibody mimetic (nanobody or DARPin) could be produced to directly target the protein or epigenetic mark of interest.

A fully *in vitro* nanobody discovery approach has been described, in which a yeast-display library is used to screen for nanobodies that exhibit high affinity for the target of interest[61]. Once the best nanobody is identified from the yeast-display screen, it can be readily expressed in bacteria and purified with high yield[60]. Design ankyrin repeat proteins (DARPins) are a second class of small protein binders. An *in vitro* ribosome display screening approach has also been described for the selection of DARPins with affinity towards the target of choice[62], followed by expression in *E. coli* and high-yield purification[63]. Thus, producing a fusion between Tn5 and a custom procured nanobody or DARPin would be readily achievable.

1.2.3. Epigenomic applications of NanoTag

NanoTag is a multimodal epigenomic profiling strategy because it provides the possibility of obtaining information regarding chromatin accessibility, DNA methylation and chromatin-associated protein binding sites simultaneously in one experiment. This unlocks the ability to study interactions between several epigenetic factors and their effects on the same genomic loci. Compared to methyl-ATAC-seq, which provides information on DNA methylation and chromatin accessibility at random sites within the genome, NanoTag would provide this information for the specific sites at which the protein of interest is located. This is advantageous because fewer reads are then required to obtain the desired information and the signal-to-noise ratio is improved, resulting in superior sequencing quality. To obtain such methylome sequencing data, we would anneal GBN-Tn5 to methylated sequencing adapters and use these transposomes to

conduct NanoTag. Once the DNA is tagmented and purified, it would be subjected to bisulfite treatment, followed by library preparation and sequencing.

In addition, using NanoTag, more than one chromatin-associated protein could be profiled in a single experiment, by treating the lysed nuclei with a mixture of different types of transposomes. Distinct types of transposomes could be prepared by annealing different nanobody-Tn5 fusions (eg. GBN-Tn5 and mCherry nanobody-Tn5, see section 4.2.2) with barcoded sequencing adapters. After sequencing, the barcodes serve as identifiers, which are used to distinguish which of the two Tn5 fusions was bound at each locus in the genome. The concurrence of two or more barcodes at one locus could then suggest interactions of multiple targets at specific loci.

To illustrate the kinds of biological questions NanoTag would help answer, a study by Ji et al. is a noteworthy example[64]. To investigate the epigenetic switch and transcription factors involved in transformation of epithelial breast cells, the AccessTF computational approach was developed, which uses the information provided by DNase-seq and ChIP-seq experiments to predict which transcription factors are bound to their DNA motifs *in vivo* in a particular cellular state. A score schema (TFScore) was developed, which uses the information provided by AccessTF to predict the most functionally relevant transcription factors involved in a particular biological process[64]. Using NanoTag, one would be able to obtain information regarding chromatin accessibility, DNA methylation, and either two of the histone modifications investigated by Ji et al. (eg. H3K4me3 and H3K27ac) or two of the most important TFs predicted by TFScore (eg. STAT3 and JUNB) within a single experiment. A second biological process whose investigation could be facilitated by NanoTag is the interaction between obesity-related transcription factors (such as PPARs) and epigenetic modifications (mCpGs and histone PTMs such as H3K9ac or H3K27ac) in adipocyte differentiation[65]. Thirdly, NanoTag could also be used to study the interactions between epigenetic mechanisms (DNA methylation and H3K4 methylation) and transcription factors, such as NFKB, in the induction of autoimmunity[66]. Thus, by providing unique insight into epigenomic interactions, NanoTag will further understanding of physiological and pathological cellular processes.

To summarize, NanoTag facilitates the study of interactions between epigenetic factors in the most native-like state, enabling an epigenomic and transcriptomic characterization of specific cellular states. In this way, NanoTag could further the understanding of disease-underlying epigenomic events, and may, ultimately, facilitate the search for novel drug targets.

1.3. Optimizing NanoTag

The in-house purification of Tn5 (and of Tn5 fusion derivatives), while enabling immense flexibility in experimental design and being cost-effective, does present an important challenge. Variability in tagmentation activity between different Tn5 batches may result in different library preparation efficiencies and may constitute a confounding factor for inter-experiment comparisons. This is why, for example, ATAC-seq protocols recommend the preparation of samples using different concentrations of transposomes and the subsequent selection of the appropriate concentration based on assessment of library quality.

To account for such variability in enzymatic activity between batches, Rykalina et al. propose a qPCR-based strategy (Figure 4)[67]. In this approach, the amplification rate of Tn5-treated plasmid DNA is compared to that of untreated DNA and used to calculate the difference in amplification latency between the two samples. This difference in amplification latency is used to quantify the tagmentation efficiency of a particular transposase batch.

Treating linearized plasmid DNA with active Tn5 leads to cleavage of the DNA molecules at random sites. By chance, some molecules become cleaved within the region that is amplified in qPCR, leading to disruption of the amplification template in those molecules and, overall, slower amplification of transposase-treated DNA than of untreated plasmid[67]. This difference in amplification latency can be quantified for each batch of transposomes produced and used to adjust the concentration of transposase that is applied during subsequent ATAC-seq and NanoTag experiments, respectively. So far, we have used this approach to qualitatively assess whether the GBN-Tn5 transposomes

we produced are functional. In the future, we will also apply this strategy to estimate the proper concentration of GBN-Tn5 for NanoTag.

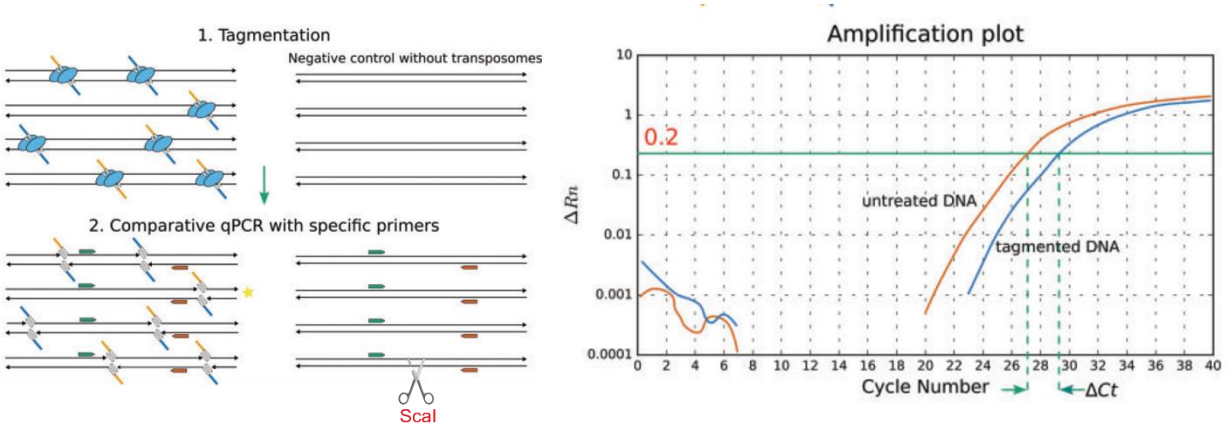


Figure 4 | **qPCR-based strategy for tagmentation efficiency test.** Transposase-treated DNA contains molecules in which the region acting as PCR template is disrupted, leading to later amplification during qPCR of tagmented DNA (blue line in amplification plot). ScaI cleaves the pUC19 sequence within the region of the amplification template. In ScaI digested pUC19, all templates should thus theoretically be disrupted and amplification should be maximally delayed (not depicted in amplification plot). Thus, ScaI digested pUC19 can serve as a positive control for successful tagmentation. Adapted from Rykalina et al. (2017)[67].

Aside from a proper concentration of transposomes, other factors may influence the efficiency of NanoTag and, therefore, require further consideration for optimal sequencing results. Inspired by the CUT&Tag protocol, we designed the NanoTag protocol to include a two-step transposition process, with the purpose of minimizing untargeted tagmentation by the hyperactive Tn5 (see section 2.6 for technical details). First, the binding of GBN-Tn5 to its targets is allowed to occur during a one-hour room temperature incubation of the permeabilized nuclei with transposomes. By manipulating the buffer conditions, inactivity of Tn5 during this step is promoted, so that the nanobody can interact with its target first. This minimizes unspecific tagmentation. Then, we perform three short washing steps with the goal of removing unbound GBN-Tn5 from the DNA. Finally, we add to the nuclei a buffer containing $MgCl_2$, which acts as a catalyst for transposition, and incubate the reaction at 37ºC. This temporal control over enzymatic activity can be used to optimize incubation times for minimal untargeted transposition. In addition, harsher detergents can be employed if insufficient tagmentation is observed during NanoTag.

To summarize, standardization of GBN-Tn5 transposome production and optimization of the NanoTag protocol over time will perfect our strategy and allow us to achieve superior DNA library quality.

1.4. Chromatin Readers

To test the NanoTag approach in a proof-of-concept experiment we employed a system put in place by the group of Tuncay Baubec. They have developed a series of "engineered chromatin reader" proteins (eCRs), which can recognize and bind specific chromatin modifications and have established mouse embryonic stem cell lines that stably express eGFP-tagged eCRs[68] (Figure 5).

The mCpG-binding reader MBD1 and the H3K4me3 reader TAF3 were the first eCRs to be tested with NanoTag. The mCpG eCR was derived from the MBD1 member of methyl-CpG binding proteins, which acts as a transcriptional repressor[69]. The H3K4me3 eCR was derived from the metazoan TAF3 subunit of the basal transcription factor TFIID, which contains a plant homeodomain (PHD) finger that selectively binds to H3K4me3 sites[70]. An important prerequisite for the ability of our approach to specifically target loci in the genome was the use of a nanobody that recognizes and binds its target with high affinity. The high affinity with which the GFP-binding nanobody binds to GFP had already been characterized[71]. To confirm the specificity of NanoTag, it was essential to include negative controls in our experiment. Fortunately, the same group had also designed inactive mutants of each eCR, which were, similarly, expressed in mESC lines as fusions to eGFP. To prevent MBD1 from recognizing DNAme, the Arginine 22 residue in the active site of MBD1 was substituted with Alanine (R22A), resulting in an inactive DNAme eCR[68,72]. The H3K4me3 eCR was only observed to be functional by Villaseñor et al.[68] when expressed as a dual chromatin reader domain (see Figure 5). When a single copy of the taf3 gene was expressed in cells, the eCR was unable to recognize H3K4me3, thus showing the inactivity of the monomeric TAF3 reader. Thus, as negative controls in our NanoTag experiments, we included: wild-type mESC, mESC expressing inactive eCRs, and, for the mCpG profiling experiment, cells devoid of DNA methylation.

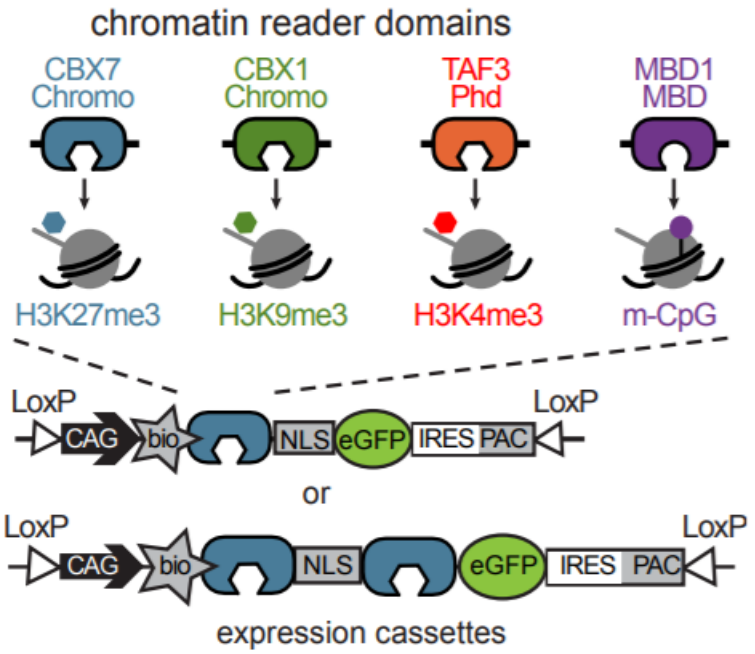


Figure 5 | **Engineered chromatin reader domains.** Engineered chromatin readers recognize specific histone modifications and DNA methylation (top) and express these proteins as a fusion to eGFP in mouse embryonic stem cells. Engineered chromatin reader (eCR)-expressing mES cells constitute a suitable model to test NanoTag in – Tn5 should be targeted to the loci within the genome at which the modification recognized by the eCR is located via the anti-GFP nanobody-eGFP interaction. Reproduced from Villaseñor et al. (2020)[68].

Conducting NanoTag on eCR expressing cells should, thus, result in genome-wide profiling of mCpGs and of H3K4me3 binding sites and provide chromatin accessibility information in mESC. After sequencing DNA libraries generated from cells expressing active and inactive eCRs, the signal difference would provide information about the enrichment of the respective eCR at specific locations. The enrichment profiles can then be compared to previously published ChIP-seq data to verify the specificity and efficacy of NanoTag.

To summarize, NanoTag is a multimodal epigenomic profiling strategy derived from two well characterized and established epigenomic sequencing approach. We discussed how NanoTag could be applied for the investigation of epigenome-related biological processes and diseases. Based on existing recommendations and studies, a strategy to optimize GBN-Tn5 activity has been formulated. Finally, we described a proof-of-concept experiment in which NanoTag was applied in transgenic cell lines to profile epigenetic features.

2. Methods

2.1. Molecular cloning

pGEX6P1-GFP-Nanobody was a gift from Kazuhisa Nakayama (Addgene plasmid # 61838; http://n2t.net/addgene:61838; RRID:Addgene_61838)[73], and 3XFlag-pA-Tn5-Fl was a gift from Steven Henikoff (Addgene plasmid # 124601; http://n2t.net/addgene:124601; RRID:Addgene_124601)[32]. The GFP nanobody-Tn5 (GBN-Tn5) gene was cloned by replacing the protein A gene in the 3X Flag pA-Tn5-Fl plasmid with the GFP nanobody gene from the pGEX6P1-GFP-Nanobody plasmid. In this way, a new plasmid (the GBN-Tn5 plasmid) was designed and cloned (for complete sequence see Appendix), which encoded the fusion between the GFP nanobody and Tn5. In our plasmid, GBN-Tn5 is fused to a 3X Flag tag at the N-terminus and to a intein-CBD tag at the C-terminus, as was the case for the protein A-Tn5 gene encoded by the original plasmid. This cloning strategy was employed to allow for usage of the same purification method for GBN-Tn5 as that described for pA-Tn5[32] and for Tn5[74].

The first step of the cloning strategy was to conduct PCR in order to attach *HindIII* and *EcoRI* restriction sites to the ends of the GFP nanobody gene, which would allow for subsequent ligation into the host plasmid. Two sets of primers (provided by Microsynth, Table 1) were designed that contained a 18-21 bp region complementary to the end of the GFP nanobody gene and the restriction site for the *HindIII* (New England Biolabs # R0104S) and *EcoRI* (New England Biolabs # R0101S) restriction enzymes, respectively. 1-4 nucleotides were included between the restriction site and the GFP nanobody-complementary region in each primer to ensure the maintenance of the open reading frame (ORF) in the new plasmid and the integrity of the inserted gene.

Table 1 | **Oligonucleotides used for the production and testing of GBN-Tn5.** The first 4 rows list the primers used for PCR (first 4 rows) to amplify the GBN gene. The following three rows list the oligonucleotides annealed to GBN-Tn5. The last two rows list the primers used in the 1PCR-based tagmentation efficiency test.

	Oligonucleotide	Sequence
Primers for PCR of GBN	Set 1_FWD primer	GATCAAGCTTGGGCGTTCAGCTGGTTGAAAGCGG
	Set 1_REV primer	CTAGGAATTCGCCGCTGCTAACGGTAAC
	Set 2_FWD primer	GATCAAGCTTGGGCGTTCAGCTGGTTGAAAGCG
	Set 2_REV primer	CTAGGAATTCGCCGCTGCTAACGGTAACCTG
Oligonucleotides annealed to GBN-Tn5	Tn5ME-A	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
	Tn5ME-B	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
	Tn5ME-rev	[phos]CTGTCTCTTATACACATCT
Primers for qPCR of pUC19	610 bp_FWD primer	CCTATCTCAGCGATCTGTCTATTTC
	610 bp_REV primer	GCGCGGTATTATCCCGTATT

50-μL reactions were set up for PCR using 10 ng of template (pGEX6P1-GFP-Nanobody plasmid), 0.5 μM primers, 200 μM dNTPs, 1 U of Phusion High-fidelity DNA Polymerase (Thermo Scientific # F530S) and 10 μL of 5X Phusion buffer. Temperature gradient PCR (using a C1000 Thermal Cycler from Bio-Rad) was conducted for each set of primers with annealing temperature varied as follows: set 1 – 49.9ºC - 59.7ºC and set 2 –-54.5ºC - 64.5ºC. The cycling conditions used were: initial denaturation for 30s at 98ºC; 25 cycles of: denaturation at 98ºC for 10s, annealing (50ºC - 60ºC /55ºC -65ºC), extension for 15s at 72ºC; final extension for 7 min at 72ºC. Out of each 50 μl reaction, 10 μl were used for electrophoresis to check for amplification. The PCR products of the 64.5ºC and 63.8ºC annealing with primer set 2 were pooled and used further.

Next, the PCR product and the host (3X Flag pA-Tn5-Fl) plasmid were digested with *HindIII* and *EcoRI* in parallel reactions, to create sticky ends for subsequent ligation. 50-μl double digestion reactions were set up containing: 1 μg of DNA (PCR product / host plasmid, respectively), 30 U of *HindIII*, 30 U of *EcoRI*, 5 μl of 10X NEBuffer 3.1. Since each of the restriction enzymes was reported to display 50% activity in the common buffer, 30 U of enzyme were used. Digestion reactions were incubated at 37ºC for 1.5 hours.

The digestion products were separated through 0.8% agarose gel electrophoresis. The band size of the digested plasmid was 8240 bp, with a small band visible at 426 bp corresponding to the pA-Tn5 excised region, and the digested PCR product band size was 360 bp. The bands corresponding to the digested PCR product and to the linearized host plasmid (lacking the pA-Tn5 gene) were removed from the gel under UV light, weighed and the DNA extracted using the QIAquick gel extraction kit (Qiagen # 28115).

To insert the GFP nanobody gene into the digested host plasmid, ligation reactions were set up using different insert : plasmid molar ratios (3:1, 5:1, 7:1). 1 U of T4 DNA ligase (Thermo Scientific #EL0014), 2 μl of 10X T4 DNA Ligase Buffer, 40 ng of host plasmid and 5.2, 8.7 and 12.2 ng, respectively, of insert were used to set up the three 20-μl reactions. Ligation reactions were incubated for 10 minutes at 22ºC.

The ligation product was then used for transformation of Subcloning Efficiency DH5α competent cells (Invitrogen LS18265017). This type of competent cells exhibits high transformation efficiency, increasing the chance that the ligated plasmid is incorporated by bacteria. Competent cells were thawed on ice for 30 minutes. 10 μl of ligation product was added to 50 μl competent cells, mixed gently and incubated on ice for 30 minutes, followed by heat shocking at 42ºC for 20 seconds. Samples were immediately incubated on ice for 2 minutes. 950 μl of LB medium with no antibiotic was added to each reaction and incubated at 37ºC with shaking at 225 rpm for 1 hour. 20 μl and 200 μl of each reaction were spread on separate LB/Ampicillin (100 μg/ml) plates and incubated overnight at 37ºC. Each colony was picked and used to inoculate 8 mL of LB + Ampicillin (100 μg/ml). Cultures of every colony were incubated overnight at 37ºC with 225 rpm shaking, then centrifuged at 5,000 RCF, 4ºC for 15 minutes. Pellets were then used to extract the plasmid using the QIAprep Miniprep kit (Qiagen #27104) and each plasmid was sequenced. Plasmids containing mutations were discarded, and one plasmid was selected for further use. To enable subsequent protein expression, the plasmid was used to transform *T7 express* competent *E. coli* (New England Biolabs #C2566I). Competent cells were thawed on ice for 30 minutes, 100 ng of plasmid DNA was added to 50 μl of cells, mixed gently and incubated on ice for 30 minutes. The mixture was then heat shocked at 42ºC for 10 seconds and incubated on ice for 5 minutes. 950 μl of SOC medium (20 g/l Tryptone, 5 g/l Yeast Extract, 4.8 g/l $MgSO_4$, 3.603 g/l dextrose, 0.5g/l NaCl, 0.186 g/l KCl) was added and the reaction incubated at 37ºC for 1 hour with mixing at 250 rpm. Then, 20 μl and 200 μl of each reaction was spread onto separate LB + Ampicillin plates and incubated overnight at 37ºC. One colony was used to inoculate 8 ml of LB, incubated overnight at 37ºC and used to prepare 50% glycerol stocks for long term storage at -80 ºC and for protein expression.

