A Bacterial Surface Display Platform for the Discovery of Cytosine Modification Readers from cDNA Libraries

Dissertation

Submitted for the degree of Doctor of Natural Sciences

(Dr. rer. nat.)

Presented by

Damian Schiller

at the

ccb fakultät für chemie
und chemische biologie

of the

tu technische universität
dortmund

Dortmund 2023

The past is written, but the future is left for us to write.

from Star Trek

The work for this thesis was executed between September 2019 and November 2023 in the group of Prof. Dr. Daniel Summerer at the TU Dortmund University. The work was funded by the TU Dortmund University and the ERC Horizon 2020 program (ERC EPICODE Grant Nr. 723863 and ERC COMBICODE Grant Nr. 101100794).

Acknowledgments

At very first, I would like to thank Prof. Dr. Daniel Summerer for giving me the opportunity to work on this interesting project in the field of epigenetics. Moreover, I would like to express my gratitude for his constant support and constructive discussions, as well as the opportunity for creative freedom and independence during my thesis. Further thanks go to Prof. Dr. Hannes Mutschler for agreeing to act as second reviewer of this thesis.

I would like to extend my gratitude to the members of my thesis advisory committee, Prof. Dr. Andrea Musacchio and Dr. 't Hart, for the fruitful discussions and valuable input. At this point, I would also like to thank the coordinators of the International Max-Planck-Research School for Living Matter, Dr. Lucia Sironi and Christa Hornemann, for enabling an excellent scientific and social environment from which I benefited throughout the entire time.

Of special importance to me are the current and former members of the AG Summerer, namely Dr. Anna Witte, Dr. Jan Wolfgramm, Dr. Shubhendu Palei, Dr. Benjamin Buchmuller, Dr. Álvaro Muñoz López, Dr. Tzu-Chen Lin, Dr. Anne Jung, Simone Eppmann, Nadine Schmidt, Sudakshina Banerjee, Lena Engelhard, Brinja Kosel, Zeyneb Cakil, Katrin Bigler, Lejla Maksumic, Kotryna Keliuotyte and Sayan Sil. These people have not only ensured the best possible working atmosphere but also provided helpful advice and support. My special thanks go to Dr. Benjamin Buchmuller for the helpful discussions and for introducing me into R, Sudakshina Banerjee for the positive and supporting attitude and Simone Eppmann for the helpful work inside and outside the lab.

Furthermore, I would like to thank all the members of the neighboring and collaborating groups. I would also like to thank my bachelor students Natalie Jácobo Goebbels, Nadeshda Kataev, and Leon Wagner who supported my work by cloning vectors and libraries and conducting insightful FACS experiments.

Not to forget, Martina Reibner, Petra Alhorn, Maria Sergani, Alicia Wachholz, and Ulrich Schoppe from the faculty's administration dealt with all the bureaucracy and general management and allowed everything to be carried out smoothly.

I can't express my gratitude enough for Jascha, Laura, Paul, Kristina, Veronika, Jakob, Frank, Milan, Niklas, David, Johanna, Peter, Lukas, and Alex, who are the people that supported me during this journey and helped me get through even the hardest times. I would like to extend this gratitude to the former colleagues and fellow students Natalie, Dominic, Anna, Sven, Mirko, and Christian, who became friends over the years.

At last, I would like to thank my parents Monika and Lothar, who always encouraged and supported me and who have offered me immense opportunities. At this point, I express my hopes that my late mother is in some way taking part in the last steps of my PhD journey.

Table of Contents

1 List of Figures

List of main figures

List of supplementary figures

2 List of Tables

List of main tables

List of supplementary tables

3 List of Equations

List of main equations

List of supplementary equations

4 List of Abbreviations

5caC	5-carboxyl cytosine
5fC	5-formyl cytosine
5hmC	5-hydroxymethyl cytosine
5mC	5-methyl cytosine
6mA	N6-methyl adenine
A	Adenine
AB	Antibody
APC	Allophycocyanin
AT	Autotransporter
Bam	β-barrel assembly machinery
BDH1	D-beta-hydroxybutyrate dehydrogenase
BER	Base excision repair
bHlH	Basic helix–loop–helix
bp	Base pair
Brd3	Bromodomain-containing protein 3
BS-seq	Bisulfite sequencing
C	Cytosine
C9orf85	Chromosome 9 open reading frame 85
cDNA	Complementary DNA
CDS	Coding sequence
cfu	Colony-forming units
CHD	Chromodomain/helicase/DNA binding domain
ChiP	Chromatin immunoprecipitation
CHMP6	Charged multivesicular body protein 6
CIRBP	Cold-inducible RNA-binding protein
CpG	Cytosine-phosphate-Guanine
CRIP1	Cysteine-rich protein 1
DHHC	Aspartate-histidine-histidine-cysteine tetrapeptide
DHX38	Pre-mRNA-splicing factor ATP-dependent RNA helicase
DNA	Deoxyribonucleic acid
DNMT	DNA methyl transferases
dNTP	Deoxyribonucleotide triphosphate
dsbA	Disulfide oxidoreductase
dsDNA	Double stranded DNA
EEF1A1	Elongation factor 1-alpha 1
E-ep.	E-epitope

EF	Enrichment factor
EGFR	Epidermal growth factor receptor
EHEC	Enterohemorrhagic *E. coli*
EMSA	Electrophoretic mobility shift assay
EPEC	Enteropathogenic *E. coli*
ER	Endoplasmic reticulum
ESCRT	Endosomal sorting complexes required for transport
ESYT2	Extended synaptotagmin-2
FACS	Fluorescence-activated cell sorting
FAM50B	Family With Sequence Similarity 50 Member B
FDA	U.S. Food and Drug Administration
FITC	Fluorescein isothiocyanate
G	Guanine
Gbp	Giga base pairs
gDNA	Genomic DANN
GPI	Glycosylphosphatidylinositol
GTF	General transcription factor
HA	Human influenza hemagglutinin
HDAC	Histone deacetylase
HhH	Helix-hairpin-helix
hMBD2	Human Methyl-CpG-binding domain protein 2
HNRNPA0	Heterogeneous nuclear ribonucleoprotein A0
HNRNPDL	Heterogeneous nuclear ribonucleoprotein D-like
IM	Inner membrane
IMUP	Immortalization up-regulated protein
Int'	Truncated intimin
Ire	Initiator elements
Kac	Acetylated lysine
LB	Lysogeny broth
LIM	LIN-11, Isl-1 and MEC-3
LINE	Long interspersed element
lncRNA	Long non-coding RNA
Lpp	Major outer membrane lipoprotein
LTR	Long tandem repeat
MACS	Magnetic-activated cell sorting
MBD	Methyl binding domain
MeCP2	Methyl-CpG-binding protein 2
mESC	Mouse embryonic stem cells