2.2. Protein expression and purification

A pipette tip was dipped into glycerol stock of *T7 express E coli* transformed with GFP Nanobody-Tn5 plasmid and used to inoculate 50 ml of LB/Ampicillin (100 µg/mL) medium. The primary culture was incubated overnight at 37ºC and 10 ml of the primary culture was then used to inoculate two 2.5-liter flasks containing 500 ml of LB/Ampicillin. The secondary culture was incubated at 37ºC with shaking at 225 rpm for approximately 3 hours, until OD600 reached at least 0.5, then cultures were cooled to 10ºC (through incubation on ice in the 4ºC room for 10 minutes) and then protein expression was induced with 0.25 mM IPTG. Cultures were incubated at room temperature with shaking (130 rpm) for another 4 hours. This incubation was performed sub-optimally due to lack of an incubator equipped with refrigeration. Ideally, the incubation would have been performed at 23ºC with 225 rpm shaking. Finally, the bacterial cells were collected by centrifugation at 5000 RCF at 4ºC for 15 minutes and pellets were stored at -80ºC until further processing (Figure 6).

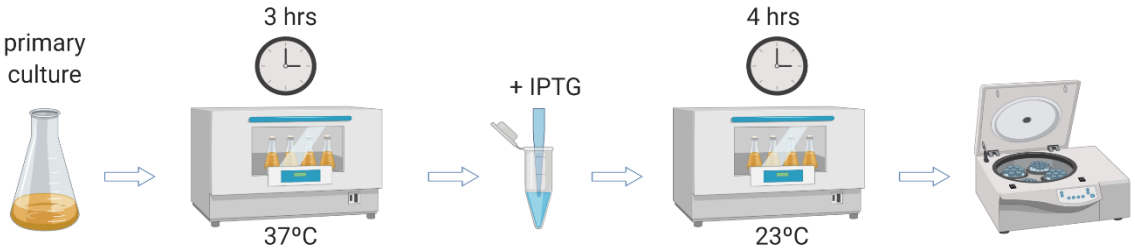


Figure 6 | **Summary of bacterial culture for protein expression**. The primary culture was used to inoculate a larger secondary culture, incubated for approximately 3 hours, protein expression was then induced with 0.25 mM IPTG. Incubation was continued for 4 additional hours, followed by centrifugation to collect the bacterial pellets.

GBN-Tn5 purification was conducted according to the protocol described by Picelli et al.[74], summarized in Figure 7. The bacterial pellet was resuspended in 80 ml HEGX buffer (20 mM HEPES-KOH at pH 7.2, 0.8 M NaCl, 1 mM EDTA, 10% glycerol, 0.2% Triton X-100), with cOmplete Protease Inhibitor Cocktail (Roche 11697498001), transferred to a cold 100-ml glass bottle and lysed via sonication on ice in a Branson 250 Sonifier in two cycles at 50% duty cycle, output 5 for 15 minutes. The two 15-minute sonication cycles were

interrupted for 15 minutes, to ensure the lysate remained cool. The lysate was centrifuged at 15,000 RCF for 30 minutes at 4ºC. The supernatant was transferred to a cold glass bottle, and 2 ml of 10% PEI, pH 7.9 were added dropwise while stirring at 4ºC to precipitate DNA. The mixture was centrifuged at 12,000 RCF for 10 minutes at 4ºC and the supernatant was loaded onto a 10-ml chitin column.

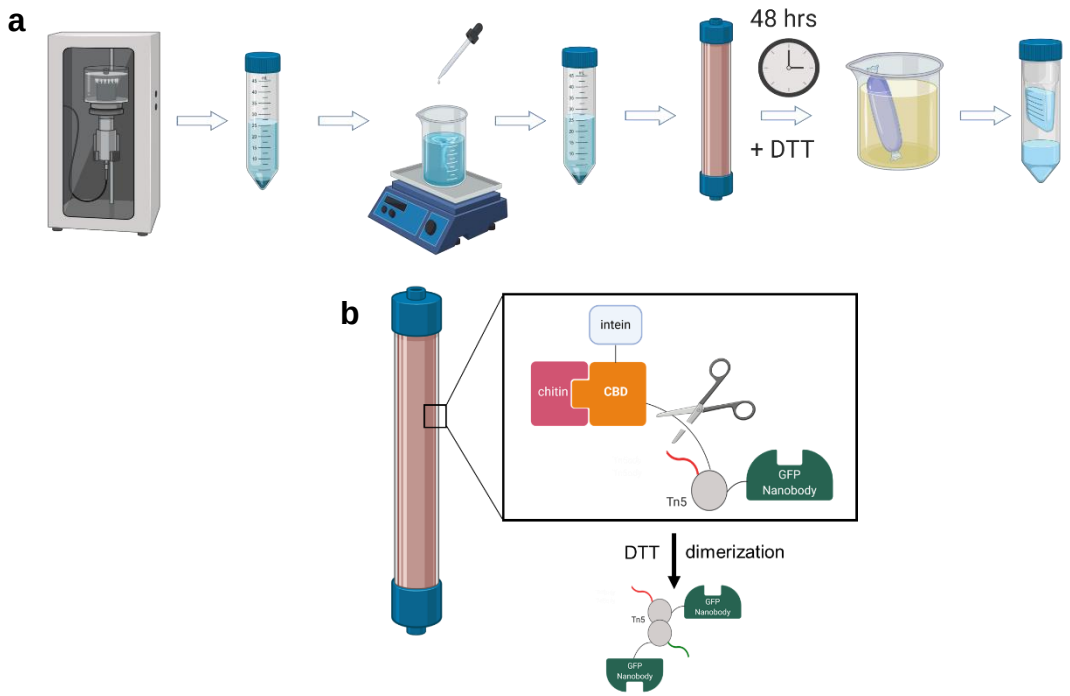


Figure 7 | **GBN-Tn5 purification workflow**. Summary of the steps involved in the purification protocol **(a)**: bacterial cells resuspended in lysis buffer are lysed by sonication, centrifuged, PEI is added dropwise while stirring to the mixture, which is then centrifuged again, and the supernatant is loaded onto the chitin column and incubated overnight. DTT is then added to the column and incubated for 48 hours. Finally, the protein is eluted, dialyzed and concentrated. The purification protocol involves affinity chromatography; **(b)** using a chitin column, to which the protein binds via its intein-CBD tag. This tag is then self-cleaved in the presence of DTT to release the pure protein.

The protein purification setup we used (including the chromatography column) is depicted in Figure 8. The column was washed overnight with 500 ml HEGX buffer. Self-cleavage of the intein-CBD tag was induced by loading 15 ml of HEGX 100 mM DTT onto the column and incubating for 48 hours. Another 15 ml of HEGX 100 mM DTT buffer was used to elute the protein from the column. The protein solution was then dialyzed against 1L of dialysis buffer (100 mM HEPES-KOH at pH7.2, 0.2 M NaCl, 0.2 mM EDTA, 2mM DTT, 0.2% Triton X-100, 20% glycerol) for two hours. The buffer was exchanged, and the

dialysis continued for another two hours. Following dialysis, the protein sample was concentrated approximately 10-fold, bringing the protein sample from about 10.5 ml to 1.5 ml. The protein was stored as a 50% glycerol stock at -20ºC.

Figure 8 | **The protein purification setup we used to produce GBN-Tn5**. A peristaltic pump was connected to the column to ensure continuous flow.

2.3. Production of functional transposomes

To verify the identity and molecular weight of the GBN-Tn5 we produced, a protein aliquot was used for Western Blot, taking advantage of the N-terminal 3X Flag tag of the GBN-Tn5. 20-µl protein samples were prepared (by diluting with water) using 4X loading buffer (200mM Trizma Base pH 6.8, 8% SDS, 0.4% Bromophenol blue, 40% glycerol) containing 10% β-mercaptoethanol). Samples were vortexed, centrifuged and incubated at 95°C for 5 minutes, then loaded onto a precast 4-20% polyacrylamide preMini-PROTEAN TGX gel (Bio-Rad #4561094) in parallel with 3 µl of MW marker (Precision Plus Protein™ Dual Color Standards, Bio-Rad #1610374). Electrophoresis was conducted on ice, using cold buffer (25 mM Trizma base, 250 mM Glycine, 0.1% SDS, pH 8.3) at 90 V (constant voltage) for 20 min, then the voltage was increased to 120 V and electrophoresis was

continued until the dye front reached the bottom of the gel. The gel was removed from the electrophoresis chamber and proteins were transferred to a nitrocellulose membrane (Trans-Blot Turbo Mini Nitrocellulose Transfer Pack, Bio-Rad #170-4158) using the TransBlot Turbo device (Bio-Rad) set to default protocol for high MW proteins (1.3 A, 25 V, 10 minutes). The membrane was then incubated in 5% SureBlock (Lubio Science # SB232010) in TBS-T, on a rocking platform at room temperature for 60 minutes, followed by incubation in a 50-ml tube containing 4 ml of the primary antibody solution (4 ml blocking solution, 8.6 µl anti-Flag antibody, Abcam #ab18230) and incubated over night at 4°C on a rotator. The membrane was washed 3 times for 10min in TBS-T and then incubated in 15 ml secondary antibody solution (15 ml blocking solution, 1.5 µl Goat anti rabbit IRDye 800CW (LI-COR, # 926-32211)) on a rocking platform for 1 hour at room temperature. The membrane was again washed 3 times for 10min in TBS-T and then scanned using an Odyssey infrared imaging scanner (LI-COR). The scanned membrane was analyzed using the Image Studio Lite software (LI-COR).

Tn5 and its derivatives act by cleaving DNA and attaching adaptors to the ends of the fragments. Thus, the functional unit of Tn5 is a transposome: a Tn5 dimer, in which each monomer is annealed to a distinct oligonucleotide duplex. These oligonucleotide duplexes contain one identical strand (Tn5ME-rev) and one distinct strand (Tn5ME-A or Tn5ME-B). The sequences of these oligonucleotides (provided by IDT) are shown in Table 1 and are according to Picelli et al[74]. In order to create functional transposomes, the GBN-Tn5 protein must be mixed in a 1:1 molar ratio with the oligonucleotide duplexes. Thus, quantification of the protein concentration is crucial for this step.

We attempted different methods of quantifying the protein concentration, including using NanoDrop 1000, Qubit, the Pierce BCA Protein Assay Kit (Thermo Scientific 23225) and SDS-PAGE electrophoresis, however they all proved problematic (see section 3.3 for details) due to various components of the protein storage buffer: 100 mM HEPES-KOH at pH7.2, 0.2 M NaCl, 0.2 mM EDTA, 2mM DTT, 0.2% Triton X-100, 20% glycerol. Quantification by NanoDrop 1000 was deemed unreliable due to excessive inconsistency between protein dilutions. Quantification by Qubit was not appropriate due to sensitivity to detergent. Quantification by BCA assay was then attempted due to insensitivity to

detergent. All other possibly interfering substances were in the protein buffer at concentrations with which the BCA assay was compatible, according to the manufacturer instructions. However, measurements taken with this assay were also unreliable. We attempted this assay in different ways, in an effort to find a suitable strategy: using very dilute protein samples to minimize interference and comparing with water-diluted standard; using 40-fold dilutions of protein sample and adding the same amount of buffer to the standards. However, regardless of the strategy employed, reliable measurements could not be obtained. After concluding that the NanoDrop 1000, Qubit and BCA assays fail to provide reliable results, we attempted to quantify the protein sample by comparing it to lysozyme standards of known concentration in SDS-PAGE electrophoresis, however this is an unreliable method due to the fact that the Coomassie stain used in this method binds with different affinities to different proteins. To circumvent this quantification issue, we prepared several transposome batches, by mixing equal volumes of protein diluted to different degrees and of oligonucleotide duplexes at 25 μM and used the qPCR-based tagmentation efficiency assay described by Rykalina et al.[67] to identify which of the prepared transposome batches displayed the highest tagmentation efficiency (see sections 2.4 and 3.4.1). The most efficient transposome batch was used further to conduct NanoTag and was assumed to constitute a 1:1 molar ratio mixture of GBN-Tn5 and oligonucleotide duplexes. From this, the concentration of the protein stock was estimated (see section 4.1 for potential solutions).

To obtain oligonucleotide duplexes, lyophilized oligonucleotides from IDT were resuspended in 10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 1 mM EDTA. Equal volumes of Tn5ME-rev and Tn5ME-A, Tn5ME-rev and Tn5ME-B, respectively, were mixed in two separate tubes, incubated at 95ºC for 2 minutes and cooled over 45 minutes to 25ºC to create duplexes, which were then mixed in equal volumes and stored as a 50% glycerol solution. The oligonucleotide mixture was then mixed in a 1:1 molar ratio with GBN-Tn5 50% glycerol stock and incubated at room temperature for 1 hour. The transposome solution was stored at -20ºC until use.

2.4. Quantifying tagmentation activity

To assess the tagmentation ability of GBN-Tn5 transposomes, we employed the approach designed by Rykalina et al[67]. in which qPCR is used to compute the difference in amplification (ΔCp) of a specific region between tagmented plasmid DNA and untreated DNA. EcoRI-digested pUC19 (Carl Roth #X911.1) was used for tagmentation. Rykalina et al.[67] describe four different amplification regions and four corresponding sets of primers for this experiment. In our experiments, we always employed amplification of a 610-bp region, as this provided satisfying results while minimizing the reaction time (amplifying larger fragments would significantly increase the reaction time). The primers used to amplify this 610-bp region are listed in Table 1 and are according to Rykalina et al.[67] In addition, we introduced a positive control in this experiment – ScaI-digested pUC19. The 610-bp amplification region contains one ScaI restriction site. Thus, amplification of ScaI/pUC19 occurs very slowly, due to disruption of template molecules. 20-µl tagmentation reactions were prepared containing: 50 ng EcoRI/pUC19 DNA, 4 µl of 5X TAPS-DMF (50 mM TAPS-NaOH at pH 8.5, 25 mM MgCl2, 50% DMF) and variable volumes of GBN-Tn5 transposomes. Reactions were incubated at 55ºC for 7 minutes and inactivated by addition of 5 µl of 0.2% SDS, followed by incubation at 55ºC for 7 minutes. Tagmentation products were purified in the Diagenode IP-Star machine, using the IPure protocol, with 20 µl AMPure XP beads, and DNA was eluted in 10 mM Tris-HCl, pH 7.5. 10-µl qPCR reactions were prepared containing: 2 µl pure tagmented DNA, 5 µL LightCycler 480 SYBR Green I Master (Roche Diagnostics #04707516001) and 0.1 µM primers. qPCR was conducted with the following conditions: 95ºC for 10 min followed by 50 cycles of 95ºC for 15 s and 63ºC for 15 sec and 72ºC for 60 s. Once the qPCR was completed, ΔCp values for each sample of tagmented DNA were computed by subtracting the Cp of untreated DNA from that of tagmented DNA. Transposome batches showing the highest ΔCp values, and the highest differences in amplification rate between transposome concentrations were assumed to be the most efficient.

2.5. Tagmentation of cDNA

To further test the activity of GFP nanobody-Tn5 transposomes, cDNA was tagmented, according to Picelli et al.[74] 1 µl of diluted mouse liver RNA was reverse-transcribed,

preamplified using 18 PCR cycles to mimic the single-cell protocol for cDNA preparation, purified with 25 μl AMPure XP beads, following precisely the smart-seq2 protocol by Picelli et al[75]. The cDNA library was assessed for fragment size distribution using the Bioanalyzer. The Bioanalyzer profile was also used to quantify the cDNA concentration. Subsequently, 20-μl tagmentation reactions were prepared containing: 4 ng cDNA, 4 μl 5X TAPS-DMF buffer and variable concentrations of transposomes. Only reactions containing transposomes prepared with unconcentrated GBN-Tn5 exhibited tagmentation activity in this assay (see section 3.4.1 for details). Tagmentation reactions were incubated at 55ºC for 7 minutes and inactivated by addition of 5 μl 0.2% SDS followed by incubation at room temperature for 5 minutes. 25 μl of 2X KAPA HiFi HotStart ReadyMix (Roche # 7958927001) and 1 μl each of Illumina Nextera adaptors (i7 and i5, illumine FC-121-1031) were added to the 25 μl tagmentation reactions and enrichment PCR was conducted under the following conditions: 3 minutes at 72ºC, 30s at 95ºC, and 20 cycles of 10 s at 95ºC, 30 s at 55ºC and 30 s at 72ºC. The use of the KAPA HiFi DNA Polymerase (KAPA Biosystems # KK2101), a non-hot start enzyme, was recommended by Picelli et al.[74] but in our case failed to generate libraries. The 52-μl PCR products were purified using 52 μl of AMPure XP beads and DNA was eluted in 15 μl EB buffer. Finally, tagmented cDNA libraries were assessed using the Agilent High Sensitivity DNA Kit (Agilent #5067-4626) on the Bioanalyzer.

2.6. NanoTag

NanoTag was performed on mES cells expressing a fusion between eGFP and the corresponding chromatin reader domain, as described in sections 3.5 and 3.6.

The NanoTag protocol is derived from the Omni-ATAC-seq protocol by Corces et al.[76] and the single-cell CUT&Tag protocol by Kaya-Okur et al.[32] Transposition is performed in two steps: first, the lysed nuclei are incubated with transposomes in the absence of $MgCl_2$ for 1 hour at room temperature, to allow for interaction between the GFP nanobody and eGFP and, therefore, of targeting of the transposome complexes to specific loci in the genome. This step is followed by three washes to remove any unbound GBN-Tn5. Only then is $MgCl_2$ added to the mixture and transposition is allowed to occur. This is done to minimize unspecific transposition, similarly to the CUT&Tag protocol. Spermidine is used

as a chromatin-stabilizing agent in the absence of $MgCl_2$, again according to Kaya-Okur et al[32].

Detailed NanoTag Protocol

Required equipment:

Table-top refrigerated centrifuge, thermomixer, thermocycler, qPCR machine

Reagents:

Table 2 | **Reagents required to conduct NanoTag experiments.**

Reagent	Manufacturer - Catalog number	Further preparation
GBN-Tn5 transposomes	in-house produced	
1 M Tris-HCl Buffer, pH 7.5	Invitrogen #15567027	
5 M NaCl	Invitrogen #AM9760G	
Spermidine	Sigma-Aldrich #S0266	200 mM solution
10% NP40	Sigma/Roche # 11332473001	
10% Tween-20	Sigma/Roche #11332465001	
Digitonin	Sigma-Aldrich #D141	2% in DMSO
1 M $MgCl_2$	Invitrogen #AM9530G	
0.5 M EDTA	Fischer Scientific #PR-V4233	
MinElute Reaction Cleanup Kit	Qiagen #28204	
Nextera XT Index Kit	Illumina #FC-131-1001	
NEBNext High-Fidelity 2X PCR Master Mix	NEB #M0541L	
SYBR Gold Nucleic Acid Gel Stain	Thermo #S-11494	

Buffer preparation:

Table 3 | **Preparation of buffers required to conduct NanoTag experiments.**

Buffer	Preparation
Resuspension buffer	Dilute 500 µl of 1 M Tris-HCl pH 7.5 and 100 µl of 5 M NaCl in 49.4 mL nuclease-free water.
Initial wash buffer - **Buffer A**	Add 5 µl 200 mM Spermidine to 1995 µl resuspension buffer per sample.
Lysis buffer - **Buffer B**	Add 0.5 µl 10% NP-40, 0.5 µl 10% Tween-20, 0.25 µl 2% Digitonin and 0.125 µl 200 mM Spermidine to 48.625 µL of resuspension buffer per sample
Wash buffer - **Buffer C**	Add 40 µl 10% Tween-20 and 10 µl 200 mM Spermidine to 3950 µl resuspension buffer per sample
Transposition buffer 1 – **Buffer D**	Add 5 µl 10% Tween-20, 2.5 µl 2% Digitonin, 1.25 µl 200 mM Spermidine and 10 µl GBN-Tn5 transposomes to 481.25 µl resuspension buffer per sample
Transposition buffer 2 – **Buffer E**	Add 5 µl 10% Tween-20, 2.5 µl 2% Digitonin, 1.25 µl 200 mM Spermidine and 1.5 µl 1M $MgCl_2$ to 139.75 µl resuspension buffer per sample
Inactivation buffer – **Buffer F**	Add 4.04 µl 0.5 M EDTA to 45.96 µl nuclease-free water per sample

Procedure:

All steps until incubation in transposition buffer D are performed at 4ºC on ice.

Cell lysis & DNA transposition

- Pellet 50,000 viable cells by centrifuging at 500 RCF, 4ºC for 5 minutes;
- Remove supernatant, wash cells twice in 1 ml Buffer A;
- Aspirate entire supernatant, carefully avoiding the small pellet;
- Add 50 µl cold Buffer B; pipette up and down gently three times to mix;
- Incubate on ice for 3 minutes to allow cellular lysis to proceed;
- Wash lysis buffer by adding 1 ml Buffer C and invert tube three times gently to mix;
- Centrifuge at 500 RCF, 4ºC for 10 minutes to pellet nuclei;
- Aspirate entire supernatant, carefully avoiding the nuclear pellet;
- Resuspend pellet in 500 µl Buffer D by gently pipetting up and down 6 times;
- Incubate at room temperature for 1 hour on a rotator at minimum speed to allow for interaction between the nanobody and GFP to occur;
- Centrifuge at 300 RCF for 3 minutes; remove supernatant; add 1 ml Buffer C and incubate for 5 minutes to wash off excess transposomes;
- Repeat previous step twice;
- After removing the supernatant, resuspend the nuclear pellets in 150 µl Buffer E;
- Incubate at 37ºC for 1 hour with mixing at 1000 rpm to allow transposition to occur;
- Stop the reaction by adding 50 µl Buffer F and incubate on ice for 30 minutes;
- Split each sample into two, and purify DNA with the MinElute Reaction Cleanup Kit (Qiagen #28204) (this is done because the manufacturer recommends that only 100 µl of reaction volume be loaded onto each column); at the end, elute the DNA originating from the same initial sample into the same tube in 10 µl EB buffer;
- Possible stopping point: store DNA at -20ºC;

Library preparation

Limited-cycle amplification:

- Prepare PCR samples: add the entire 10 µl purified DNA, 25 µl 2X NEBNext master mix, 2.5 µl each of i5 and i7 Nextera indexes and 10 µl nuclease-free water

- Run PCR under the following conditions: 72ºC for 5 min, 98ºC for 30 sec, followed by 5 cycles of: 98ºC for 10 sec, 63ºC for 30 sec, 72ºC for 1 min;

qPCR to determine number of additional cycles required:

- Prepare samples for qPCR: add 5 µl of PCR product, 5 µl 2X NEBNext master mix, 0.15 µl 100X SYBR Gold Nucleic Acid Gel Stain, 0.5 µl each of the same i5 and i7 Nextera indexes as used for PCR, 3.85 µl nuclease-free water;
- Run qPCR under the following conditions: 98ºC for 30 sec, followed by 30 cycles of: 98ºC for 10 sec, 63ºC for 30 sec, 72ºC for 1 min;
- For each sample, plot fluorescence versus number of cycles; check the number of cycles at which 1/3 of the maximum fluorescence is reached – this is the number of additional cycles required;

Final amplification:

- Use the remainder of the PCR product (45 µl) to conduct PCR under the following conditions, for the required number of cycles: 98ºC for 30 sec, followed by x cycles of: 98ºC for 10 sec, 63ºC for 30 sec, 72ºC for 1 min;

Purification:

Perform double-sided purification of the PCR products, to remove primer dimers and large (>1 kb fragments):

- Add 22.5 µl AMPure XP beads, mix, incubate for 10 minutes, incubate in magnetic rack for 5 minutes, transfer supernatant to new tube;
- Add 58.5 µl AMPure XP beads, mix, incubate for 10 minutes, incubate in magnetic rack for 5 minutes and discard supernatant;
- Wash beads with 200 µl 80% EtOH by pipetting over beads 10 times and discard EtOH; keep tube open in magnetic rack for 10 minutes to ensure all EtOH is removed;
- Resuspend beads in 20 µl nuclease-free water, mix, incubate in magnetic rack for 5 minutes; transfer supernatant to new tube;
- Quantify libraries with Qubit, assess library quality with Bioanalyzer.

3. Results

3.1. Cloning the GFP nanobody-Tn5 gene

The multimodal epigenomic sequencing strategy we propose – NanoTag - requires the use of a GFP nanobody-Tn5 protein fusion. To produce the plasmid encoding this protein fusion, we employed a three-step strategy shown in Figure 9. We extracted the GBN gene from the donor plasmid (pGEX6P1 GFP nanobody) using PCR and restriction digestion. We used the 3X Flag pA-Tn5 FI plasmid, previously used by Kaya-Okur et al.[32] to express a protein A-Tn5 fusion for CUT&Tag, as the recipient plasmid. Once the GBN gene was amplified and extracted, we used it to substitute the protein A sequence in the recipient plasmid (Figure 9).

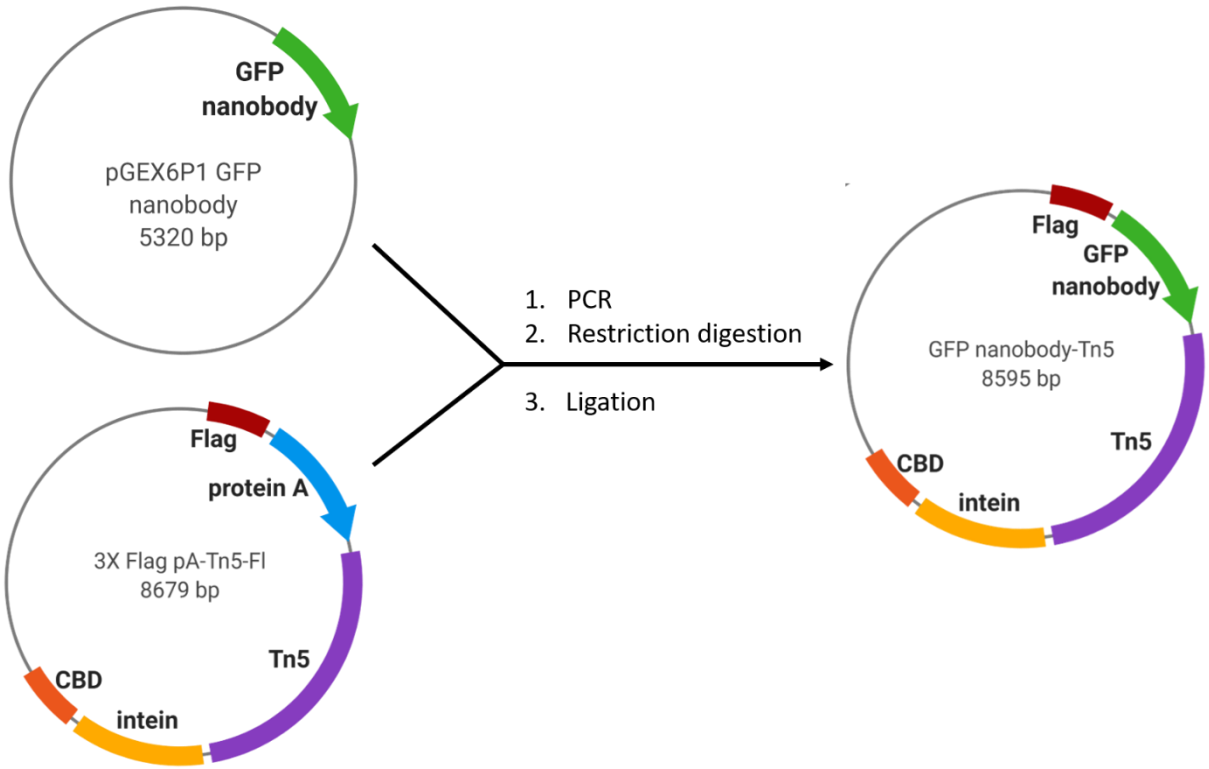


Figure 9 | **Cloning the anti-GFP nanobody – Tn5 (GBN-Tn5) fusion gene**. PCR was conducted to attach the same restriction sites at the ends of the GBN gene that flank the protein A gene within the 3X Flag pA-Tn5 FI (host) plasmid. Restriction digestion of both the host and donor plasmids was followed by purification of the GBN gene and of the digested host plasmid lacking the protein A gene and their ligation, resulting in a plasmid encoding the GBN-Tn5 fusion with an N-terminal Flag tag and a C-terminal intein-CBD tag.

3.1.1. Preparation of the anti-GFP nanobody insert

To insert the GBN gene into the host plasmid, we first attached restriction sites to the ends of the GBN gene that were identical to those flanking the pA gene (*HindIII* at the N-terminal and *EcoRI* at the C-terminal). For this, we amplified the GBN gene using primers that contained sequences complementary to the ends of the gene followed by the *HindIII* and *EcoRI* restriction sites, respectively (Figure 10a).

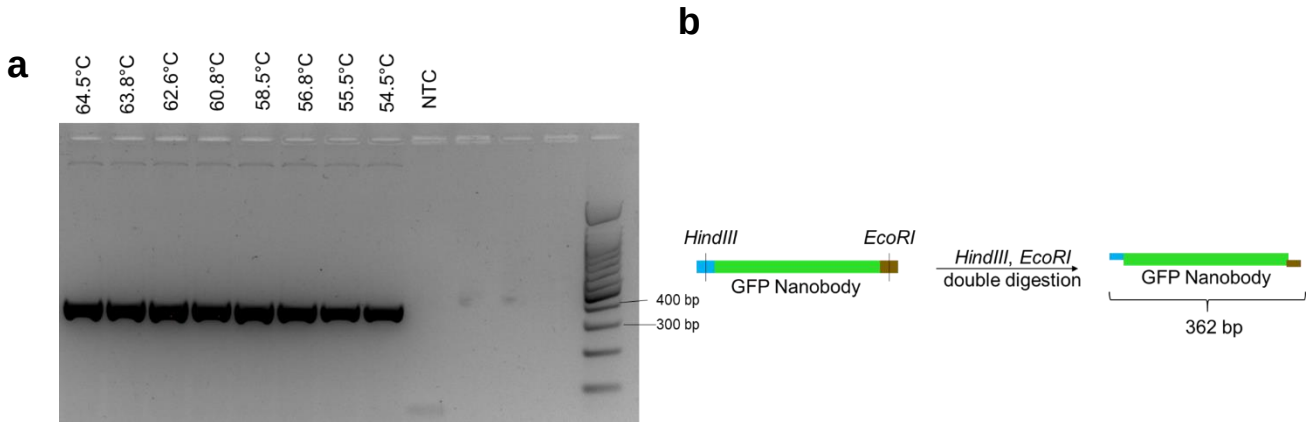


Figure 10 | **PCR to add *HindIII* and *EcoRI* restriction sites to the anti-GFP nanobody gene, followed by double digestion of the PCR product**. We conducted temperature-gradient PCR (left) to attach the proper restriction sites to the ends of the GBP gene. NTC = non-template control (no DNA). We digested the PCR product with the HindIII and EcoRI restriction enzymes (right) to create sticky ends, in preparation for subsequent insertion of the gene into the host plasmid.

All PCR temperature conditions yielded unique bands of the expected size of 362 bp (Figure 10), indicating successful amplification of the GBN gene, flanked by *HindIII* and *EcoRI* restriction sites. We then prepared the GBN gene for insertion by subjecting the PCR product to double digestion with *HindIII* and *EcoRI*. This produced appropriate sticky ends to enable ligation of the GBN gene into the recipient plasmid (Figure 10b).

3.1.2. Preparation of the Tn5 recipient plasmid

In order to remove the protein A sequence and to provide the recipient plasmid with sticky ends, which were necessary for subsequent ligation, we also digested the host plasmid with *HindIII* and *EcoRI* (*Figure 11*a). The resulting linearized plasmid retained the Tn5 gene as well as the sequences encoding its 3X Flag, intein and CBD tags. The double digestion products yielded bands of the expected sizes: digested PCR product - 362 bp; digested host plasmid - 8240 bp. We excised the bands from the agarose gel and purified

them for subsequent ligation. We separated the digestion products by gel electrophoresis (*Figure 11*b), excised the appropriate bands from the agarose gel, purified the DNA and used it for ligation.

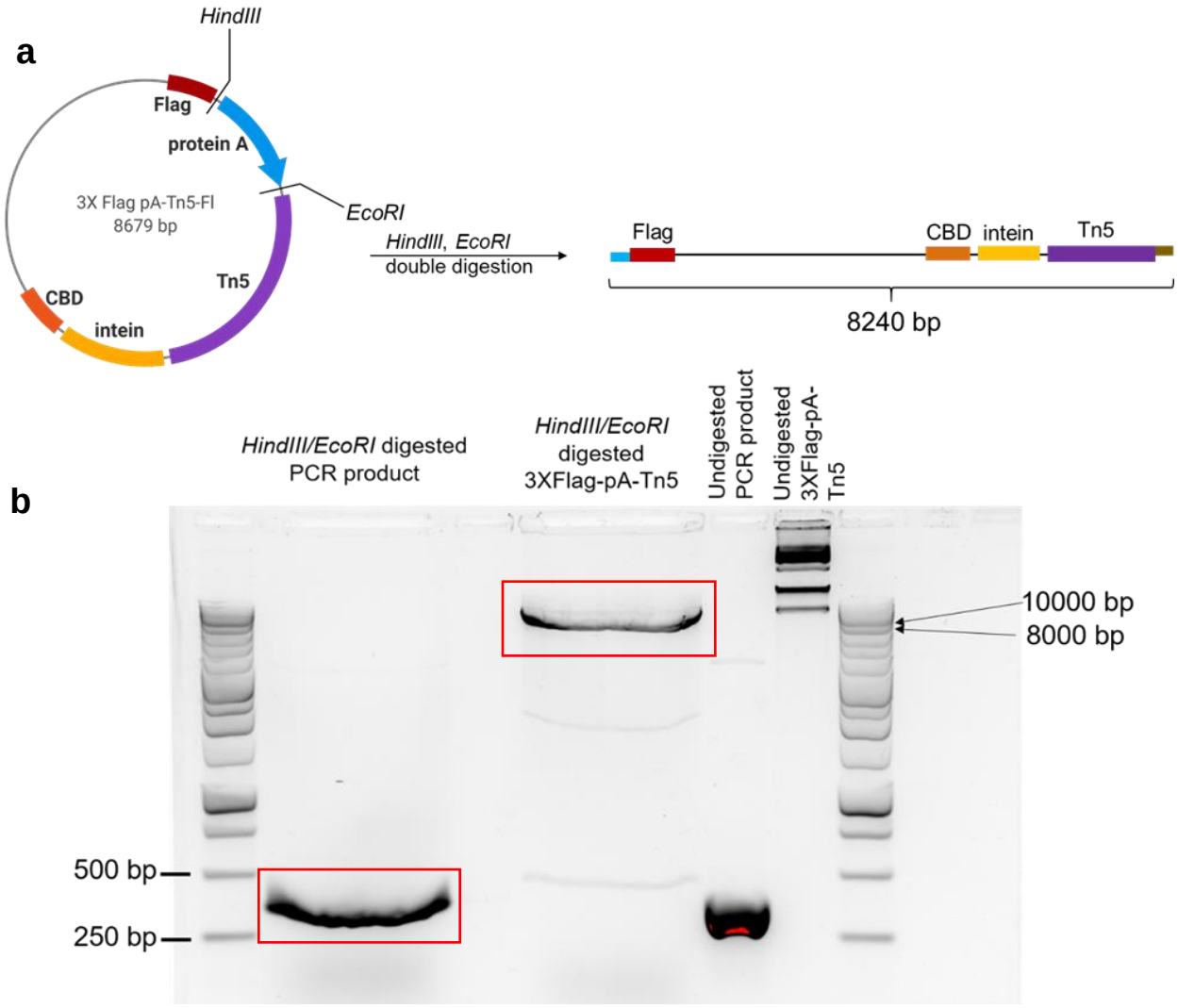


Figure 11 | **Restriction digestion with HindIII and EcoRI of the host plasmid**. Double digestion creates sticky ends and removes the protein A sequence from the linearized host plasmid, in preparation for subsequent ligation. We separated the digestion products by electrophoresis, and gel purified DNA from the bands corresponding to the digested PCR product and the linearized host plasmid lacking protein A (red). We used unusually wide wells in order to be able to separate the entire volume of digestion product (50 μL).

3.1.3. Ligation of the GBN and Tn5 genes

To produce the GBN-Tn5 plasmid, we ligated the GBN insert into the recipient host plasmid. We then used the resulting plasmids to transform Subcloning Efficiency DH5α competent bacterial cells and selected colonies that incorporated the plasmid on Ampicillin-containing LB agar plates. Seven colonies of transformed bacteria resulted. To confirm the selection of colonies carrying a single copy of the GBN-Tn5 plasmid and that the GBN gene became inserted in the correct orientation, DNA from all seven colonies was extracted and sequenced. The resulting sequences were aligned to the expected sequence (the *in silico* designed plasmid sequence). The plasmid from only one colony (C2) was used further. The alignment corresponding to this colony is shown in Figure 12. The sequencing results from all other colonies are shown in the Appendix (Figure 24). The alignment shows a complete match to the expected sequence, and the correct orientation. This shows that the cloning of the GFP nanobody-Tn5 fusion into the host plasmid was successful.

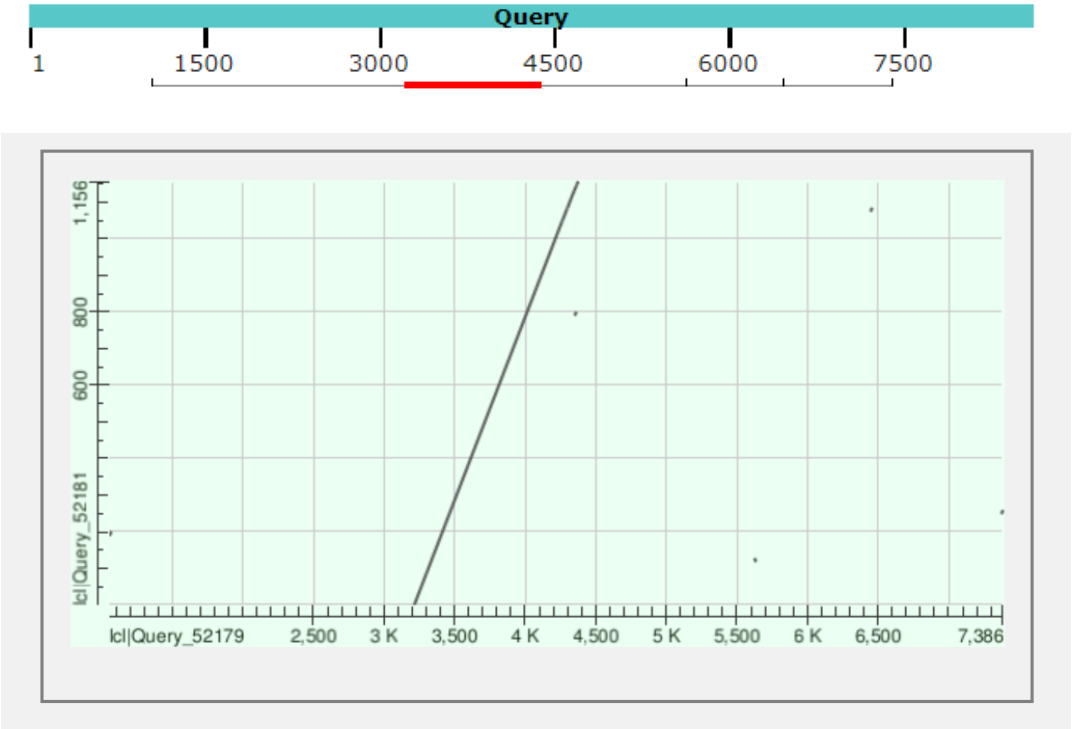


Figure 12 | **Sequence alignment between designed and actual GFP nanobody-Tn5 plasmid.** Graphical summary (top) and dot plot (bottom) of the results of BLAST alignment between the target sequence of the designed GFP nanobody-Tn5 plasmid and the real sequence of the resulting plasmid DNA. The sequenced region covers the Flag, GFP nanobody genes and part of the Tn5 gene.