miRNA	Micro RNA
mRNA	Messenger RNA
MS	Mass spectrometry
MT2A	Metallothionein-2
MTE	Motif ten elements
Myc	Myc proto-oncogene protein
MYPN	Myopalladin
NCOR2	Nuclear receptor corepressor 2
NGS	Next generation sequencing
NPC	Neuronal progenitor cells
NTHL1	Endonuclease III-like protein 1
NuRD	Nucleosome Remodeling Deacetylase complex
OM	Outer membrane
OmpA	Outer membrane protein A
OMPT	Outer membrane protease T
ORF	Open reading frame
PAL	Peptidoglycan-associated lipoprotein
PCR	Polymerase chain reaction
PD	Periplasmic domain
PE	Phycoerythrin
PhoE	Phosphate-limitation-inducible outer-membrane protein
PIC	Pre initiation complex
poly-A	Poly adenine
poly-dT	Poly thymine
PPI	Protein-protein interactor
PTM	Post-translational modification
PYHIN1	Pyrin and HIN domain-containing protein 1
RACK1	Small ribosomal subunit protein
RBM3	RNA-binding protein 3
RCOR3	REST corepressor 3
RGG	Arginine–glycine–glycine repeat
RNA	Ribonucleic acid
RNAi	RNA interference
RPL24	Large ribosomal subunit protein L24
RRM	RNA recognition motif
rRNA	Ribosomal RNA
SAH	S-adenosyl homocysteine
SAM	S-adenosyl methionine

SAv	Streptavidin
scFv	Single-chain variable fragment
SINE	Short interspersed element
siRNA	Small interfering RNA
snRNA	Small nuclear RNA
SP	Signal peptide
S-phase	Synthesis phase
SRA	SET and ring finger associated domain
ssDNA	Single stranded DNA
T	Thymine
Tam	Translocation and assembly module
TCF25	Transcription Factor 25
TDG	Thymine DNA glycosylase
TEG	Triethylene glycol
TET	Ten-eleven translocation dioxygenase
TEV	Tobacco etch virus
TF	Transcription factor
TFIID	Transcription factor IID
TM	Transmembrane domain
TPM1	Tropomyosin alpha-1 chain
TRD	Target-recognition domain
tRNA	Transfer RNA
TSS	Transcription start site
U	Uracil
uORF	Upstream ORF
UTR	Untranslated region
UV-Vis	Ultraviolet–visible
WD	Tryptophan-aspartate repeat
ZC3H3	Zinc finger CCCH domain-containing protein 3
ZDHHC4	Zinc Finger DHHC-Type Palmitoyltransferase 4
Znf	Zinc finger protein

5 List of Publications

Parts of this thesis will be published in:

[1] **Schiller, D.**, Kataev, N., Jácobo Goebbels, N., Summerer, D. A bacterial surface display platform for the discovery of cytosine modification readers from cDNA libraries. Manuscript in preparation.

Further publications:

[2] Nowacki, J., Malenica, M., Schmeing, S., **Schiller, D.**, Buchmuller, B., Amrahova, G., & 't Hart, P. (2023). A translational repression reporter assay for the analysis of RNA-binding protein consensus sites. *RNA Biology*, *20*(1), 85–94.

[3] Borgelt, L., Haacke, N., Lampe, P., Qiu, X., Gasper, R., **Schiller, D.**, Hwang, J., Sievers, S., & Wu, P. (2022). Small-molecule screening of ribonuclease L binders for RNA degradation. *Biomedicine and Pharmacotherapy*, *154*(113589).

[4] Hwang, J., Qiu, X., Borgelt, L., Haacke, N., Kanis, L., Petroulia, S., Gasper, R., **Schiller, D.**, Lampe, P., Sievers, S., Imig, J., & Wu, P. (2022). Synthesis and evaluation of RNase L-binding 2-aminothiophenes as anticancer agents. *Bioorganic and Medicinal Chemistry*, *58*(116653).

6 Abstract

5-methylcytosine (5mC) occurs in palindromic cytosine guanine dyads (CpGs) of mammalian genomes and is a key element of epigenetic transcription regulation. Central to these regulatory functions is the ability of cytosine modifications to modulate the interaction of chromatin proteins with DNA. In this context, reader proteins are known to selectively recognize the cytosine modification. Interactome profiling studies based on pulldowns of nuclear extracts in combination with mass spectrometry-based (MS) proteomics have been the main approach to discovering readers of 5mC. However, this approach may miss important (anti-)readers due to the competition of proteins for the probe and low expression levels. Moreover, direct readers cannot be distinguished from indirect binders of protein complexes.

In the scope of this thesis, a bacterial surface display system for discovering novel reader candidates from human cDNA was established. In detail, our approach offers screenings of cDNA-encoded protein libraries independent of their endogenous expression levels and without competition with other proteins. It further allows for rapid iterative FACS selections with defined, fluorescently labeled on- and off-target DNA probes. Furthermore, the system benefits from the fast protein expression machinery and growth of bacterial cells. We created surface display libraries of full-length and fragmented coding sequences (CDS) from human cDNA and demonstrated the compatibility of the Intimin surface display platform with displaying human proteins and protein fragments. Moving on, we proved the functionality of our selection system by enriching known 5mC reader candidates from a display library of human thyroid cDNA. In addition, we were able to identify novel reader candidates. Therefore, this study offers a promising tool to complement existing efforts in discovering 5mC reader candidates. Furthermore, the tool bears the potential to be extended to the oxidized derivates of 5mC and their combinations in CpGs, which have not been investigated so far.