3.2. Producing the GFP nanobody-Tn5 fusion protein

In order to express and then purify the GFP nanobody-Tn5 fusion protein, we transformed the GBN-Tn5 plasmid into *T7 express* bacteria (see section 2.2 for details). We optimized the bacterial growth and purification processes in order to maximize the protein yield. For example, initially we incubated the bacteria for 5 hours before inducing protein expression, but subsequent trials showed that reducing this time to 3 hours yielded a higher amount of protein. In addition, we initially attempted the bacterial lysis using a Diagenode Bioruptor – an instrument designed for chromatin shearing, but the protein yield was lower than expected. Thus, we concluded this lysis method was likely ineffective and subsequently attempted to use a Branson sonifier for bacterial lysis. This proved to be more effective and increased the protein yield.

At the end of the purification process, we used an aliquot of the pure protein for Western blot to confirm the unique presence of the GBN-Tn5 fusion in our protein sample, based on its molecular weight and on the detection of its Flag tag. While the C-terminal intein-CBD tag was cleaved during purification, the N-terminal Flag tag was retained and we exploited its presence to conduct Western blot using an anti-Flag mouse antibody (Figure 13).

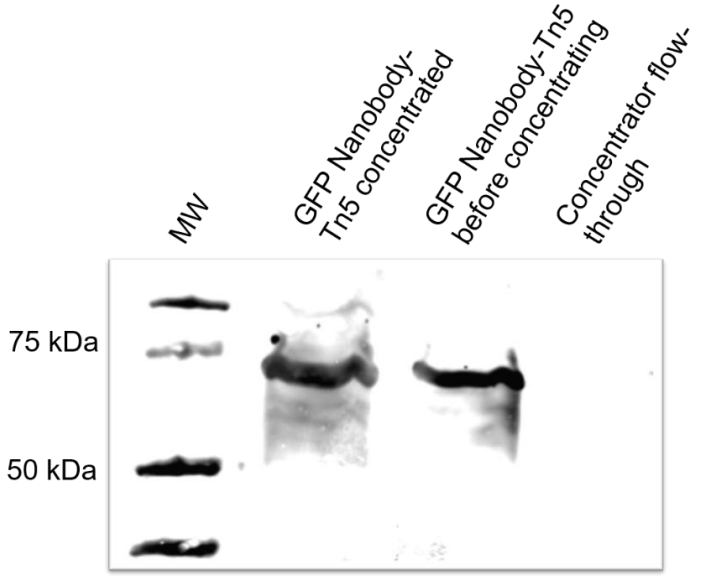


Figure 13 | **Anti-Flag Western blot of pure anti-GFP nanobody-Tn5 before and after concentration.** To achieve the desired concentration of protein for subsequent steps, the protein sample is concentrated using a centrifugal filter (concentrator). The resulting flow-through should be devoid of any protein.

The Western blot showed a unique band at 71.2 kDa, which matched the expected size of the GBN-Tn5 fusion. The Western blot revealed a band of identical size corresponding to the unconcentrated protein sample, but no bands originating from the concentrator flow-through, as expected. We confirmed the successful, high-yield purification of the GBN-Tn5 fusion, based on its display of a unique band of the expected size and on the ability to detect this band using the protein's Flag tag.

3.3. Quantification of GFP Nanobody-Tn5 protein concentration

Accurate quantification of the GBN-Tn5 protein concentration is required for proper annealing with oligonucleotides (see section 2.3).

Picelli et al.[74] described the protocol for in-house purification of Tn5 onto which our purification protocol for GBN-Tn5 was based. To quantify the Tn5 concentration they use NanoDrop, which measures the absorption at 280 nm (A280) of the protein for detection. To convert this value into a concentration, the use of the following formulas in required, which account for the molecular weight and primary sequence of the protein (A280 is determined by the amount of Trp, Tyr residues and disulfide bonds in the protein):

$$(1)\ E_{1\%} = \frac{E_\mu * 10}{MW}$$

$$(2)\ \frac{A_{280}}{E_{1\%}} * 10 = \text{concentration [mg/mL]}$$

where $E_{1\%}$ is the percent solution extinction coefficients and E_μ is the molar extinction coefficient[77]. Thus, the conversion factor is obtained (Table 4).

Table 4 | **Quantification of protein concentration with NanoDrop.** The molar extinction coefficient (dependent on protein sequence) and the molecular weight are used to compute the percent extinction coefficient, which then yields a multiplication factor used to convert A280 into concentration [mg/mL].

Protein	MW	Eμ [$M^{-1}cm^{-1}$]	E1%	Abs280	mg/mL
GBN-Tn5	71235.6	118050	16.57	1	0.603

However, when measuring the absorption of different samples of protein to check for the reliability of the measurement, the values obtained were not as expected. Specifically, samples that were obtained by diluting the initial protein aliquot by the same factor

displayed different concentrations. Thus, we attempted to quantify the protein concentration through other techniques, but we encountered difficulties in achieving proper quantification using a variety of other methods due to some components of the buffer in which the protein is purified (Table 5, section 2.3). The GBN-Tn5 storage buffer contains Triton X-100 (detergent), DTT (reducing agent), EDTA (chelating agent) and glycerol, all of which can interfere with one or more protein quantification methods. Qubit was used by Rykalina et al[67]. to quantify the in-house purified Tn5 samples. However, Qubit is very sensitive to detergent in protein samples and is therefore not suitable for our purposes. The BCA assay is insensitive to detergent but is sensitive to reducing agents, such as DTT, glycerol and EDTA[78]. Using the manufacturer's indication for compatible concentrations of interfering substances, it seemed that by diluting the protein by a large enough factor, all of the interfering substances would be brought to minute concentrations that should not affect the results. However, as seen in Figure 14 and described in more detail below, the results were unreliable.

Table 5 | **Summary of different quantification methods used in the attempt to quantify the GBN-Tn5 concentration.** Each of the quantification methods we attempted presented with drawbacks that made it unreliable for accurate quantification.

Quantification method	Concentration (mg/mL)	Concentration (μM)	Issue
Qubit	2.1	29	Sensitive to detergent
Nanodrop	5.25	74	? Unreliable
BCA Assay	15.7	220	Sensitive to DTT
SDS-PAGE	1.8	25	Sensitive to type of protein

The BCA assay results showed almost identical concentrations for the GBN-Tn5 sample before and after concentration and for the concentrator flow-through. However, the concentrator flow-through does not contain any protein (Figure 13). This makes the BCA assay unreliable for quantification of our protein, likely due to interactions between the interfering substances contained in the GBN-Tn5 storage buffer. The manufacturers mention that the potentially interfering substances were tested independently, but that these may interact if present simultaneously in the protein sample buffer[78]. Dissolving the BSA standards in the GBN-Tn5 buffer failed to correct for this as the colorimetric reaction

yielded almost identical colors for all standards due to the overwhelming effect of the buffer.

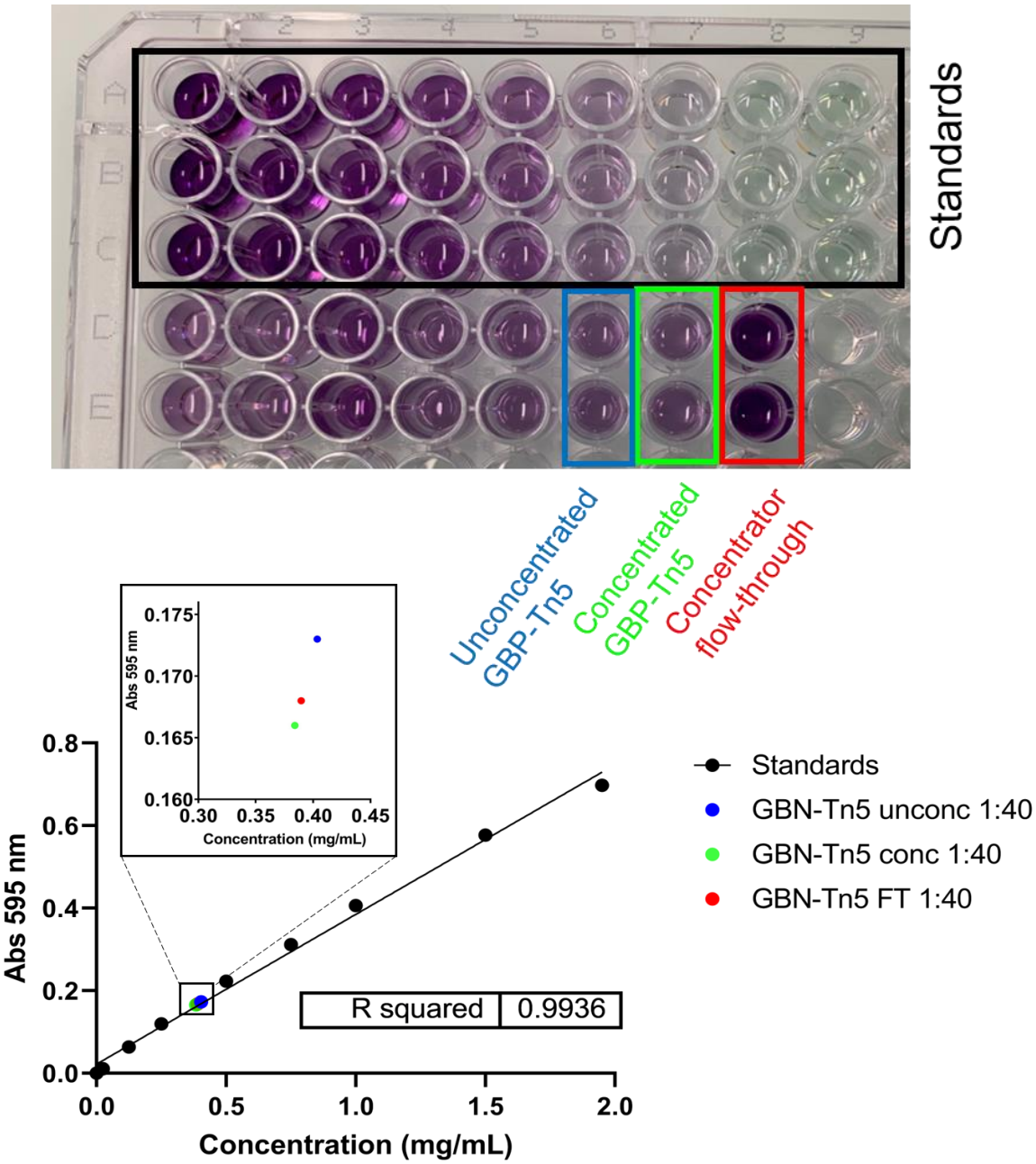


Figure 14 | **Quantification of GFP Nanobody-Tn5 concentration with BCA Assay.** The BCA assay uses a colorimetric reaction to quantify protein concentration (top). The absorption at 595 nm of the reaction product for each protein sample is proportional to the protein concentration. 9 BSA standards are used to establish the linear relationship between the absorption at 595 nm and concentration and this relation is used to extrapolate the concentration of the unknown samples (bottom). Samples were diluted 40 times to minimize the concentration of interfering agents in the buffer.

Given the failure to find a proper quantification assay, we used the results of the tagmentation activity assay (see sections 2.4 and 3.4.1) for different GBN-Tn5 to oligonucleotides annealing ratios to estimate the protein concentration. As the ideal molar ratio between protein and oligonucleotides in the annealing reaction is one to one, the reaction that yielded the best tagmentation activity was assumed to satisfy this requirement. We estimated the GBN-Tn5 concentration based on the known concentration of oligonucleotides that we added to a known protein volume in the reaction that displayed the highest tagmentation activity. However, this method is wasteful and time-consuming and we therefore will attempt other protein quantification strategies in the future (see section 4.1).

3.4. Tagmentation activity of GFP Nanobody-Tn5

To assess the enzymatic activity of the produced GFP Nanobody-Tn5 fusion, we performed two initial tests. First, we verified the tagmentation activity of the GBN-Tn5 on bacterial plasmid, according to the qPCR-based approach Rykalina et al.[67] Second, we checked whether GBN-Tn5 would also tagment cDNA, according to a similar approach by Picelli et al[74], who first described the in-house production of Tn5.

3.4.1. GBN-Tn5 tagmentation of plasmid DNA

Rykalina et al.[67] developed a qPCR-based approach to characterize the tagmentation efficiency of in-house produced Tn5 transposomes, which can be used to correct for variability in efficiency of different protein batches (see section 1.4 for details). Here, we use this procedure as a first way to assess the transposition activity of GBN-Tn5, and to select the appropriate ratio of protein to oligonucleotides used during annealing, in the absence of an accurate method to quantify the protein concentration. Specifically, four different ratios of GBN-Tn5 and sequencing adapters were used to set up four annealing reactions, and the resulting products (transposomes) were used to linearize pUC19 DNA (see section 2.4). The results of this test - the difference in amplification latency between tagmented and untreated DNA, expressed as ΔCp values – for each of the four categories of transposomes (Table 6) were used to select the most efficient batch of transposomes

for further use, and to infer the concentration of the unloaded protein, assumed to be the same as the oligonucleotide concentration used (the mixture was 1:1 v/v).

Table 6 | **Plasmid DNA tagmentation efficiency of GFP Nanobody-Tn5 transposomes.** Four different annealing reactions were set up to load GBN-Tn5 with adaptors, by using different dilutions of enzyme mixed with an equal volume of 25 µM oligonucleotides. For each annealing product, three different tagmentation reactions were set up, using three different volumes of transposomes. Tagmentation products were then used as qPCR templates, and a 610-bp region was amplified. The amplification latency was compared between tagmented and untagmented DNA, yielding the ΔCp values.

Sample	Protein : oligos annealing ratio	Cp	ΔCp
GBN-Tn5 –– 3 µl	Concentrated protein; 25 µM oligos	17.66	9.63
GBN-Tn5 –– 7 µl		19.25	11.22
GBN-Tn5 –– 12 µl		24.76	16.73
GBN-Tn5 –– 3 µl	Concentrated protein, 3X diluted; 25 µM oligos	11.68	3.65
GBN-Tn5 –– 7 µl		13.57	5.54
GBN-Tn5 –– 12 µl		13.52	5.49
GBN-Tn5 –– 3 µl	Concentrated protein, 5X diluted; 25 µM oligos	12.13	4.10
GBN-Tn5 –– 7 µl		13.76	5.73
GBN-Tn5 –– 12 µl		16.46	8.43
GBN-Tn5 –– 3 µl	Unconcentrated protein; 25 µM oligos	11.58	3.55
GBN-Tn5 –– 7 µl		14.89	6.86
GBN-Tn5 –– 12 µl		21.91	13.88
CTRL - no enzyme		8.03	

According to Table 6, the most efficient transposomes (showing the highest ΔCp values, in green) were generated by mixing equal volumes of concentrated protein and 25 µM oligonucleotides, suggesting that the unloaded GBN-Tn5 stock (50% glycerol) used for the reaction also had an approximate concentration of 25 µM. The amplification plots corresponding to the tagmentation products of this transposome batch are shown in Figure 15.

Untreated plasmid DNA was the earliest to amplify (after 5 cycles), as all template molecules were intact. Rykalina et al.[67] did not describe in their paper a positive control for this experiment, in which all of the plasmid would be disrupted and, therefore, mimic 100% tagmented DNA. In our search for such a positive control, we found that a *ScaI* restriction site was located within the pUC19 amplification template. Thus, *ScaI* digested pUC19 should be the latest to amplify, constituting a suitable positive control for this experiment. Indeed, we observed *ScaI*-digested pUC19 to show very late amplification

(after 25 cycles), indicating the successful disruption of most template molecules. Thus, the newly expanded qPCR-based tagmentation efficiency assay provides users with a lower and an upper boundary for the amplification expected for Tn5-treated DNA, facilitating the interpretation of results. GBN-Tn5 tagmented plasmid DNA showed intermediate amplification times, as expected. This effect was concentration dependent, with more GBN-Tn5 leading to more tagmentation and slower amplification. Thus, our first test indicated that GBN-Tn5 was functional, and that its concentration was sufficient to enable efficient transposition of plasmid DNA. This was also a first indication that attaching the GFP nanobody to Tn5 did not interfere with the dimerization or enzymatic activity of Tn5.

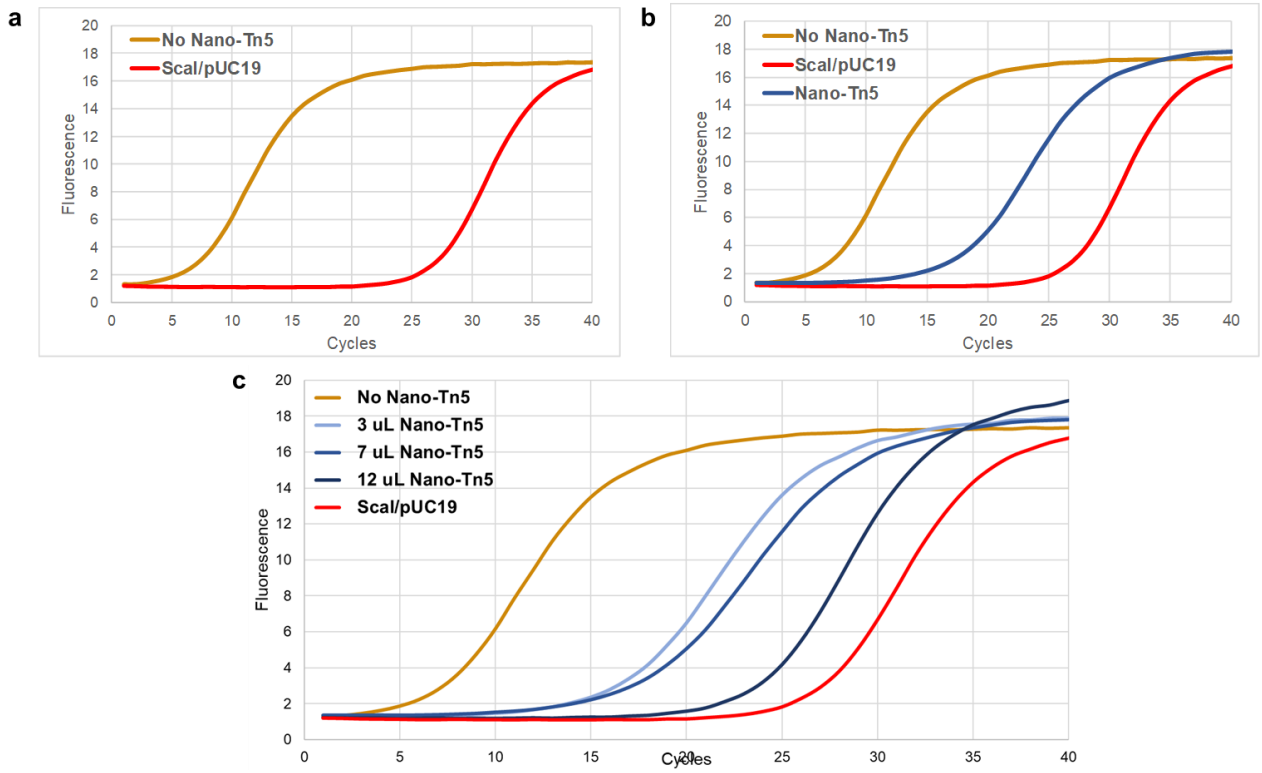


Figure 15 | **Amplification latency of plasmid DNA tagmented by GFP Nanobody-Tn5 transposomes.** Amplification plots showing the amplification stage of untreated EcoRI-digested pUC19 DNA (brown, negative control), of ScaI-digested pUC19 DNA (red, positive control) and of EcoRI-digested pUC19 DNA tagmented by increasing amounts of Nano-Tn5 (blue).

3.4.2. GBN-Tn5 tagmentation of cDNA

To test the ability of the GBN-Tn5 to successfully insert sequencing adapters, we exposed preamplified cDNA (with an average size of 8000 bp) to GBN-Tn5 transposomes,

according to Picelli et al.[74] (Figure 16). The random cleavage of cDNA followed by insertion of sequencing adapters by the GBN-Tn5 would yield cDNA fragments shorter than 8000 bp. Only fragments successfully attached with two sequencing adapters would be amplified during PCR conducted using Illumina indexes.

As shown in Figure 16, GBN-Tn5 successfully tagmented cDNA, leading to fragments with a broad size distribution, of which small fragments (150-600 bp) constituted the vast majority, similarly to what Picelli et al.[74] report for Tn5. However, this result was only obtained in the case of the transposome batch prepared by annealing unconcentrated GBN-Tn5 to 25 μM oligonucleotides and only at this particular concentration. Higher concentrations of GBN-Tn5 transposomes may have over-tagmented the very small amount of cDNA, impeding amplification as the Bioanalyzer failed to detect any cDNA in these cases. Thus, for this test it seems that the amount and concentration of transposomes used is more important than their efficiency, rendering this test less appropriate for quantifying tagmentation efficiency of different transposome batches. To conduct experiments unrelated to NanoTag, we also purified pA-Tn5 in-house, the fusion protein used by Kaya-Okur et al.[32] for CUT&Tag. Similar results were obtained when assessing the cDNA tagmentation activity of different batches of pA-Tn5 transposomes: only the unconcentrated adaptor-annealed pA-Tn5 yielded cDNA fragments and only at a single concentration of those tested, confirming the excessive sensitivity of this test to enzyme concentration.

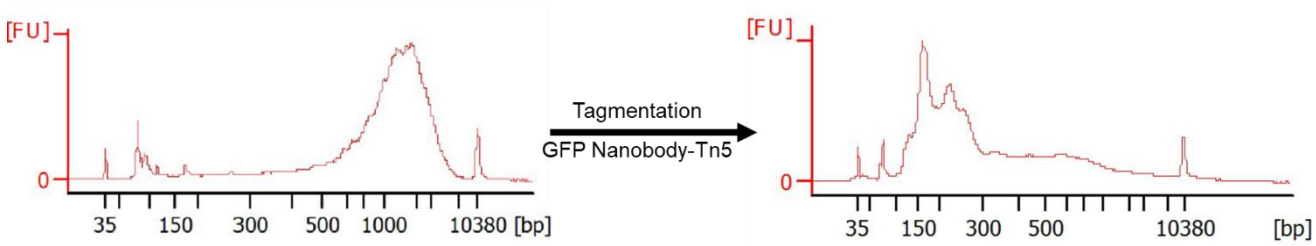


Figure 16 | **GBN-Tn5 transposomes tagment cDNA.** 4 ng of preamplified cDNA (left) was tagmented with 1 µL of transposomes obtained by annealing unconcentrated GBN-Tn5 to 25 µM oligonucleotides. The tagmentation product was amplified, purified and the library assessed on the Bioanalyzer (right). The peaks at 35 and 10380 bp correspond to the internal DNA marker.

To sum up, we confirmed the ability of GBN-Tn5 to cleave plasmid DNA in a concentration-dependent manner using the improved qPCR-based tagmentation assay,

and to insert sequencing adapters to cDNA. Importantly, both tests showed the attachment of the nanobody to not interfere with the enzymatic activity of Tn5.

3.5. Whole genome profiling of mCpG sites with NanoTag

To test the NanoTag approach on an epigenetic *in vitro* model, we collaborated with the group of Tuncay Baubec at the University of Zurich, who designed a series of chromatin reader domains that can recognize and bind specific chromatin features, such as histone PTMs and mCpG sites. They attached engineered chromatin readers (eCRs) to eGFP and stably expressed the fusions in mouse embryonic stem cell (mESC) lines. Conducting NanoTag on these cells would result in GBN-Tn5 binding to the eGFP-eCRs, followed by tagmentation of the cellular DNA by GBN-Tn5. Subsequent library preparation and sequencing would enable the profiling of the epigenetic marks recognized by the eCRs in mESC.

First, we conducted NanoTag on mESCs expressing a methyl binding domain (MBD)-eGFP fusion, which would provide a whole-genome profile of mCpG sites, similar to that obtained by MeDIP-seq. Preparation of NanoTag generated DNA libraries involved, as for ATAC-seq, a limited-cycle amplification of the DNA fragments, followed by qPCR to determine the appropriate amount of additional cycles required (Figure 17). As mentioned in section 1.2.2, the Tn5 transposase acts via cleavage of double-stranded DNA and ligation of sequencing adapters to the ends of the DNA fragments. During subsequent library preparation, the DNA fragments are amplified using primers that anneal to the adaptors attached by Tn5. Thus, the more fragments are generated by Tn5, the faster amplification occurs. Amplification of DNA from MBD-expressing cells (violet, Figure 17) occurred much earlier than that of untagmented DNA (orange, Figure 17) and also earlier than that of DNA from inactive MBD-expressing cells, which suggests tagmentation occurs at a higher rate in the former than in the latter cases. This is an indication that the GFP nanobody-Tn5 is successfully being targeted to eGFP-MBD sites, and not merely cutting randomly at all accessible DNA regions.

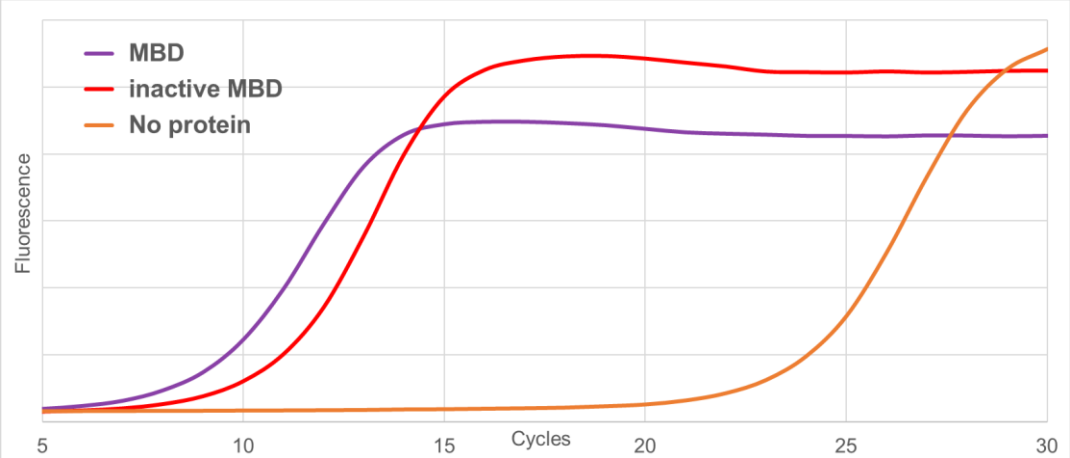


Figure 17 | **Amplification plots of DNA fragments produced by NanoTag from mES cells.** NanoTag was conducted on mES cells expressing MBD and on cells expressing an inactive MBD point mutant (R22A). No protein = cells processed identically but to which no Nano-Tn5 was added.

Next, we assessed the quality of the libraries generated by NanoTag in the same experiment, by DNA fragment analysis using the Bioanalyzer. We chose to focus our analysis of small fragments on the ones ranging from 150 to 600 bp. This particular size range was chosen based on reports by Krasnenko et al.[79] that fragment sizes between 280 and 330 bp (including the adapter sequence) were most efficiently sequenced and mapped. We thus chose a broader size range centered around this interval for our analysis. While DNA treated with no transposase is not fragmented and, therefore, not amplified and not detected, DNA treated with transposase is tagmented, producing fragments with a broad size distribution. There is a four-fold enrichment in small fragments (150-600 bp) in the tagmented DNA of MBD-expressing cells (Figure 18b), which is reduced in the DNA of inactive MBD-expressing cells (Figure 18c) and in that of cells lacking DNA methylation (Figure 18d). This is a further indication that Nano-Tn5 fragments DNA more frequently in cases where it is being tethered to the DNA by the nanobody-GFP interaction. This also suggests that NanoTag is successfully tagmenting mammalian DNA in a targeted manner. It is important to emphasize that Nano-Tn5 also fragments DNA in the control samples, where it is not being targeted to the DNA. This is to be expected, as Tn5 by itself is a hyperactive transposase that cuts accessible DNA randomly and is used to generate libraries in ATAC-sequencing. Thus, background (untargeted) DNA fragmentation by Nano-Tn5 is to be expected. Still, background

tagmentation by pA-Tn5 also occurs in the CUT&Tag approach, without this impeding the interpretation of sequencing results. Given this, we expect that, upon analyzing the sequenced libraries, subtraction of the signal readout for the control samples (eg. inactive MBD) from that of the test sample (eg. MBD) will be sufficient to allow identification of the target areas, as they would present with significant signal enrichment.

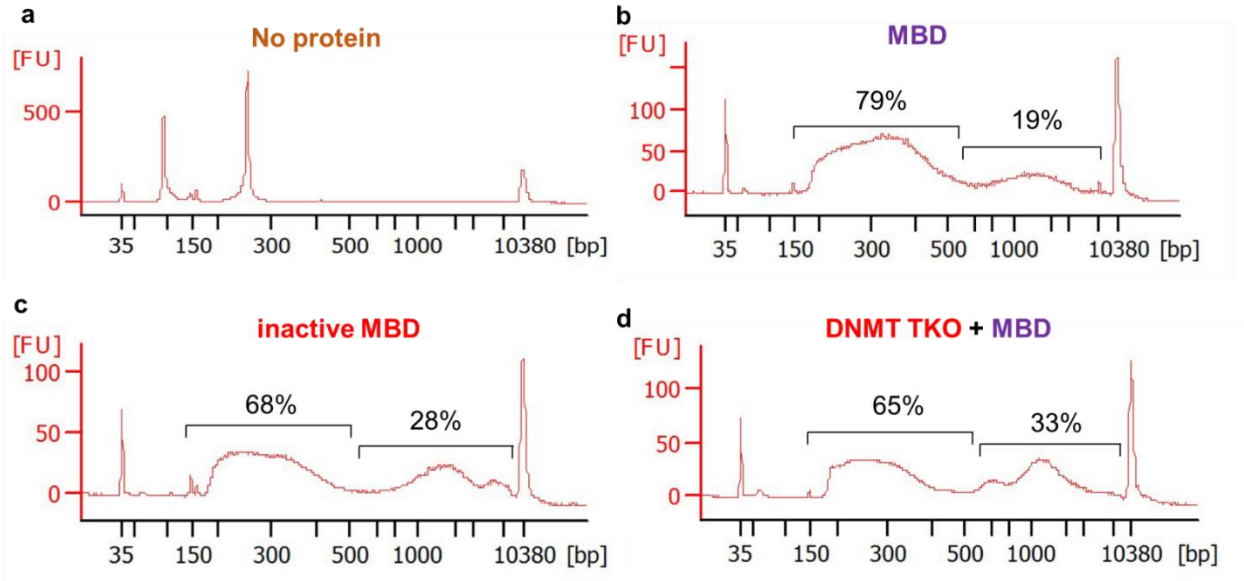


Figure 18 | **DNA fragment size distribution of libraries generated by NanoTag from mES cells expressing MBD and mutant MBD.** DNA libraries were generated from mES cells expressing MBD (a, b) and mutant MBD, respectively (c), and from cells lacking DNA methylation (DNMT triple knock-out cells) expressing MBD by NanoTag, with a 50X dilution of GBN-Tn5. The amounts of fragments ranging from 150 and 600 bp, and from 600 and 10000 bp, respectively, are shown as percentages. Colors match those in figure 14.

3.6. Whole-genome profiling of H3K4me3 sites with NanoTag

To probe another, less abundant, epigenetic mark, we used NanoTag to profile the H3K4me3 sites in the mES cell genome, using the TAF3 chromatin reader, and its corresponding inactive mutant.

Using NanoTag in TAF3-expressing cells, we generated DNA libraries, which showed a very significant enrichment in small (150-600 bp) fragments, indicating successful tagmentation by GBN-Tn5 (Figure 19). This enrichment was less pronounced in the tagmented DNA of mutant TAF3-expressing cells, a result similar to that obtained when profiling the mutant MBD reader. This suggests that, while some untargeted tagmentation

by GBN-Tn5 did occur, it seems that targeted tagmentation was achieved by the nanobody-GFP interaction, enabling profiling of chromatin readers.

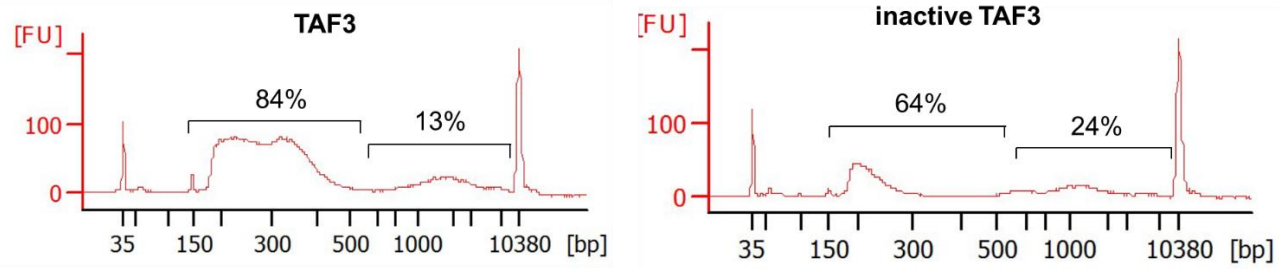


Figure 19 | **DNA fragment size distribution of libraries generated by NanoTag from mES cells expressing TAF3 and mutant TAF3.** NanoTag was conducted on DNA of mES cells expressing TAF 3 and mutant TAF3, respectively, using a 50X dilution of Nano-Tn5. The amounts of fragments in the libraries ranging from 150 and 500 bp, and from 500 and 10000 bp, respectively, are shown as percentages.

ATAC-seq protocols advise optimizing the tagmentation reaction by attempting several different concentrations of Tn5. In a similar manner, we prepared each NanoTag sample using a different concentration of Nano-Tn5, to identify the one yielding the highest quality of libraries. Comparison between different samples was always done on samples containing the same amount of transposomes from the same batch. However, different amounts of transposomes did not yield libraries of similar quality. A 50-fold dilution of Nano-Tn5 appeared to be most successful in consistently generating libraries of good quality and thus, we will use this concentration further in NanoTag experiments.

The most reliable proof that NanoTag successfully enables targeted epigenomic profiling will be provided by performing next generation sequencing of the libraries we generated.

4. Discussion

4.1. Production of GBN-Tn5 and proof-of-concept NanoTag experiment

Here we presented the successful cloning and production of a novel protein fusion, GBN-Tn5, and showed its enzymatic activity in tagmenting plasmid DNA, cDNA and mammalian DNA. Furthermore, we used GBN-Tn5 in a first, proof-of-concept, NanoTag experiment with promising results that suggest NanoTag to be a functional method for epigenomic profiling in mammalian cells.

Fusions of nanobodies to various enzymes have been described. For example, fusions of nanobodies to alkaline phosphatase[80], to horseradish peroxidase[81] or to nanoluciferase[82] have been successfully used for the development of novel immunoassays, and fusion of a nanobody to O-Linked N-acetylglucosamine transferase has been produced and applied for targeted protein glycosylation[83]. However, GBN-Tn5 is, to our knowledge, the first fusion between a nanobody and a transposase that has been attempted, and its success proves that fusing a nanobody to Tn5 does not interfere with the enzymatic activity of the transposase. The molecular cloning procedure for the GBN-Tn5 gene was straight-forward and fast. We encountered no major difficulties at any of the steps despite our limited experience with molecular cloning and we successfully produced the GBN-Tn5 plasmid after only two weeks of work. One limitation of our approach is that the GBN-Tn5 gene was only sequenced partially, such that the known sequence covered the GBN and a part of the Tn5 gene. To be completely certain that the entire sequence is correct, we could have sequenced the plasmid also in the opposite direction. However, given the confirmed enzymatic ability of the GBN-Tn5, any mutations in the Tn5 sequence are highly unlikely. Given that our cloning strategy was successful, feasible and fast, the prospect of designing and producing a large variety of nanobody-Tn5 fusions becomes readily achievable. The bottleneck of this process will, thus, be not the cloning of new fusions but the potential design and production of novel antibody mimetics.

The process of purifying the GBN-Tn5 protein was also successful and highly efficient. While the quantification of GBN-Tn5 concentration proved problematic (see below), we

estimate that from each 1L of bacterial culture, we obtained approximately 2.5-3 mg of GBN-Tn5 (1.5 ml stock at 25 μM), which is sufficient for approximately 300 samples in the NanoTag experiment, according to the dilution that, so far, has yielded the best results. The GBN-Tn5 purification strategy we employed involves low costs and is technically accessible and straight-forward. The purification of GBN-Tn5 does not require expensive reagents or automated protein purification systems, using instead only a chromatography column. While our protein purification setup (see Figure 8) involved a low-cost automated pump that ensured continuous flow through the column, this could also be done manually. Despite the fact that our research group did not have prior experience with protein purification, we encountered no problems in following this purification protocol, which was originally described by Picelli et al.[74] The high yield and low cost facilitate large-scale production of GBN-Tn5 and other Tn5 variants. In addition, one batch of GBN-Tn5 is sufficient for approximately 300 NanoTag samples. Thus, the entire volume of purified protein could be used to produce a single batch of transposomes. This transposome stock could be used for dozens of NanoTag experiments, without concern for batch-to-batch variability. What constitutes a bottleneck in the production of GBN-Tn5 transposomes is the moderately high cost of the commercially available sequencing adapters, which could become a considerable expense in the case of large-scale production.

Efficient lysis of the bacterial cells is a yield-determining step in the purification of proteins[84]. Many bacterial lysis methods exist, ranging from mechanical (e.g. using a French press), osmotic, thermal and acoustic (e.g. using sonication) to chemical and enzymatic approaches[85], so choosing the best lysis method can be a challenge. In our case, optimization of bacterial lysis was also required. Initially, using the Bioruptor device for bacterial lysis proved ineffective, so we instead employed an ultrasonication device, which significantly improved the subsequent protein purification yield. Finally, while expressing GBN-Tn5 with an intein-CBD tag and using a chitin resin for its purification proved efficient, different strategies may be employed to produce GBN-Tn5. For example, in-house purification of Tn5 using Chitin Magnetic beads[67] precludes the use of a column and shortens the purification time but is not readily amenable to large-scale purification. Alternatively, Tn5 can also be produced using a His tag purification strategy, with

reportedly higher yields[86]. However, we decided to follow this particular existing protein purification approach to simplify our cloning strategy, since the plasmid encoding a fusion between protein A and Tn5 had already been described, which also contained the Flag and intein-CBD tags. In addition, this approach allows the expression and purification of GBN-Tn5 expression, as well as its use for transposome assembly and its testing within two weeks. Thus, while one may envision purification strategies with higher yields, our strategy resulted in sufficient GBN-Tn5 amounts for many experiments, while keeping technical requirements and costs at a minimum. Finally, large-scale production of GBN-Tn5 is achievable and may be outsourced to external contractors.

Reliable quantification of GBN-Tn5 is required for its annealing to sequencing adapters in order to produce high efficiency transposomes. While an excess of adapters does not interfere with the activity of GBN-Tn5, it is wasteful, as any unannealed oligonucleotides are simply discarded during library processing. As already mentioned, the outsourced oligonucleotide synthesis represents the financial bottleneck of GBN-Tn5 transposome production. As reported in section 3.3, we tested various protein quantification methods. We found commonly used quantification methods to yield unreliable results, mostly due to the interference of compounds present in the buffer used for purification and storage of the GBN-Tn5. The most widely used, and one of the simplest, protein quantification methods is to measure the absorption at 280 nm of the protein sample, using a UV-Vis spectrophotometer, in combination with quartz cuvettes, which do not absorb UV light. We performed these measurements using a NanoDrop device, which provides similar quantification capabilities but is designed for small volume measurements of 1 μl. Converting the absorption at 280 nm to concentration of a protein sample involves taking into account the molar extinction coefficient of the protein of interest[77] (see section 3.3), which can be estimated using the ProtParam tool from ExPASy[87], based on the amino acid sequence of the protein. However, accurate determination of the molar extinction coefficient should account for additional factors, such as the chemical environment the amino acids are in[88]. Thus, the molar extinction coefficient of GBN-Tn5 may have been improperly estimated, leading to inaccurate protein concentration estimates.