7 Zusammenfassung

5-Methylcytosin (5mC) tritt in dem Genom von Säugtieren in palindromischen Cytosin-Guanin Dyaden (CpGs) auf und ist ein Schlüsselelement in der epigenetischen Transkriptionsregulation. Von zentraler Bedeutung für diese regulatorischen Funktion ist die Eigenschaft der Cytosinmodifikation die Interaktion von Chromatinproteinen mit der DNA zu modulieren. In diesem Zusammenhang sind sogenannte „Reader“-Proteine dafür bekannt, die Modifikation selektiv zu erkennen und an diese zu binden. Die Anreicherung von 5mC-„Reader“-Proteinen aus dem nuklearen Lysat und deren anschließender Identifizierung mit massenspektrometrie-gestützter Proteomanalyse ist die verbreitetste Methode für das Interaktom-Profiling von Cytosinmodifikationen. Allerdings können niedrige Expressionslevel und die Konkurrenz der Proteine um die DNA-Sonde dazu führen, dass wichtige (Anti-)“Reader“ übersehen werden. Darüber hinaus können mit dieser Methode direkte „Reader“ nicht von indirekten Bindern aus Proteinkomplexen unterschieden werden.

In Rahmen dieser Thesis wurde ein System für die Entdeckung neuer „Reader“-Kandidaten aus humaner cDNA mittels Präsentation von Proteinen auf der bakteriellen Oberfläche etabliert. Genauer gesagt ermöglicht unser Ansatz eine Sondierung von cDNA-kodierten Proteinbibliotheken unabhängig von ihrem endogenen Expressionslevel und ohne Konkurrenz der Proteine untereinander. Darüber hinaus ermöglicht er schnelle, iterative FACS-Selektionen mit definierten, fluoreszenzmarkierten Sonden für die Zielmodifikation und unerwünschte andere Modifikationen. Das Selektionssystem profitiert des Weiteren von der schnellen Proteinexpression und dem schnellen Wachstum bakterieller Zellen.

Wir konnten cDNA-basierte Proteinbibliotheken für die Oberflächenpräsentation erstellen und damit die Kompatibilität des intimin-Systems hinsichtlich der Präsentation von humanen Proteinen und Proteinfragmenten demonstrieren. Weiterführend konnten wir die Funktionalität des Selektionssystems durch die Anreichung von bekannten und neuen 5mC-„Reader“-Kandidaten aus humaner Schilddrüsen cDNA zeigen. Aufgrund dieser Ergebnisse stellt unser System ein vielversprechendes neues Werkzeug zur Komplementierung bereits existierender Studien zur Entdeckung von 5mC-Reader Kandidaten dar und kann auch auf die oxidierten Derivate von 5mC und ihre bisher unerforschten Kombinationsmöglichkeiten in CpGs angewandt werden.

8 Introduction

8.1 Genome organization and composition in eukaryotes

8.1.1 Storage of genetic information in deoxyribonucleic acid sequences

All living organisms on our planet, prokaryotes, eukaryotes, and archaea, store information about their molecular composition and functional organization in their genome. This genetic information is encoded by defined chemical building blocks, the so-called nucleobases, and saved as the sequential arrangement of bases in a linear polymer - the deoxyribonucleic acid (DNA). The four different nucleobases in DNA are composed of heterocycles that are derivates of purine, adenine (A), and guanine (G), or pyrimidine, thymine (T), and cytosine (C) (Figure 1).[1,2]

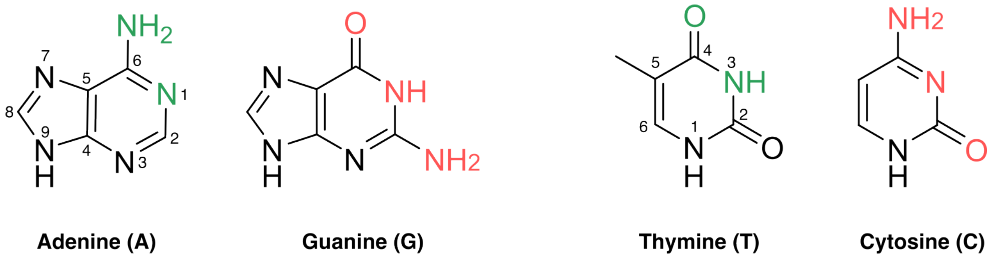

Figure 1: Chemical structures of the DNA nucleobases. Numbering of the heterocyclic atoms of purine and pyrimidine bases is done exemplarily for adenine and thymine. Hydrogen bond-forming atoms or groups are colored.

These nucleobases are attached to deoxyribose, or in ribonucleic acid (RNA) to ribose sugars, via N^9 (for purines) or N^1 (for pyrimidines) and are then called nucleosides. Nucleosides form nucleotides when additional phosphate groups are attached to the hydroxy group of 5' carbon. Multiple nucleotides form the DNA polymer via a covalent linkage. The sugar molecules within these DNA polymers are connected through phosphodiester bonds at their 3' and 5' carbons. This linkage creates the negatively charged sugar-phosphate backbone, which imparts directionality to a DNA strand, resulting in a distinct 5'- and 3'-end (Figure 2).[1]

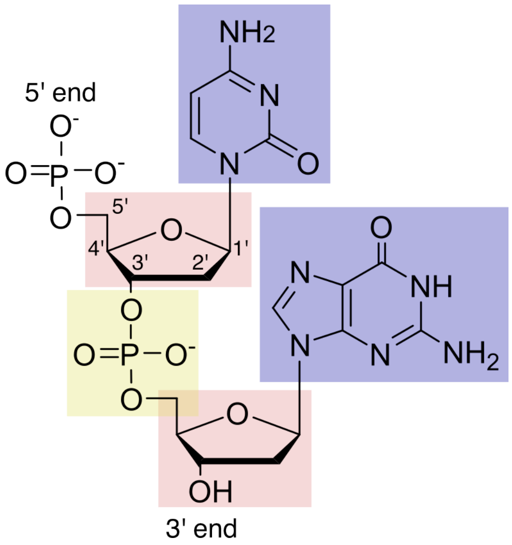

Figure 2: Chemical structure of the dinucleotide CG. The phosphodiester bond is highlighted in yellow, deoxyribose in red, and the respective nucleobases in blue.