To address this quantification issue, we used the qPCR-based approach described by Rykalina et al.[67] to test the tagmentation efficiency of several transposome batches, prepared using different concentrations of GBN-Tn5, as described in section 3.4.1. This approach was successful in allowing us to choose the transposome batch with the highest efficiency for conducting NanoTag, and we used this as a guide for preparing new transposome batches from the same GBN-Tn5 stock. However, this method is time-consuming and inefficient. A potential solution is protein quantification by mass spectrometry, a very powerful method that can accurately quantify a target protein even in complex mixtures. Another potential solution is to store an aliquot of GBN-Tn5 in buffer lacking detergent. This would allow accurate quantification by fluorometry, using the Qubit device, which is highly sensitive to detergents but not to other possibly interfering substances. This can be achieved either by removing the detergent (Triton X-100) at the end of the purification procedure, using Bio-Beads (supplied by Bio-Rad), or by excluding detergents throughout the entire purification procedure[86]. We expect that one of these methods will provide an accurate and reliable strategy for the quantification of GBN-Tn5. Still, due to the high-yield production of GBN-Tn5, it is possible to conduct many NanoTag experiments without concern for batch effects. A unique batch of transposomes could simply be prepared from the entire stock of protein and used for many NanoTag experiments.

We improved the qPCR-based tagmentation efficiency assay described by Rykalina et al.[67] by including a positive control mimicking fully tagmented DNA. We generated the positive control by digesting the amplification template DNA with a restriction enzyme (*ScaI*) whose restriction site is located within the amplification region. This advances the assay by defining minimum and maximum boundaries for the amplification latency of a tagmented DNA sample. We used this newly advanced strategy to assess the tagmentation ability of GBN-Tn5. In our tests, GBN-Tn5 tagmented DNA showed increasingly delayed amplification with increasing amounts of GBN-Tn5, consistent with the results of Rykalina et al.[67] In addition, all GBN-tagmented DNA samples showed amplification latencies intermediate between those of untagmented DNA and *ScaI*-digested DNA, rendering the newly improved qPCR-based tagmentation assay even more useful than its original version.

We presented the results of the NanoTag experiments on cells expressing eCRs in sections 3.5 and 3.6. In qPCR, libraries generated from cells expressing inactive eCR showed later amplification than libraries from cells expressing the active eCR. Upon DNA fragment analysis using the Bioanalyzer, we found the DNA library fragments to be broadly distributed in size, with a strong enrichment for small fragments. This enrichment was less pronounced in libraries from cells expressing the inactive eCR. Thus, comparisons between the libraries generated from cells expressing active eCRs versus inactive eCRs indicated that the nanobody-eGFP interaction was successful in tethering the nanobody-Tn5 fusion to the loci in the genome at which the eCR was bound. These preliminary results suggest NanoTag to be a functional targeted epigenomic profiling strategy.

The limitation of the results described so far is the lack of genomic information. The preliminary results do not provide information about the locations in the genome where the GBN-Tn5 cut DNA and inserted adapters. We will sequence the NanoTag DNA libraries to address this limitation and provide definite proof that GBN-Tn5 is indeed localized at target loci in the genome. First, we will repeat the NanoTag experiments to obtain at least three replicates for each eCR. Then, we will simultaneously sequence libraries generated from active eCR-expressing mES cells, inactive eCR-expressing mES. To analyze the sequencing results, we computationally subtract the signal of the inactive eCR-expressing cells from that of the active eCR-expressing cells. This will correct for unspecific (untargeted) transposition, in a similar manner to the subtraction of input signal in ChIP-seq[89] (*Figure 20*). The remaining signal would originate from transposition sites and map to regions in the genome where the eCRs were bound. Simple bioinformatic analysis will suffice for the identification of these enrichment regions: we will align the reads to the mouse reference genome, visualize the enrichment regions with a genome data viewer and check whether the enrichment peaks match with the expected occurrence of the specific epigenetic mark. We will compare our results against ChIP-seq results for each eCR reported by Villaseñor et al.[68], but also to ChIP-seq data sets available in the literature for mCpGs[90] and H3K4me3[91], respectively. For example, we expect H3K4me3 to be greatly enriched at gene TSS[29]. Ultimately, we should obtain whole-genome profiles of mCpGs and H3K4me3 sites in mESCs if NanoTag was

successful. Importantly, these profiles will likely deviate from previously reported ChIP-seq profiles because NanoTag would identify the target binding sites restricted to regions of accessible chromatin.

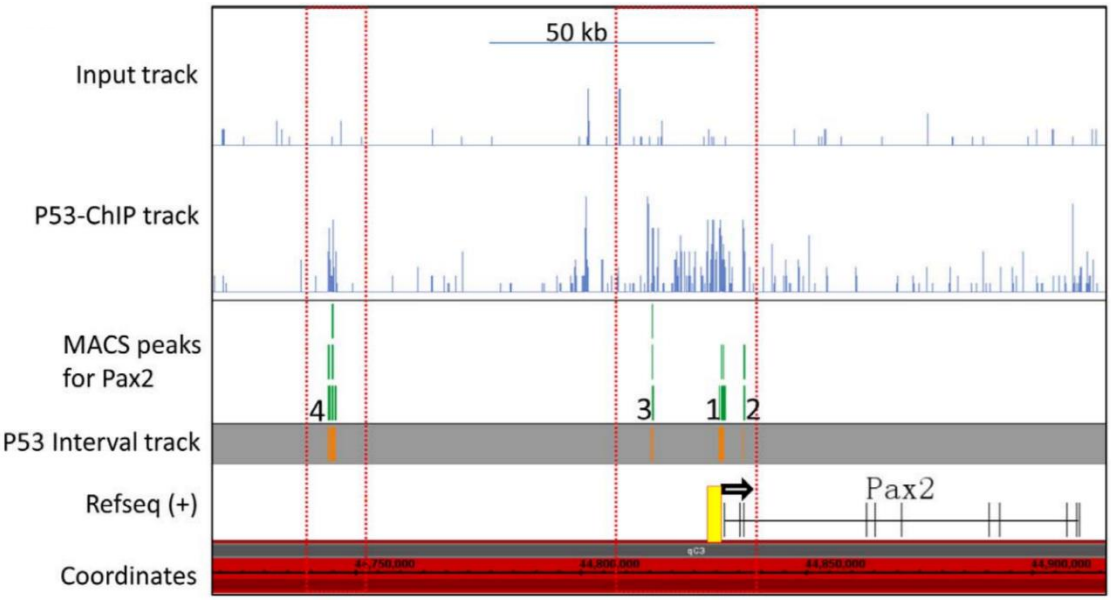


Figure 20 **| Illustration of background subtraction to show enriched regions of a ChIP target**. In ChIP-seq, the input constitutes the negative control (Input track), whose signal is subtracted from that of the test sample (P53-ChIP track) to correct for any experimental bias. In ChIP-seq, this would mainly result from bias in DNA shearing and unspecific antibody binding. The difference between the first two tracks constitutes the actual presence of P53 (P53 Interval track). Reproduced from Saifudeen et al. (2012)[92].

We will also conduct NanoTag on mESCs expressing other eCRs, including chromodomain-containing CBX7/dPC (against H3K27m3) and CBX1 (against H3K9me3), as well as a bivalent reader recognizing H3K27m3 and H3K4me3[68].

We envisioned NanoTag as a technique that draws on advantages of several techniques - Omni-ATAC-seq (and its variation, Methyl-ATAC-seq) and CUT&Tag – to ultimately and uniquely enable simultaneous multimodal epigenomic profiling without the use of antibodies. Thus, NanoTag overcomes existing antibody limitations, such as specificity, batch-to-batch variation and permeability, that hamper both ChIP-seq and CUT&Tag, while simultaneously providing more information than either method. In NanoTag, the GBN-Tn5 is targeted *in situ* to loci in the genome at which the GFP-tagged target is located. Once bound, GBN-Tn5 cleaves the DNA at the specific sites where it was targeted as long as these sites are within a region of accessible chromatin. Thus,

NanoTag provides a genome-wide profile of target binding sites along with information about chromatin accessibility and DNA methylation characterizing these sites. The use of GFP-tagged proteins is very frequently employed in biological research. Importantly, many GFP-tagged DNA-associated proteins have been described. For example, the modENCODE and modERN projects created databases of transcription factor binding sites, using transgenic lines expressing transcription factor-GFP fusions[93,94]. They have generated 403 transgenic strains of *C. elegans* and 427 strains of *Drosophila* expressing fusions between GFP and transcription factors[94]. Thus, using a GFP-binding nanobody provides NanoTag with increased versatility, enabling the profiling of a wide range of targets without any adjustments to the described protocol.

4.2. Potential NanoTag applications

The versatility and modularity of NanoTag open the possibility for a myriad of other experiments and applications in which variations of NanoTag would result in new epigenomics approaches. Here, we discuss several potential applications of NanoTag that could provide novel and unprecedented insight into the epigenome.

Figure 21 presents a first conceptual overview of envisioned NanoTag-derived applications.

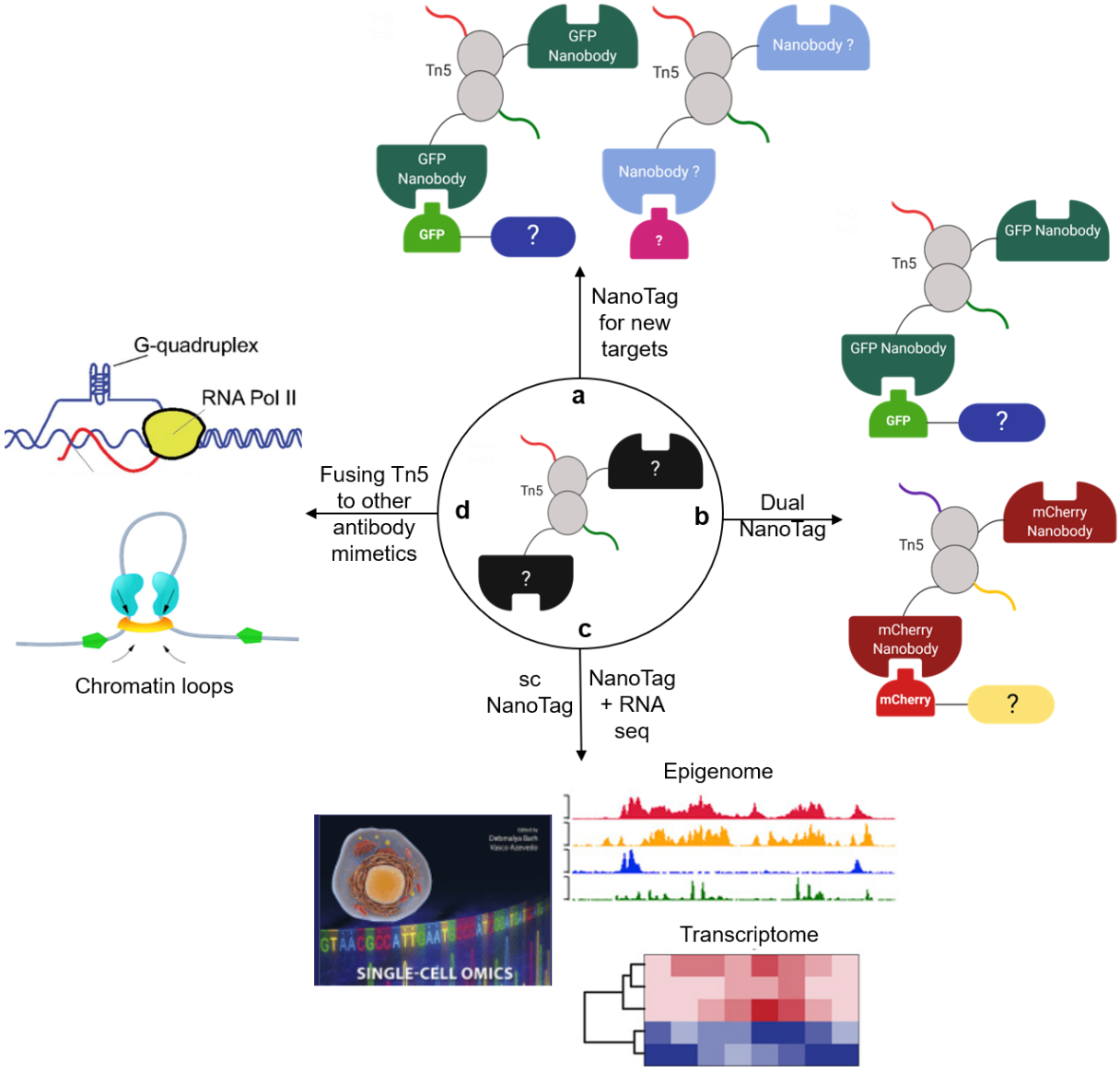


Figure 21 | **Envisioned NanoTag applications. (a)** NanoTag profiles different fluorescent targets, such as other chromatin reader domains, or native proteins directly using fusions of Tn5 to a selective nanobody. **(b)** NanoTag profiles two targets simultaneously by simultaneous treatment of DNA with two types of transposomes: GFP nanobody-Tn5, and mCherry nanobody-Tn5 transposomes. **(c)** RNA is extracted concomitantly with conducting NanoTag to combine epigenomic profiling with transcriptomic analysis. This approach is also applicable at the single cell level. **(d)** Fusing Tn5 to other antibody mimetics, such as DARPins should enable direct profiling of other chromatin features, such as G-quandruplexes, or chromatin loops.

4.2.1. NanoTag for other targets

NanoTag could be used to profile other chromatin-associated GFP-tagged targets by using the same protein fusion, or to profile proteins tagged with other labels, by using

different nanobodies. A large variety of nanobodies against different tags have been produced, including mCherry[95], Spot tag[96] and ALFA tag[97]. In addition, DARPins against the same or other tags, such as the mTFP1 fluorophore[98], have also been developed. Based on the experience of producing GBN-Tn5, testing it on plasmid DNA and cDNA and applying it for NanoTag, we estimate that generating a new antibody mimetic-Tn5 fusion that is ready for NanoTag can be achieved within a month. This makes the prospect of generating a series of other fusions and using them to profile new targets readily achievable. More generally, a tag-free target could be profiled by using a fusion to an antibody mimetic specifically designed for that target (*Figure 21*a).

4.2.2. Dual NanoTag

Dual NanoTag (*Figure 21*b) would enable the whole-genome profile for two distinct targets simultaneously, facilitating the study of interactions between two chromatin-associated proteins as well as their potential effect on DNA methylation and chromatin accessibility. To exemplify how this approach may be conducted, we describe a potential application within the already described eCR system. To conduct this experiment, we would first produce a second type of nanobody-Tn5 fusion, for example using an mCherry-binding nanobody, and anneal it with a distinct set of adapters. In addition, this experiment requires a model system in which the two target proteins are expressed as fusions to GFP and mCherry, respectively. For this, we would use the mammalian expression vector used by Villaseñor et al. for expressing their eCRs[68]. We will replace the eGFP tag of an eCR with an mCherry tag and transfect this vector into cells that already express one of the eGFP-eCR fusions. Ultimately, this will result in a mES cell line that expresses one eGFP-tagged eCR (e.g. the H3K4me3 reader) and one mCherry-tagged, eCR (e.g. the H3K27me3 reader). Once the cells are prepared, we will subject them to NanoTag and simply apply a mixture of the two transposome types containing different indexes during the transposition step.

Additionally, given the availability of antibody mimetics against a wide range of different tags, theoretically even more than two targets could be profiled simultaneously. The adapters can be readily expanded for barcoding, providing each type of Tn5 fusion with a clear identifier for bioinformatic deconvolution. Assuming there would be a transgenic

model system in which different targets are expressed as fusions to different tags, using multiple Tn5 fusions to nanobodies/DARPins against the same tags would enable the simultaneous profiling of all targets in the same experiment. This may, of course, prove to be technically cumbersome but may also provide a unique opportunity to study epigenomic network interactions.

4.2.3. Single-cell NanoTag for multi-omic approaches

Combining NanoTag with RNA extraction from the same cells, followed by RNA sequencing (*Figure 21*c) would enable a combined epigenomic and transcriptomic profiling approach. Reyes et al. showed that ATAC-seq and RNA-seq can be combined for simultaneous profiling of chromatin accessibility and gene expression in the same cells[99]. We could employ a similar strategy to combine NanoTag and RNA-seq. Specifically, cells could first be fixed with DSP, which preserves mRNA, then subjected to the usual NanoTag protocol, followed by mRNA capture using oligo-dT beads[99]. DNA libraries would then be generated as usual, while the smart-seq2 protocol[75] could be used to generate the RNA-seq libraries. Such a combined epigenomic and transcriptomic profiling approach could be used to study the effect of epigenetic factors and their interaction on gene expression.

The NanoTag protocol was derived from the Omni-ATAC-seq[76] and from the CUT&Tag protocol[32], two approaches applicable at the single-cell level. Given this, we expect that a single-cell NanoTag strategy (scNanoTag) would also be feasible. To test scNanoTag, we would sort individual cells of a target population into a microwell-plate by FACS[100]. Then, we would apply the NanoTag protocol to all microwells simultaneously. This is a low-cost method that would enable a proof-of-principle scNanoTag experiment. Next, we would attempt an ultra-high-throughput scNanoTag through droplet microfluidic partitioning and barcoding of individual cells using commercially available platforms such as ICELL8 from Takara Bio or Chromium from 10X Genomics[101]. scNanoTag would allow for the investigation of multiple epigenetic features within the same DNA molecule simultaneously. This would provide insight into the functional roles of the epigenomic network in defining unique cell identities and in allowing for heterogeneity within a defined cell population. In addition, scNanoTag could enable the characterization of functional

consequences of epigenomic heterogeneity and contribute to ongoing efforts to compile epigenomic atlases of every cell type within a tissue.

4.2.4. Structural chromatin interrogation using NanoTag

NanoTag could also be modified so as to enable the investigation of structural chromatin features (*Figure 21*d).

4.2.4.1. Probing DNA G-quandruplexes using NanoTag

G-quadruplexes (G4) are a DNA conformation that occurs primarily in telomeres, but also in other functional regions of the genome. The frequency and non-random distribution of G4s in the genome indicates an important biological function, even though causal evidence is sparse[102]. For example, G4s have been implicated in gene regulation and cancer progression[102,103]. While the physical and chemical properties of basket-type and propeller-type quandruplex structures have been extensively investigated, their sites of occurrence and function in the genome have remained enigmatic[102]. A sequencing-based approach for probing G4s, G4-Seq, has been recently developed[104]. However, in G4-Seq, the DNA is sequenced in conditions that promote the formation of G4 structures, which does not reflect the molecular crowding within the nucleus of a cell[104]. Thus, this technique does not provide a way to probe *in situ* existing G4 structures as they occur in the cell, but merely reveals the sites in the genome at which G4 conformations can potentially occur. In addition, this method uses a comparative sequencing approach, which requires that the DNA molecule be sequenced twice, rendering this approach expensive.

The group of Andreas Plückthun have pioneered the field of designed ankyrin repeat binding proteins (DARPins) and they have engineered DARPins that distinguish between different types of quadruplex DNA[105]. We envision using a G4-DARPin-Tn5 fusion to probe G-quadruplexes (G4) and their potential interactions with other epigenetic marks. In contrast to G4-Seq, our approach would, for the first time, provide a genome-wide map of structurally defined G-quadruplexes in their native state in the cell. In addition, this approach may also be scalable to the single-cell level, which would provide important insights into the heterogeneity of G-quadruplexes. To our knowledge, whole-genome G4 profiling in single cells has so far not been reported.

4.2.4.2. Probing chromatin loops using NanoTag

Another structural chromatin feature that could be investigated by NanoTag is chromatin loops. Cohesin is a protein complex that was shown to be responsible for the formation of chromatin loops, which are involved in the replication, transcription and repair of DNA[106]. While Hi-C and related methods are currently used to probe chromatin loops in cells, the epigenetic information gained through such techniques is limited to interactions between genomic loci. In contrast, a NanoTag approach would enable the simultaneous profiling of cohesins but also of the surrounding chromatin accessibility and DNA methylation. This would require the design of a cohesin-recognizing DARPin, followed by its fusion to Tn5. In this manner, NanoTag could be used to profile chromatin loops and their interaction to other epigenetic features under different cellular contexts.