Genomic DNA (gDNA) consists of two opposing DNA strands, which align in an antiparallel fashion to create a higher-order double-helical structure that was discovered by Watson and Crick[2] in 1953 (Figure 3a and b). Within the double helix, the sugar-phosphate backbone faces outwards while the inward-facing nucleobases of the double-stranded DNA (dsDNA) are held together by π- π interactions and by hydrogen bonds. Notably, these hydrogen bonds are exclusively energetically favorable between specific nucleobase pairs of the opposing strands, namely C-G and A-T, which are also referred to as Watson-Crick base pairs.[2,3] This ensures the complementarity of both DNA strands according to the rule of Chargaff[4] which states that the complementary nucleobase pairs occur in equimolar amounts in dsDNA. Cytosine and guanine are connected via three hydrogen bonds whereas the adenine-thymine base pair forms two bonds (Figure 3c).[2,3] The complementarity of the two strands of dsDNA is the basis for the inheritability of genetic information which is illustrated in section 8.1.2.

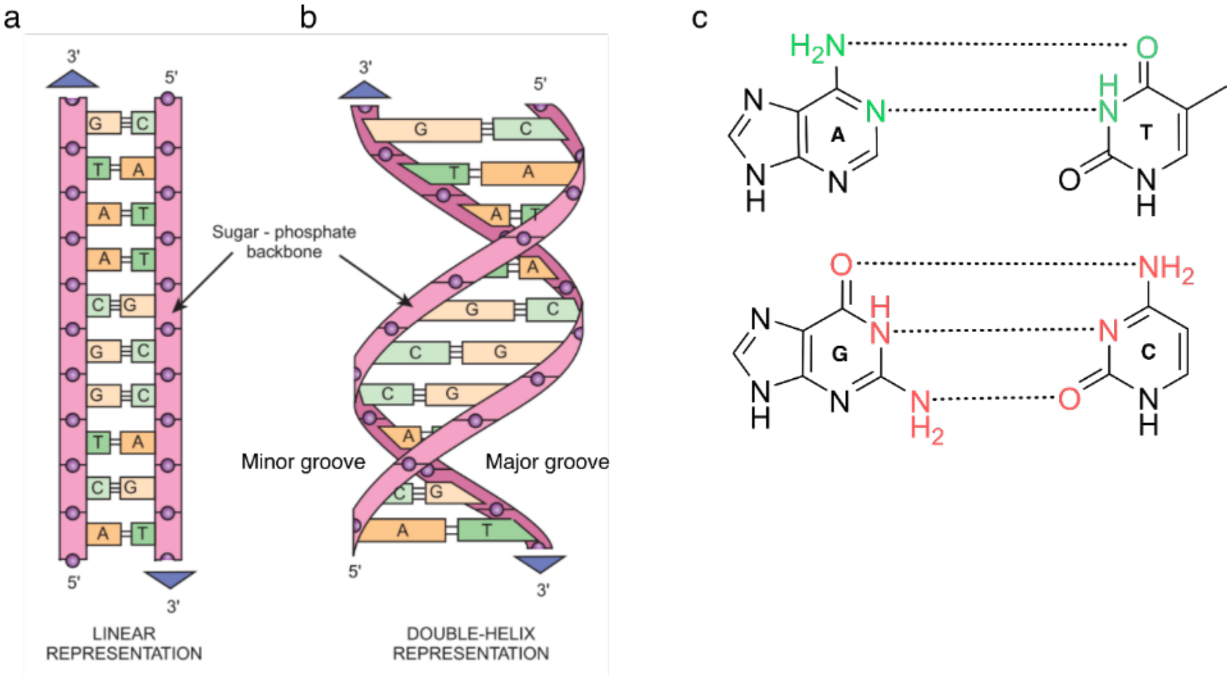

Figure 3: **The structure of the DNA helix.** Schematic representation of the DNA double-strand (a) and the DNA double helix (b). (c) Depiction of the chemical groups involved in the hydrogen bonds connecting the Watson Crick base pairs in dsDNA. Figures a and b are adapted from Clark et al.[5]

The physiologically relevant form of dsDNA in organisms is B-DNA. This form features a right-handed double helix with a diameter of approximately 23.7 Å, accommodates about 10.4 base pairs per helical turn, and contains minor and major grooves (Figure 3b).[6] Both grooves, especially the major groove, contain essential chemical and structural information, including sequence-specific hydrogen bond donor and acceptor sites. In particular, the major groove is an essential region for specific protein-DNA interactions of transcription factors (TFs) and epigenetic reader proteins, which will be discussed in more detail in section 8.2.3.[7]

8.1.2 DNA replication – inheritability of genetic information

As previously mentioned, the two strands of the DNA double helix align complementarily, allowing one strand to store the information of the opposing strand in its sequence. This enables a copying mechanism that uses the sequence of the strands as templates and that was already postulated by Watson and Crick in the early 1950s.[2]

In cells, the process of genome copying is referred to as replication and occurs during the synthesis phase (S-phase) of the cell cycle. Its fundamental role in life necessitates stringent regulation, involving numerous different proteins of which only the key players are highlighted here. First, the two strands are separated under ATP hydrolysis by helicase proteins to form the replication fork – the place where replication takes place. The resulting single-stranded DNA (ssDNA) serves as the template for DNA polymerases to copy the strand information by base pair complementarity. DNA polymerases are enzymes that generate DNA strands by catalyzing the phosphodiester bond formation between the 3′-hydroxy ends of the DNA strand and the 5′ end of nucleotides. Additionally, most replicative polymerases possess an exonuclease-based proofreading mechanism that removes mismatching incorporated bases, thereby enhancing the precision of replication by approximately 2-3 orders of magnitude.[8] Substrates of the polymerases are nucleotides in their triphosphorylated, active precursor state. As previously mentioned, DNA polymerases require a free 3′-hydroxy end for strand elongation and cannot initiate DNA strand synthesis from scratch. To initiate replication, DNA polymerase α, also known as primase, synthesizes a complementary RNA primer molecule to whose 3′-end other DNA polymerases can attach the first nucleotide.[9] The initiation of Replication occurs at 30,000-50,000 AT-rich DNA replication origins, enabling efficient copying of the over 6 gigabase pairs (Gbp)[10] in the human genome.[9,11] After replication, and before mitosis, each cell contains two chromosome sets, each bearing one original strand and a newly synthesized one. This is the final result of the semi-conservative replication process which was first explored by Meleson and Stahl in 1958.[12]

8.1.3 Flow of genetic information – the central dogma

For life to exist it is essential that the genetic information is transferred into a functional form which is in many cases a folded polypeptide chain – a protein. First, the DNA sequence is transcribed into a mobile molecule that can exit the nucleus and serves as a template for the cytoplasmic translation – the messenger RNA (mRNA). The transcription of genetic information in mRNA and the subsequent translation into a polypeptide sequence was postulated by Crick in 1958 as the 'central dogma of molecular biology' (Figure 4).[13]

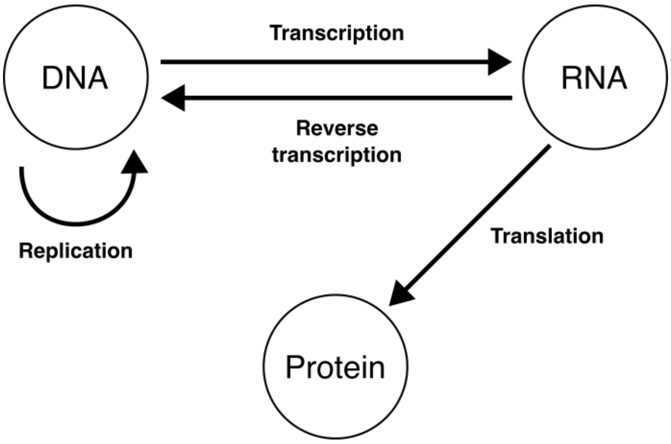

Figure 4: The central dogma of molecular biology as postulated by Crick. The informational flow is facilitated by template-based polymerization of biomolecules. Figure adapted and modified from Crick.[14]

Additionally, reverse transcription is used by retrotransposons and retroviruses to transfer information from an RNA template into DNA.[14] These processes, including replication, are the fundaments of the informational flow within cells and function based on template-based polymerization of biomolecules. Moreover, additional processes influence structural and functional information by other mechanisms including the chemical modification of the biopolymers.[15]

8.1.4 Transcription

Transcription is the cellular process in which the DNA sequence information is transferred in a complementary RNA sequence. RNA differs from DNA by an additional 2' hydroxy group at its ribose sugar moiety and the substitution of thymine by another purine base called uracil (U, Figure 5).

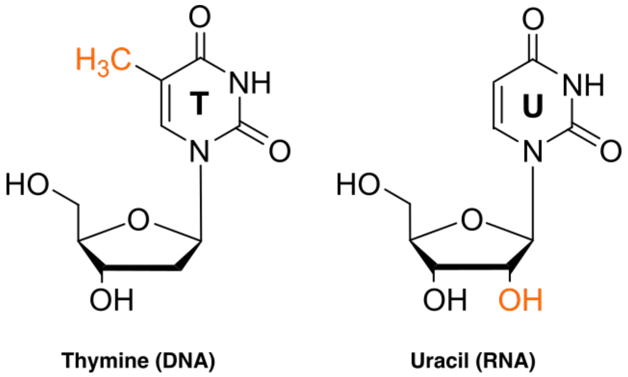

Figure 5: Structural differences of RNA and DNA by means of the substituted nucleotides thymine and uracil. Differences are highlighted in orange.

RNA molecules can be separated into three categories depending on their cellular functions: Protein-coding associated RNAs, regulatory RNAs, and parasitic RNAs.[16] For instance, mRNAs that encode the protein sequence, ribosomal RNAs (rRNAs) that constitute the ribosome and catalyze protein biosynthesis, and transfer RNAs (tRNAs) that carry the specific amino acid for incorporation in the polypeptide chain belong to the first category. Regulatory RNAs include small interfering RNA (siRNA) and microRNA (miRNA) that mediate the degradation of mRNA via RNA interference (RNAi) and therefore directly regulate gene expression. Other regulatory RNAs are long non-coding RNAs (lncRNA) that are involved in the formation of subnuclear bodies called paraspeckles, in epigenetic modification and chromatin remodeling regulation e.g., during X-chromosome inactivation.[16–18]

Transcription of a gene is initiated by binding of TFs to a sequence upstream of the transcription start site (TSS), the so-called promoter sequence. TFs recognize specific sequence elements of the promoter, like AT-rich TATA-box motifs, and are key transcriptional regulators. They trigger the formation of the pre-initiation complex via recruiting of RNA polymerase II which then facilitates the synthesis of the precursor mRNA (pre-mRNA) starting at the TSS. The synthesis of RNA by RNA polymerases occurs in similarity to the DNA polymerase mechanism.[19] There are several ways in which the pre-mRNA is further processed before it is exported into the cytoplasm (Figure 6a). Already during early strand synthesis, a methylated guanine moiety is added to the 5′ end via the orchestrated functions of phosphatases, guanyl transferases, and guanine-N7 methyltransferases. This 5′ methyl cap is recognized by several proteins and is an important signal for distinguishing the mRNA from other RNA molecules and for further pre-mRNA processing.[20] The pre-mRNA contains protein-coding regions (exons) that are interspaced by non-coding regions (introns). For functional mRNA generation, the introns are removed and exons are joined in a process called splicing. The assembly of a whole protein and small nuclear RNA (snRNA) machinery, the spliceosome, is required to catalyze the transesterification of the exons under ATP hydrolysis. Splicing has the advantage that many different protein variants can be encoded in a single gene and are accessible upon alternative combinations of introns and exons in the mRNA – a process referred to as alternative splicing.[21] After splicing the 3′ end of the mRNA is marked by a poly adenine (poly-A) tail which is synthesized template-independently by the poly-A polymerase. The poly-A tail protects the mRNA from degradation and, together with the 3′ methyl-guanine cap, marks the mRNA as completely processed and mature.[20,22] Besides the methyl-guanine cap and the poly-A tail, the mature mRNA consists of three basic components a 5′ untranslated region (UTR) the coding sequence (CDS), and the 3′ UTR (Figure 6b). The 5′ UTR contains regulatory elements of translation, whereas the 3′ UTR harbors elements for mRNA stability and localization.[23] Hence, the mature mRNA is marked to be exported into the cytoplasm where it functions as a template for the translation into proteins.[22]