4.3. NanoTag under temporal control

Very recently, nanobodies were engineered to bind to their cognate antigen in a chemogenetically controlled manner (LAMAs, *Figure 22*)[107]. Using a GFP binding LAMA-Tn5 fusion to conduct NanoTag could improve the strategy by allowing the cells under study to constitute their own control. Thus, for example, instead of comparing the sequencing signal of DNA from active eCR-expressing cells to that of DNA from inactive eCR-expressing cells, we could instead compare the signal of DNA from cells treated with the on switched LAMA-Tn5 to DNA from the same cells but in which the LAMA-Tn5 is not switched on chemogenetically. In addition, one could imagine combining the use of LAMA-Tn5 with the Dual NanoTag approach. Thus, both types of transposomes could be applied to cells simultaneously but the chemogenetic control over the LAMAs would allow temporal control over the activity of each one, individually. In this way, by comparing cells in which both transposome types are switched on simultaneously with cells in which they are switched on at different times, one could distinguish between cases where the two targets coexist in the same cell and cases where they actually occur in different cells, without the requirement for single-cell experiments.

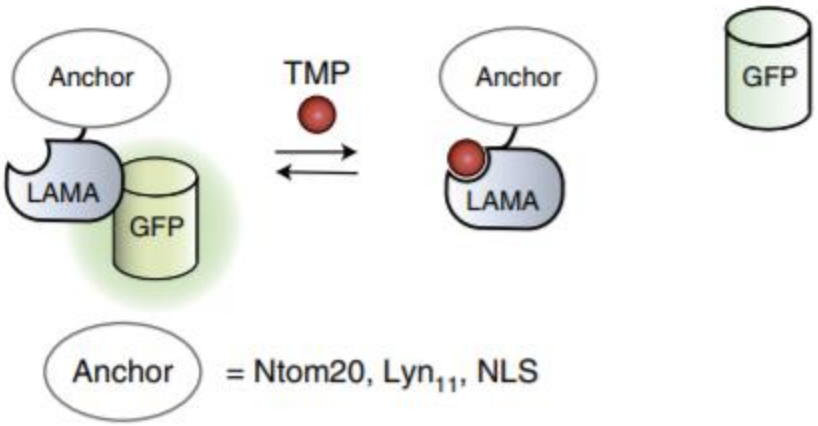


Figure 22 | **Chemogenetic control of a GFP-binding nanobody using a small molecule antibacterial drug (TMP)**. LAMAs can be expressed intracellularly and localized to specific cellular compartments by attaching an Anchor, such as a nuclear localization signal (NLS) Reproduced from Farrants et al. (2020)[107].

4.4. Alternatives to genetically-encoded protein fusions in NanoTag

An alternative way of obtaining the antibody mimetic-Tn5 fusions used for NanoTag could be using Click chemistry. This approach could enhance the modularity of NanoTag by allowing new fusions to be created without the need for molecular cloning and purification of each one separately. While here we described the generation of the GFP nanobody-Tn5 through genetic fusion, an alternative approach could be through chemical conjugation. Hudak et al. described a Copper-free click chemistry approach suitable for protein-protein conjugation, which requires that one protein be genetically tagged with an aldehyde tag, and subsequently covalently attached to a second DIBAC-functionalized protein (Figure 23)[108].

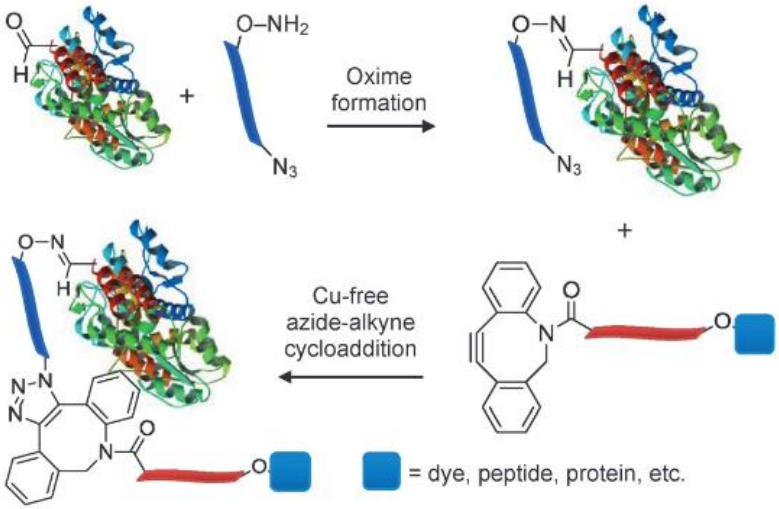


Figure 23 | **Cu-free "Click" chemistry can be used to derive bifunctional protein fusions**. This approach relies on orthogonal triazole and oxime chemistry to enable site-specific conjugation of proteins. An fGly linker (blue band) and aldehyde tag are attached to one protein (ribbon structure). The other protein (blue square) is DIBAC-functionalized (red band). Cu-free click chemistry can then be performed for covalent attachment of the two proteins. Reproduced from Hudak et al. (2012)[108]

Instead, aldehyde-tagged Tn5 could be expressed and purified, and used for each new fusion to a nanobody or DARPin that is produced directly in the DIBAC-attached form. This approach may streamline production of a wide range of distinct Tn5-antibody mimetic fusions.

4.5. NanoTag *in vivo*

The next step to prove the feasibility of NanoTag will be testing our approach *in vivo*. We are currently searching for potential collaborators who may host in their lab a transgenic animal line that expresses a fluorescently tagged chromatin-associated protein (e.g. a transcription factor). Once a suitable animal model is found, we would enrich for cells that express the tagged target by FACS. This would provide us with cells that can be subjected to NanoTag in order to obtain an *in vivo* whole-genome profile of the protein of interest.

This experiment would showcase the ability of NanoTag to provide a more native-like picture of the molecular landscape in animal cells, as, in contrast to ChIP-seq, it does not involve antibodies that can bind unspecifically or chemical fixation of the tissue. An ideal experiment to showcase the ability of our approach to provide *in vivo* information would

be to use NanoTag to profile a transcription factor for which no ChIP-grade antibodies exist but for which a transgenic animal line that expresses this transcription factor as a fusion to GFP has been generated. For example, to our knowledge, no ChIP-validated antibody exists against PGC1α, a transcriptional co-activator that regulates mitochondrial biogenesis[109]. Yeo et al. have conducted *in vivo* transfection of GFP-tagged PGC1α in mouse muscle[110]. We could envision using NanoTag on muscle cells from such mice to profile the binding sites of PGC1α.

Envisioned as described throughout this section, NanoTag should enable multimodal epigenomic profiling in various cellular contexts. We expect that NanoTag will reveal interactions between different epigenomic features and their function in a wide range of dynamic biological processes, such as malignant transformation, adipocyte differentiation, induction of autoimmunity or spermatogenesis.

5. Conclusion

In this work we achieved the successful cloning and purification of GBN-Tn5 and proved its functionality in mediating DNA transposition. We expanded a previously established qPCR assay to assess Tn5 activity by devising a positive control. We have shown that NanoTag enables the successful generation of DNA libraries from a mESC model expressing various engineered chromatin readers. These preliminary results suggest that NanoTag is a viable approach for targeted epigenomic profiling. To provide definite proof that NanoTag functions as expected, we will sequence the DNA libraries from cells expressing different eCRs and compare the results to homologous ChIP-seq data. As a highly versatile and unfixed *in situ* method, NanoTag enables targeted multimodal epigenomic profiling without the use of antibodies, constituting an important step in the effort to better characterize the epigenomic network and its roles in the cell.

6. Appendix

6.1. Plasmid Sequence

The complete sequence of the GFP nanobody-Tn5 plasmid:

5’ atacactccgctatcgctacgtgactgggtcatggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgt
ctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcatgtgtcagaggttttcaccgtcatcaccgaaacgcgc
gaggcagctgcggtaaagctcatcagcgtggtcgtgcagcgattcacagatgtctgcctgttcatccgcgtccagctcgttgagtttctcc
agaagcgttaatgtctggcttctgataaagcgggccatgttaagggcggttttttcctgtttggtcactgatgcctccgtgtaagggggattt
ctgttcatgggggtaatgataccgatgaaacgagagaggatgctcacgatacgggttactgatgatgaacatgcccggttactggaac
gttgtgagggtaaacaactggcggtatggatgcggcgggaccagagaaaaatcactcagggtcaatgccagccgaacgccagca
agacgtagcccagcgcgtcggccgccatgccggcgataatggcctgcttctcgccgaaacgtttggtggcgggaccagtgacgaag
gcttgagcgagggcgtgcaagattccgaataccgcaagcgacaggccgatcatcgtcgcgctccagcgaaagcggtcctcgccga
aaatgacccagagcgctgccggcacctgtcctacgagttgcatgataaagaagacagtcataagtgcggcgacgatagtcatgccc
cgcgcccaccggaaggagctgactgggttgaaggctctcaagggcatcggtcgagatcccggtgcctaatgagtgagctaacttac
attaattgcgttgcgctcactgcccgctttccagtcgggaaacctgtcgtgccagctgcattaatgaatcggccaacgcgcggggagag
gcggtttgcgtattgggcgccagggtggtttttcttttcaccagtgagacgggcaacagctgattgcccttcaccgcctggccctgagag
agttgcagcaagcggtccacgctggtttgccccagcaggcgaaaatcctgtttgatggtggttaacggcgggatataacatgagctgtc
ttcggtatcgtcgtatcccactaccgagatatccgcaccaacgcgcagcccggactcggtaatggcgcgcattgcgcccagcgccat
ctgatcgttggcaaccagcatcgcagtgggaacgatgccctcattcagcatttgcatggtttgttgaaaaccggacatggcactccagt
cgccttcccgttccgctatcggctgaatttgattgcgagtgagatatttatgccagccagccagacgcagacgcgccgagacagaactt
aatgggcccgctaacagcgcgatttgctggtgacccaatgcgaccagatgctccacgcccagtcgcgtaccgtcttcatgggagaaa
ataatactgttgatgggtgtctggtcagagacatcaagaaataacgccggaacattagtgcaggcagcttccacagcaatggcatcct
ggtcatccagcggatagttaatgatcagcccactgacgcgttgcgcgagaagattgtgcaccgccgctttacaggcttcgacgccgctt
cgttctaccatcgacaccaccacgctggcacccagttgatcggcgcgagatttaatcgccgcgacaatttgcgacggcgcgtgcagg
gccagactggaggtggcaacgccaatcagcaacgactgtttgcccgccagttgttgtgccacgcggttgggaatgtaattcagctccg
ccatcgccgcttccacttttcccgcgttttcgcagaaacgtggctggcctggttcaccacgcgggaaacggtctgataagagacaccg
gcatactctgcgacatcgtataacgttactggtttcacattcaccaccctgaattgactctcttccgggcgctatcatgccataccgcgaa
aggttttgcgccattcgatggtgtccgggatctcgacgctctcccttatgcgactcctgcattaggaagcagcccagtagtaggttgagg
ccgttgagcaccgccgccgcaaggaatggtgcatgccggcatgccgccctttcgtcttcaagaattaattcccaattccccaggcatca
aataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcctgagtaggacaaatccgccg
ggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcaggacgcccgccataaactgccaggaattaattcccc
aggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcctgagtaggaca
aatccgccgggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcaggacgcccgccataaactgccaggaa
ttaattccccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcctga
gtaggacaaatccgccgggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcaggacgcccgccataaact
gccaggaattaattccccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaac
gctctcctgagtaggacaaatccgccgggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcaggacgcccg
ccataaactgccaggaattaattccccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgt
cggtgaacgctctcctgagtaggacaaatccgccgggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcag
gacgcccgccataaactgccaggaattggggatcggaattaattcccggtttaaaccggggatctcgatcccgcgaaattaatacga
ctcactatagggaattgtgagcggataacaattcccctctagaaataatttgtttaactttaagaaggagatataccatgggtgattac
aaggatcacgatggcgattacaaggatcacgatatcgattacaaggatgatgatgataagatgaccatgattacgccaagcttgggc
gttcagctggttgaaagcggtggtgcactggttcagcctggtggtagcctgcgtctgagctgtgcagcaagcggttttccggttaatcgtt
atagcatgcgttggtatcgtcaggcaccgggtaaagaacgtgaatgggttgcaggtatgagcagtgccggtgatcgtagcagctatga

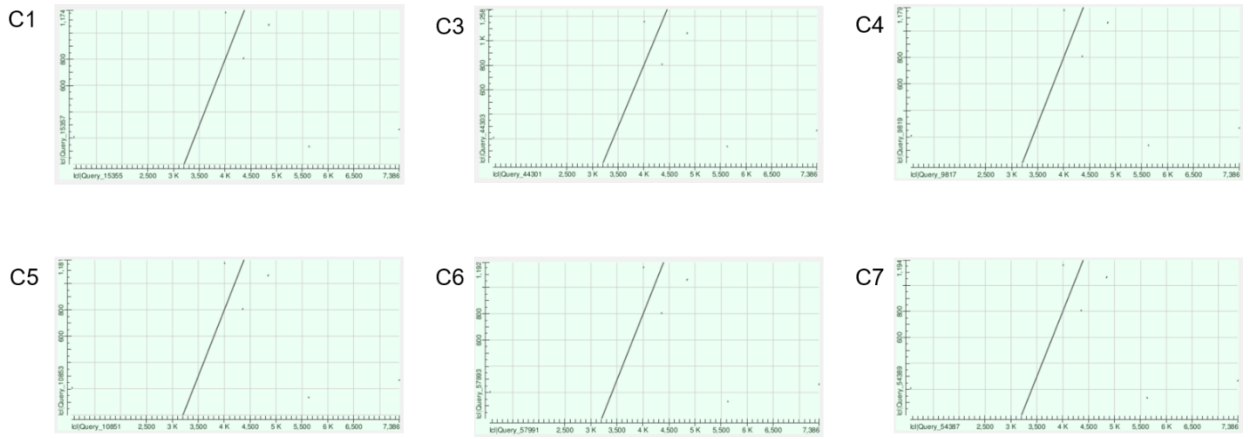
agatagcgttaaaggtcgttttaccatcagccgtgatgatgcacgtaataccgtttatctgcaaatgaatagcctgaaaccggaagata
ccgcagtgtattattgcaatgttaacgtgggctttgaatattggggtcagggcacccaggttaccgttagcagcggcgaattcggtggcg
gtggctctggcggtggtgggagtggaggtgggggatcaggaggaggcggttcccatatgattaccagtgcactgcatcgtgcggcgg
attgggcgaaaagcgtgttttctagtgctgcgctgggtgatccgcgtcgtaccgcgcgtctggtgaatgttgcggcgcaactggccaaat
atagcggcaaaagcattaccattagcagcgaaggcagcaaagccatgcaggaaggcgcgtatcgttttattcgtaatccgaacgtg
agcgcggaagcgattcgtaaagcgggtgccatgcagaccgtgaaactggcccaggaatttccggaactgctggcaattgaagatac
cacctctctgagctatcgtcatcaggtggcggaagaactgggcaaactgggtagcattcaggataaaagccgtggttggtgggtgcat
agcgtgctgctgctggaagcgaccacctttcgtaccgtgggcctgctgcatcaagaatggtggatgcgtccggatgatccggcggatg
cggatgaaaaagaaagcggcaaatggctggccgctgctgcaacttcgcgtctgagaatgggcagcatgatgagcaacgtgattgc
ggtgtgcgatcgtgaagcggatattcatgcgtatctgcaagataaactggcccataacgaacgttttgtggtgcgtagcaaacatccgc
gtaaagatgtggaaagcggcctgtatctgtatgatcacctgaaaaaccagccggaactgggcggctatcagattagcattccgcaga
aaggcgtggtggataaacgtggcaaacgtaaaaaccgtccggcgcgtaaagcgagcctgagcctgcgtagcggccgtattaccct
gaaacagggcaacattaccctgaacgcggtgctggccgaagaaattaatccgccgaaaggcgaaaccccgctgaaatggctgct
gctgaccagcgagccggtggaaagtctggcccaagcgctgcgtgtgattgatatttatacccatcgttggcgcattgaagaatttcaca
aagcgtggaaaacgggtgcgggtgcggaacgtcagcgtatggaagaaccggataacctggaacgtatggtgagcattctgagcttt
gtggcggtgcgtctgctgcaactgcgtgaatcttttactccgccgcaagcactgcgtgcgcagggcctgctgaaagaagcggaacac
gttgaaagccagagcgcggaaaccgtgctgaccccggatgaatgccaactgctgggctatctggataaaggcaaacgcaaacgc
aaagaaaaagcgggcagcctgcaatgggcgtatatggcgattgcgcgtctgggcggctttatggatagcaaacgtaccggcattgc
gagctggggtgcgctgtgggaaggttgggaagcgctgcaaagcaaactggatggctttctggccgcgaaagacctgatggcgcag
ggcattaaaatctgcatcacgggagatgcactagttgccctacccgagggcgagtcggtacgcatcgccgacatcgtgccgggtgcg
cggcccaacagtgacaacgccatcgacctgaaagtccttgaccggcatggcaatcccgtgctcgccgaccggctgttccactccgg
cgagcatccggtgtacacggtgcgtacggtcgaaggtctgcgtgtgacgggcaccgcgaaccacccgttgttgtgtttggtcgacgtc
gccggggtgccgaccctgctgtggaagctgatcgacgaaatcaagccgggcgattacgcggtgattcaacgcagcgcattcagcgt
cgactgtgcaggttttgcccgcgggaaacccgaatttgcgcccacaacctacacagtcggcgtccctggactggtgcgtttcttggaag
cacaccaccgagacccggacgcccaagctatcgccgacgagctgaccgacgggcggttctactacgcgaaagtcgccagtgtca
ccgacgccggcgtgcagccggtgtatagccttcgtgtcgacacggcagaccacgcgtttatcacgaacgggttcgtcagccacgcta
ctggcctcaccggtctgaactcaggcctcacgacaaatcctggtgtatccgcttggcaggtcaacacagcttatactgcgggacaattg
gtcacatataacggcaagacgtataaatgtttgcagccccacacctccttggcaggatgggaaccatccaacgttcctgccttgtggca
gcttcaatgactgcaggaaggggatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataa
ctagcataaccccttggggcctctaaacgggtcttgaggggtttttgctgaaaggaggaactatatccggataactacgtcaggtggc
acttttcggggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataa
atgcttcaataatattgaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcctgtttttgc
tcacccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagtgggttacatcgaactggatctcaacagcg
gtaagatccttgagagttttcgccccgaagaacgtttcccaatgatgagcacttttaaagttctgctatgtggcgcggtattatcccgtgttg
acgccgggcaagagcaactcggtcgccgcatacactattctcagaatgacttggttgagtactcaccagtcacagaaaagcatcttac
ggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgcggccaacttacttctgacaacgatcgg
aggaccgaaggagctaaccgcttttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaaccggagctgaatgaagcc
ataccaaacgacgagcgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactattaactggcgaactacttactcta
gcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctggctggtttatt
gctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccctcccgtatcgtagttat
ctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaact
gtcagaccaagtttactcatatatactttagattgatttaccccggttgataatcagaaaagccccaaaaacaggaagattgtataagca
aatatttaaattgtaaacgttaatattttgttaaaattcgcgttaaatttttgttaaatcagctcatttttttaaccaataggccgaaatcggcaaa
atcccttataaatcaaaagaatagcccgagatagggttgagtgttgttccagtttggaacaagagtccactattaaagaacgtggactc
caacgtcaaagggcgaaaaaccgtctatcagggcgatggcccactacgtgaaccatcacccaaatcaagttttttggggtcgaggtg
ccgtaaagcactaaatcggaaccctaaagggagcccccgatttagagcttgacggggaaagccggcgaacgtggcgagaaagg
aagggaagaaagcgaaaggagcgggcgctagggcgctggcaagtgtagcggtcacgctgcgcgtaaccaccacacccgccgc

gcttaatgcgccgctacagggcgcgtaaaaggatctaggtgaagatccttttgataatctcatgaccaaaatcccttaacgtgagttttc
gttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatccttttttctgcgcgtaatctgctgcttgcaaacaaaa
aaaccaccgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagat
accaaatactgtccttctagtgtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgt
taccagtggctgctgccagtggcgataagtcgtgtcttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcggg
ctgaacgggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaactgagatacctacagcgtgagctatgagaa
agcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgagggag
cttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggg
cggagcctatggaaaaacgccagcaacgcggccttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatccc
ctgattctgtggataaccgtattaccgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgag
cgaggaagctatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagt 3’