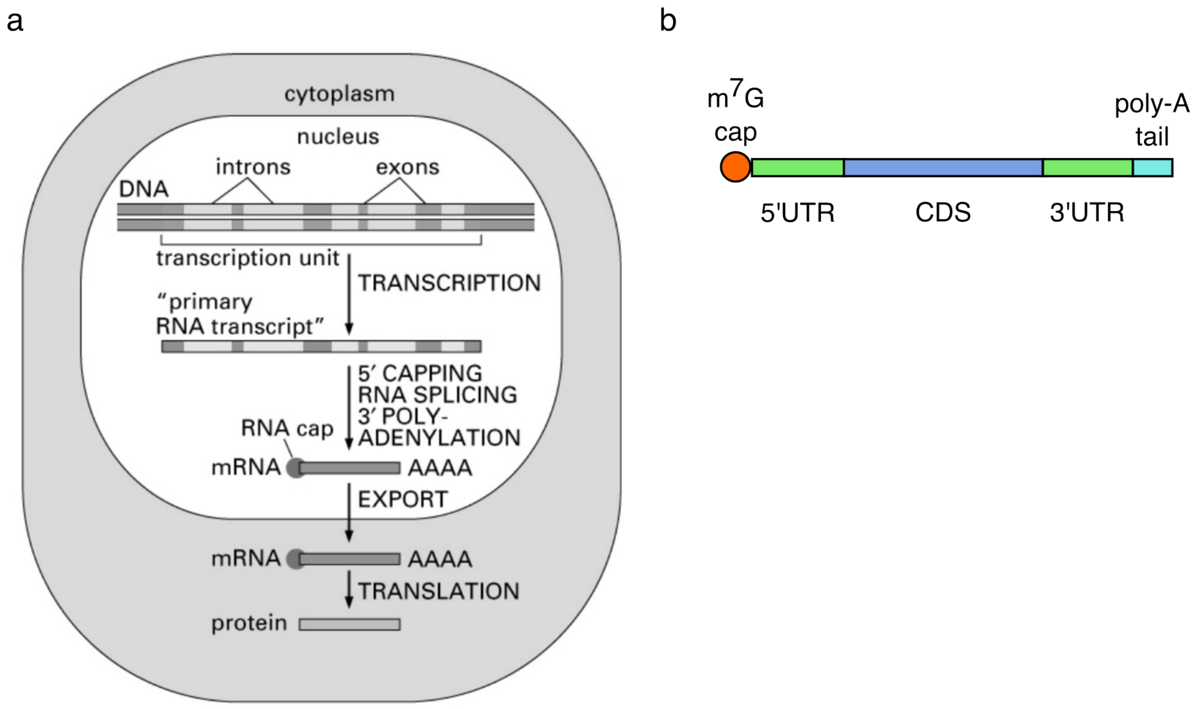

Figure 6: mRNA maturation in eukaryotic cells. (a) Schematic depiction of the mRNA processing in eukaryotic cells. During transcription, the pre-mRNA is marked as such by the addition of a 3′ methyl guanine cap. After removal of introns via splicing, the 5′ poly-A tail is introduced and the mRNA is exported to the cytoplasm. The figure is adapted from Alberts et al.[24] (b) Schematic structure of mature mRNA in eukaryotic cells containing a methyl-guanine cap (m^{7}G cap) and the poly-A tail, a 5′ and 3′ UTR and the CDS.

8.1.5 Organization and elements of the eukaryotic genome

8.1.5.1 Chromatin organization in eukaryotes

The human diploid genome contains over six Gbp that are allocated on the 23 chromosome pairs of chromosomes 1 to 22 and the sex chromosomes X and Y which together constitute the chromatin. It requires effective packing strategies to allow the chromatin, which is around two meters in total length, to fit into the nucleus of 10 µm in diameter. The packing is facilitated on multiple structural levels, the smallest being the formation of nucleosomes which are structural components of the chromatin in which the gDNA is wrapped around basic proteins, the histones (Figure 7a).[25]