6.2. Sequence alignment of GFP nanobody-Tn5 plasmid

After ligation of the GBN-Tn5 sequence into the host plasmid, we transformed the ligated DNA into Subcloning Efficiency DH5α competent bacterial cells, spread the transformed bacteria onto LB agar + Ampicillin plates and incubated overnight at 37ºC (see section 2.1). Seven colonies grew on the plates. All seven colonies were used to prepare liquid cultures and sequenced. Only the plasmid extracted from one colony (C2) was used further. The sequencing results obtained for colony C2 were shown in Figure 12. Sequencing results of plasmids purified from the other six colonies are shown in Figure 24. Sequencing results showed that, except for colony C5, which included a point mutation, the sequences of all the other plasmids matched the expected sequence.

a



b

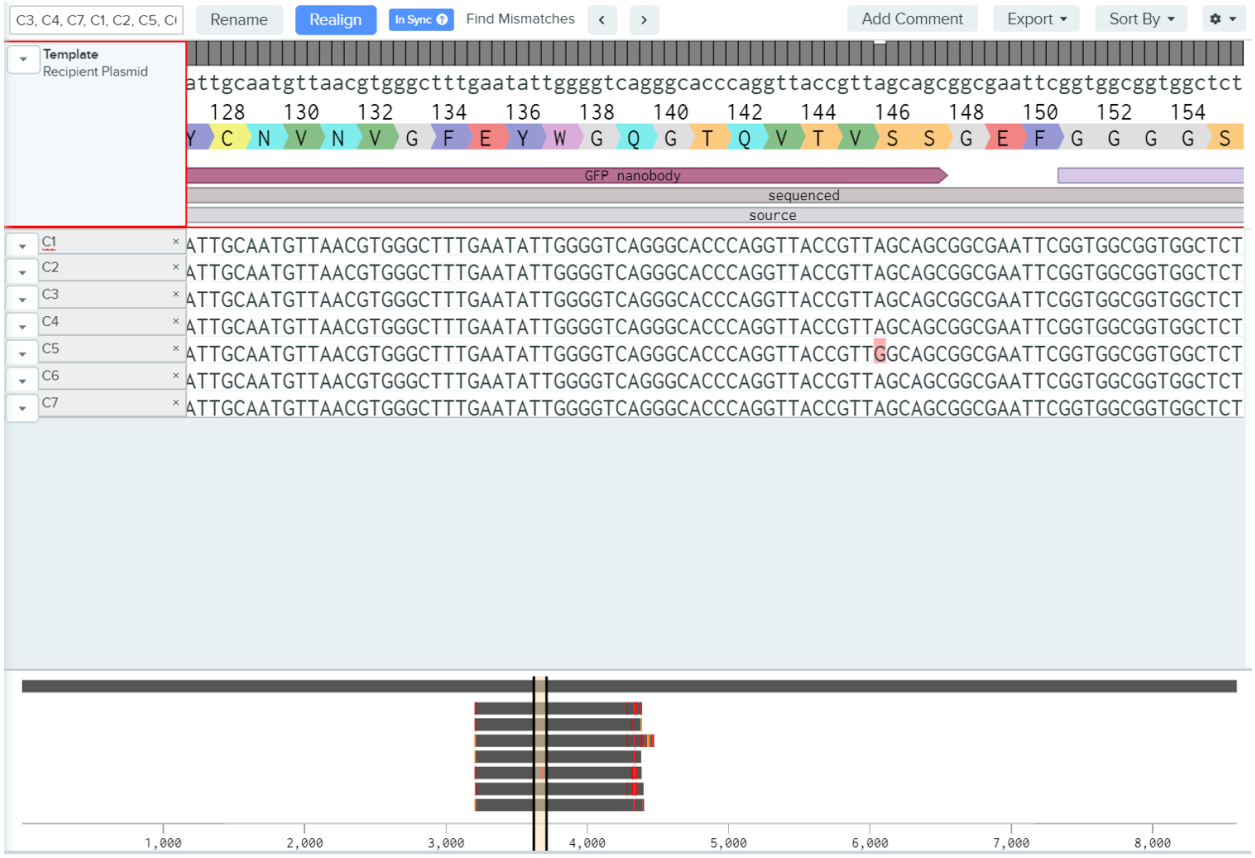


Figure 24 | **Sequencing alignments between the DNA of the colonies obtained following transformation of GBN-Tn5 plasmid into competent bacterial cells and the expected plasmid sequence. (a)** Dot plots showing alignment between the DNA of each colony and the expected plasmid sequence. **(b)** Partial alignment at nucleotide resolution, showing that the DNA of colony 5 contained a mutation.

7. References

1. Ficz, G. New insights into mechanisms that regulate DNA methylation patterning. *J. Exp. Biol.* **218**, 14–20 (2015).
2. Alegría-Torres, J. A., Baccarelli, A. & Bollati, V. Epigenetics and lifestyle. *Epigenomics* **3**, 267–277 (2011).
3. Hoppeler, H. H. Epigenetics in comparative physiology. *J. Exp. Biol.* **218**, 6–6 (2015).
4. Cardoso, S. D., Teles, M. C. & Oliveira, R. F. Neurogenomic mechanisms of social plasticity. *J. Exp. Biol.* **218**, 140–149 (2015).
5. Woldemichael, B. T. & Mansuy, I. M. Epigenetic Basis of Memory. in *Learning and Memory: A Comprehensive Reference* 247–256 (Elsevier, 2017). doi:10.1016/B978-0-12-809324-5.21116-4
6. O'Neill, C. The epigenetics of embryo development. *Anim. Front.* **5**, 42–49 (2015).
7. Zhang, Q. & Cao, X. Epigenetic regulation of the innate immune response to infection. *Nat. Rev. Immunol.* **19**, 417–432 (2019).
8. Moosavi, A. & Ardekani, A. M. Role of epigenetics in biology and human diseases. *Iranian Biomedical Journal* **20**, 246–258 (2016).
9. Tzika, E., Dreker, T. & Imhof, A. Epigenetics and Metabolism in Health and Disease. *Front. Genet.* **9**, (2018).
10. Pal, S. & Tyler, J. K. Epigenetics and aging. *Sci. Adv.* **2**, e1600584 (2016).
11. Bohacek, J. & Mansuy, I. M. Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nat. Rev. Genet.* **16**, 641–652 (2015).
12. Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. *Cell* **128**, 787–800 (2007).
13. Cremer, T. *et al.* Chromosome territories – a functional nuclear landscape. *Curr. Opin. Cell Biol.* **18**, 307–316 (2006).
14. van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780–791 (2017).
15. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16 (2020).
16. Stricker, S. H., Köferle, A. & Beck, S. From profiles to function in epigenomics. *Nat. Rev. Genet.* **18**, 51–66 (2017).
17. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nat. Methods* **4**, 895–901 (2007).

18. van Berkum, N. L. & Dekker, J. Determining Spatial Chromatin Organization of Large Genomic Regions Using 5C Technology. in *Methods in Molecular Biology* **567**, 189–213 (NIH Public Access, 2009).

19. Belton, J.-M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

20. Mieczkowski, J. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **7**, 11485 (2016).

21. Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).

22. Cook, A., Mieczkowski, J. & Tolstorukov, M. Y. Single-Assay Profiling of Nucleosome Occupancy and Chromatin Accessibility. *Curr. Protoc. Mol. Biol.* **120**, 21.34.1-21.34.18 (2017).

23. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

24. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

25. Métivier, R. *et al.* Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**, 45–50 (2008).

26. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

27. Jawaid, A., Roszkowski, M. & Mansuy, I. M. Transgenerational Epigenetics of Traumatic Stress. in *Progress in Molecular Biology and Translational Science* **158**, 273–298 (Elsevier B.V., 2018).

28. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002).

29. Okitsu, C. Y., Hsieh, J. C. F. & Hsieh, C. L. Transcriptional Activity Affects the H3K4me3 Level and Distribution in the Coding Region. *Mol. Cell. Biol.* **30**, 2933–2946 (2010).

30. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *BioEssays* **39**, e201600095 (2017).

31. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

32. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

33. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).

34. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).

35. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

36. Challen, G. A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2012).

37. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).

38. Cotton, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24**, 1528–1539 (2015).

39. Li, Y. & Tollefsbol, T. O. Impact on DNA Methylation in Cancer Prevention and Therapy by Bioactive Dietary Components. *Curr. Med. Chem.* **17**, 2141–2151 (2010).

40. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21**, 5427–5440 (2002).

41. Richardson, B. Primer: epigenetics of autoimmunity. *Nat. Clin. Pract. Rheumatol.* **3**, 521–527 (2007).

42. Urdinguio, R. G., Sanchez-Mut, J. V. & Esteller, M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *Lancet Neurol.* **8**, 1056–1072 (2009).

43. Jeong, M., Guzman, A. G. & Goodell, M. A. Genome-Wide Analysis of DNA Methylation in Hematopoietic Cells: DNA Methylation Analysis by WGBS. in *Methods in Molecular Biology* **1633**, 137–149 (Humana Press Inc., 2017).

44. Staunstrup, N. H. *et al.* Genome-wide DNA methylation profiling with MeDIP-seq using archived dried blood spots. *Clin. Epigenetics* **8**, 81 (2016).

45. Brinkman, A. B. *et al.* Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**, 232–236 (2010).

46. Lay, F. D., Kelly, T. K. & Jones, P. A. Nucleosome Occupancy and Methylome Sequencing (NOMe-seq). in *Methods in Molecular Biology* **1708**, 267–284 (Humana Press Inc., 2018).

47. Spektor, R., Tippens, N. D., Mimoso, C. A. & Soloway, P. D. methyl-ATAC-seq measures DNA methylation at accessible chromatin. *Genome Res.* **29**, 969–977 (2019).

48. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).

49. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193**, 651–669 (2013).

50. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12**, 246–258 (2011).

51. Zhao, B. S., Roundtree, I. A. & He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* **18**, 31–42 (2017).

52. Murr, R. Interplay Between Different Epigenetic Modifications and Mechanisms. in *Advances in Genetics* **70**, 101–141 (2010).

53. Morselli, M. *et al.* In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *Elife* **4**, (2015).

54. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

55. Kia, A. *et al.* Improved genome sequencing using an engineered transposase. *BMC Biotechnol.* **17**, 6 (2017).

56. Frenzel, A., Hust, M. & Schirrmann, T. Expression of Recombinant Antibodies. *Front. Immunol.* **4**, 217 (2013).

57. Bannas, P., Hambach, J. & Koch-Nolte, F. Nanobodies and Nanobody-Based Human Heavy Chain Antibodies As Antitumor Therapeutics. *Front. Immunol.* **8**, 1603 (2017).

58. Bieli, D. *et al.* Development and Application of Functionalized Protein Binders in Multicellular Organisms. in *International Review of Cell and Molecular Biology* **325**, 181–213 (Elsevier Inc., 2016).

59. Voskuil, J. Commercial antibodies and their validation. *F1000Research* **3**, 232 (2014).

60. Ruano-Gallego, D., Fraile, S., Gutierrez, C. & Fernández, L. Á. Screening and purification of nanobodies from E. coli culture supernatants using the hemolysin secretion system. *Microb. Cell Fact.* **18**, 47 (2019).

61. McMahon, C. *et al.* Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **25**, 289–296 (2018).

62. Plückthun, A. Ribosome Display: A Perspective. in *Methods in Molecular Biology* **805**, 3–28 (Humana Press Inc., 2012).

63. Stumpp, M. T., Binz, H. K. & Amstutz, P. DARPins: A new generation of protein therapeutics. *Drug Discov. Today* **13**, 695–701 (2008).

64. Ji, Z. *et al.* Genome-scale identification of transcription factors that mediate an inflammatory network during breast cellular transformation. *Nat. Commun.* **9**, 2068 (2018).

65. Huang, Q. *et al.* Mechanistic Insights Into the Interaction Between Transcription Factors and Epigenetic Modifications and the Contribution to the Development of Obesity. *Front. Endocrinol. (Lausanne).* **9**, 370 (2018).

66. Wu, H., Zhao, M., Yoshimura, A., Chang, C. & Lu, Q. Critical Link Between Epigenetics and Transcription Factors in the Induction of Autoimmunity: a Comprehensive Review. *Clin. Rev. Allergy Immunol.* **50**, 333–344 (2016).

67. Rykalina, V., Shadrin, A., Lehrach, H. & Borodina, T. qPCR-based characterization of DNA fragmentation efficiency of Tn5 transposomes. *Biol. Methods Protoc.* **2**, 1–8 (2017).

68. Villaseñor, R. *et al.* ChromID identifies the protein interactome at chromatin marks. *Nat. Biotechnol.* 1–9 (2020). doi:10.1038/s41587-020-0434-2

69. Hameed, U. F. S. *et al.* Transcriptional Repressor Domain of MBD1 is Intrinsically Disordered and Interacts with its Binding Partners in a Selective Manner. *Sci. Rep.* **4**, 4896 (2015).

70. Vermeulen, M. *et al.* Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* **131**, 58–69 (2007).

71. Kubala, M. H., Kovtun, O., Alexandrov, K. & Collins, B. M. Structural and thermodynamic analysis of the GFP:GFP-nanobody complex. *Protein Sci.* **19**, 2389–2401 (2010).

72. Ohki, I. *et al.* Solution Structure of the Methyl-CpG Binding Domain of Human MBD1 in Complex with Methylated DNA. *Cell* **105**, 487–497 (2001).

73. Katoh, Y., Nozaki, S., Hartanto, D., Miyano, R. & Nakayama, K. Architectures of multisubunit complexes revealed by a visible immunoprecipitation assay using fluorescent fusion proteins. *J. Cell Sci.* **128**, 2351–2362 (2015).

74. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

75. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

76. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

77. Thermo Scientific. *NanoDrop Spectrophotometers Protein A280*. (2010).

78. Thermo Fisher Scientific. *User Guide: Pierce BCA Protein Assay Kit*.

79. Krasnenko, A. *et al.* Effect of DNA insert length on whole-exome sequencing enrichment efficiency: an observational study. *Adv. Genomics Genet.* **Volume 8**, 13–15 (2018).

80. Sun, Z. *et al.* Nanobody-Alkaline Phosphatase Fusion Protein-Based Enzyme-Linked Immunosorbent Assay for One-Step Detection of Ochratoxin A in Rice. *Sensors* **18**, 4044 (2018).

81. Sheng, Y. *et al.* Nanobody-horseradish peroxidase fusion protein as an ultrasensitive probe to detect antibodies against Newcastle disease virus in the immunoassay. *J. Nanobiotechnology* **17**, 35 (2019).

82. Ren, W. *et al.* One-Step Ultrasensitive Bioluminescent Enzyme Immunoassay Based on Nanobody/Nanoluciferase Fusion for Detection of Aflatoxin B 1 in Cereal. *J. Agric. Food Chem.* **67**, 5221–5229 (2019).

83. Woo, C. Tools to facilitate manipulation of protein-specific glycosylation stoichiometry in cells. Available at: https://prevention.cancer.gov/funding-and-grants/funded-grants/U01CA242098. (Accessed: 5th March 2020)

84. Harrison, S. T. L. Bacterial cell disruption: A key unit operation in the recovery of intracellular products. *Biotechnol. Adv.* **9**, 217–240 (1991).

85. Shehadul Islam, M., Aryasomayajula, A. & Selvaganapathy, P. A Review on Macroscale and Microscale Cell Lysis Methods. *Micromachines* **8**, 83 (2017).

86. Hennig, B. P. *et al.* Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3: Genes|Genomes|Genetics* **8**, 79–89 (2018).

87. ExPASY. ProtParam References. Available at: https://web.expasy.org/protparam/protpar-ref.html. (Accessed: 5th March 2020)

88. Bohman, A. & Arnold, M. A. Molar Absorptivity Measurements in Absorbing Solvents: Impact on Solvent Absorptivity Values. *Appl. Spectrosc.* **71**, 446–455 (2017).

89. Liang, K. & Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**, 199 (2012).

90. Baubec, T., Ivánek, R., Lienert, F. & Schübeler, D. Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell* **153**, 480–492 (2013).

91. Greenfield, R. *et al.* Role of transcription complexes in the formation of the basal methylation pattern in early development. *Proc. Natl. Acad. Sci.* **115**, 10387–10391 (2018).

92. Saifudeen, Z. *et al.* A p53-Pax2 Pathway in Kidney Development: Implications for Nephrogenesis. *PLoS One* **7**, e44869 (2012).

93. Brown, J. B. & Celniker, S. E. Lessons from modENCODE. *Annu. Rev. Genomics Hum. Genet.* **16**, 31–53 (2015).

94. Kudron, M. M. *et al.* The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics* **208**, 937–949 (2018).

95. Fridy, P. C. *et al.* A robust pipeline for rapid production of versatile nanobody repertoires. *Nat. Methods* **11**, 1253–1260 (2014).

96. Grochowska, K. M. *et al.* A molecular mechanism by which amyloid-β induces transcriptional inactivation of CREB in Alzheimer's Disease. *bioRxiv* 2020.01.08.898304 (2020). doi:10.1101/2020.01.08.898304

97. Götzke, H. *et al.* The ALFA-tag is a highly versatile tool for nanobody-based bioscience applications. *Nat. Commun.* **10**, 4403 (2019).

98. Vigano, A. *et al.* DARPins recognizing mTFP1 as novel reagents for in vitro and in vivo protein manipulations. *bioRxiv* 354134 (2018). doi:10.1101/354134

99. Reyes, M., Billman, K., Hacohen, N. & Blainey, P. C. Simultaneous Profiling of Gene Expression and Chromatin Accessibility in Single Cells. *Adv. Biosyst.* **3**, 1900065 (2019).

100. Hu, P., Zhang, W., Xin, H. & Deng, G. Single Cell Isolation and Analysis. *Front. Cell Dev. Biol.* **4**, (2016).

101. Salomon, R. *et al.* Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip* **19**, 1706–1727 (2019).

102. Bryan, T. M. & Baumann, P. G-Quadruplexes: From Guanine Gels to Chemotherapeutics. in *Methods in molecular biology (Clifton, N.J.)* **608**, 1–16 (Humana Press, 2010).

103. Spiegel, J., Adhikari, S. & Balasubramanian, S. The Structure and Function of DNA G-Quadruplexes. *Trends Chem.* **2**, 123–136 (2020).

104. Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).

105. Scholz, O., Hansen, S. & Plückthun, A. G-quadruplexes are specifically recognized and distinguished by selected designed ankyrin repeat proteins. *Nucleic Acids Res.* **42**, 9182–9194 (2014).

106. Guillou, E. *et al.* Cohesin organizes chromatin loops at DNA replication factories. *Genes Dev.* **24**, 2812–2822 (2010).

107. Farrants, H. *et al.* Chemogenetic Control of Nanobodies. *Nat. Methods* **17**, 279–282 (2020).

108. Hudak, J. E. *et al.* Synthesis of Heterobifunctional Protein Fusions Using Copper-Free Click Chemistry and the Aldehyde Tag. *Angew. Chemie Int. Ed.* **51**, 4161–4165 (2012).

109. Austin, S. & St-Pierre, J. PGC1 and mitochondrial metabolism - emerging concepts and relevance in ageing and neurodegenerative disorders. *J. Cell Sci.* **125**, 4963–4971 (2012).

110. Yeo, D., Kang, C., Gomez-Cabrera, M. C., Vina, J. & Ji, L. L. Intensified mitophagy in skeletal muscle with aging is downregulated by PGC-1alpha overexpression in vivo. *Free Radic. Biol. Med.* **130**, 361–368 (2019).

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

TOWARDS A NOVEL APPROACH TO MULTIMODAL EPIGENOMIC PROFILING

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
DIMITRIU	MARIA-ANDREEA

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 11.03.2020	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.