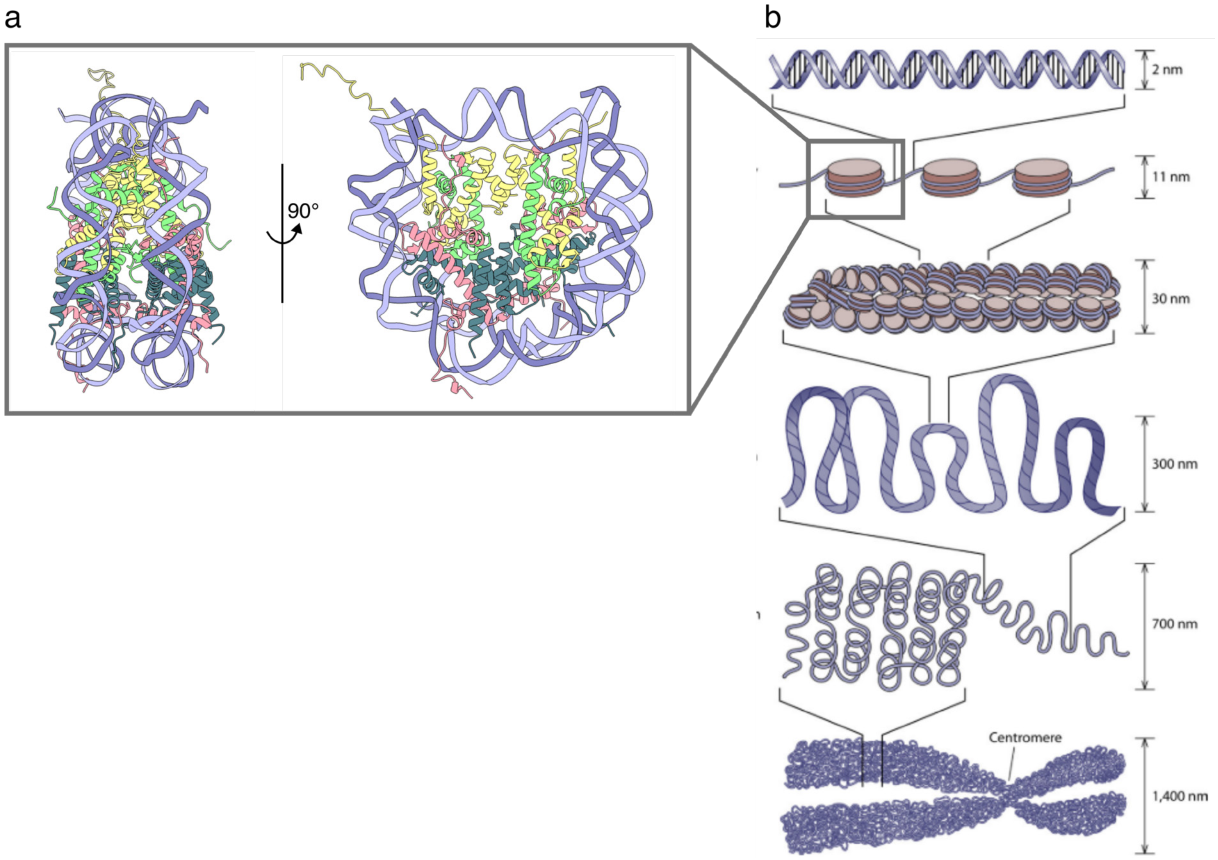

Figure 7: Nucleosomes and chromatin condensation. (a) Crystal structure of the *Xenopus laevis* nucleosome core particle consisting of 146 bp long dsDNA (purple) and histone dimers of H2A (red), H2B (dark green), H3 (yellow), and H4 (light green) as side and front view (PDB: 1AOI). (b) States of chromatin condensation starting from naked gDNA, continuing with 'beads on a string', chromatin fibers, and a packed mitotic chromosome. Figure adapted from Jansen and Verstrepen.[26]

Pioneering work on the structure of nucleosomes was done in the late 1970s and 80s. Nucleosomes were found to consist of an octameric histone core containing a dimer of each histone H2A, H2B, H3, and H4. 147 bp of the gDNA are wrapped around the histones to form the nucleosome core.[25,27,28] The nucleosomes are lined up like 'beads on a string' with a 60 bp linker DNA in between to which the histone H1 is bound. The wrapping of the gDNA reduces the length by around seven-fold.[25] Further condensation of the chromatin is achieved by the formation of 30 nm chromatin fibers (Figure 7b). Depending on the packaging of these fibers to higher order structures an ultimate size reduction of 10^4 is achieved, for example during mitosis.[29] The dynamic and variable distances of nucleosomes allow for the regulation of the accessibility of the chromatin for transcription factors, polymerases, and other DNA-binding proteins. Therefore, the chromatin can be roughly classified into accessible regions, the euchromatin, and inaccessible regions, named heterochromatin. Whereas the first is characterized by loosely packed nucleosomes and transcriptional activity, the latter exhibits high condensation, is transcriptionally inactive, and can be found in telomeric and centromeric regions and the inactivated X-chromosome. The controlled and dynamic regulation of both chromatin states is referred to as chromatin remodeling and is facilitated by protein complexes that are discussed in detail in section 8.2.4. Furthermore, eu- and heterochromatin contain specific epigenetic chemical modifications of histone tails and cytosine nucleobases whose specific functions are discussed in section 8.2.[30] Altogether, the dynamics of the chromatin packaging facilitate region-specific chromatin accessibility and condensation states thereby controlling cellular transcription specific to the cellular state.

8.1.5.2 Composition of the human genome

The human genome project was an international biological research project that was launched in 1990 to unravel the whole genomic sequence of human cells. First, it was anticipated that the human genome contains 100,000 genes. This expectation was disproven even before the project's initial completion in 2001. Surprisingly, the human genome only consisted of approximately 20,000 protein-coding genes, accounting for 1.5%-2% of the genome, while the majority of the genomic sequence was first referred to as “junk” DNA.[31] Nowadays it is known more about the relevance of this, falsely annotated, “junk” that has been found to regulate transcription in a complex manner (Figure 8). Interestingly the ratio of the non-coding and the coding part of the genome does not coincide with the organism's complexity and can even differ significantly between related species. This phenomenon is known as the 'C-value paradox', whereby the C-value represents a unit for the genomic size.[32]

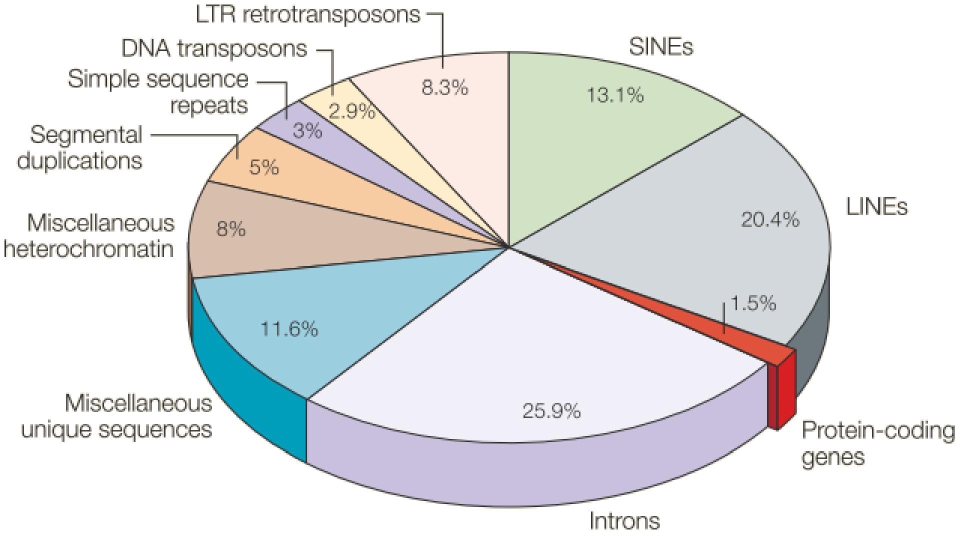

Figure 8: Composition of the human genome. Protein coding sequences (exons) represent the smallest fraction of the human genome. The majority is comprised of introns, retrotransposons, and repetitive elements. Figure adapted from Gregory based on data from the human genome project.[31,33]

More than half of the human genome is composed of repetitive regions. These comprise simple sequence repeats of one to six base pairs or transposable elements, such as long tandem repeat (LTR) retrotransposons, long interspersed elements (LINEs), short interspersed elements (SINEs), and DNA transposons.[34] Transposable elements can excise themselves (DNA transposons) or generate an RNA duplicate, reverse transcribe (LTR retrotransposons, SINEs, and LINEs), and reinsert themselves in new genomic sites. The origin of LTR retrotransposons lies in the insertion of the genetic material of retroviruses while the evolution of the other retrotransposons is poorly understood. All transposable elements act as genomic parasites without fundamental roles in the biology of the host, but have been and still are a relevant source of genomic variation, although DNA transposons are relatively inactive in mammalian cells by now.[35–37] Due to technical limitations, segmental duplications are one of the last genomic elements that have been sequenced. They represent DNA blocks of one to 400 kbp that occur multiple times in the genome with high sequence homology.[38] Segmental duplications have been substantial for the evolutionary development of the human brain and alterations of them can be responsible for developmental delay and diseases like autism.[39]

The major fraction of the non-repetitive genome is composed of introns that are over ten times more frequent compared to the protein-coding exons. Further non-coding elements contain functional non-coding RNAs, such as miRNAs, siRNAs, and lncRNAs, pseudogenes, which are non-functional copies of coding genes that can exhibit regulatory functions in gene expression, and additional regulatory elements of gene expression like promoter and enhancer sites.[40,41] The roles and functions of promoters and enhancers are described in more detail in the next section.

8.1.5.3 Elements of transcriptional regulation in eukaryotes

The effective and correct execution of biological processes within a cell or organism requires precise spatiotemporal control of gene expression for which transcriptional regulation is essential. Besides regulation via RNAi or lncRNAs, which were briefly mentioned before, transcriptional regulation involves gene regulatory regions. The regions can be divided roughly into promoter regions and upstream distal regulatory elements. Promoters are located directly upstream of the gene and are typically less than 1 kb in total length. They are composed of a core promoter region which is located at the TSS and a proximal promoter element. Distal regulatory regions can either be located several hundred base pairs upstream of the core promoter, located downstream in introns, or downstream of the whole gene (Figure 9a).[42]

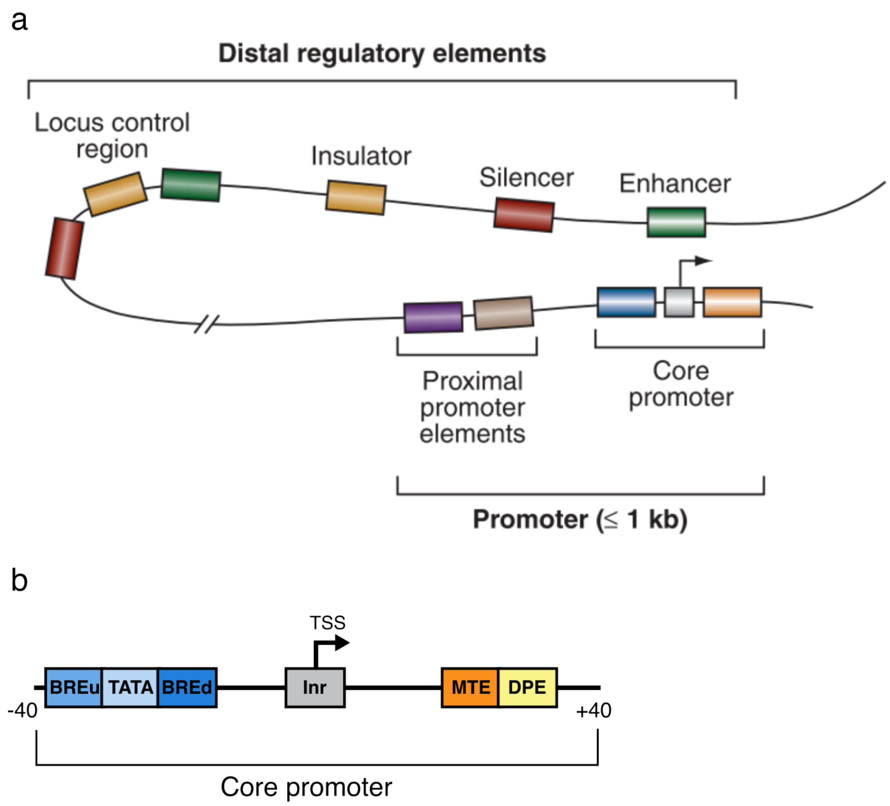

Figure 9: Composition of gene regulatory regions. (a) Schematic depiction of the positioning of promoters and distal regulatory elements in the eukaryotic genome. (b) Subelements of the core promoter consisting of upstream and downstream TFIIB binding elements (BREu, BREd), TATA box, Inr, MTE, and the downstream core promoter element. The figures are adapted and modified from Maston et al.[42] and Kugel and Goodrich[43].

The core promoter marks the TSS and reflects the establishment point for the pre-initiation complex (PIC). The PIC is set up of general TFs (GTF) that bind to the core promoter sequence and recruit the RNA polymerase II to the TSS. One of the most prominent GTFs is TFIID which binds the core promoter by selectively recognizing its TF binding, a small consensus sequence of 4-6 bp, which is the TATA box in the case of TFIID.[42,44] This sequence specificity is facilitated by the DNA binding domain of the TF. The core promoter contains multiple sub-elements that are recognized by the transcriptional activators. These include for instance the TATA box, initiator elements (Ire), and motif ten elements (MTE, Figure 9b). Ires can be found in over 50% of human core promoters and are located directly at the TSS, whereas MTE is located downstream of the TSS.[43,45] Proximal promoters are located directly upstream of the core promoter and contain several TF binding sites. 60% of the proximal promoters are located nearby or within a CpG dinucleotide-rich region (CpG island) that can dynamically modulate the binding of methyl binding domain (MBD) proteins and TFs as well as chromatin accessibility via its epigenetic modification state.[42]

Enhancers, silencers, insulators, and locus control elements represent distal regulatory elements of gene expression. Enhancers are modular long-distance regulatory elements of TF binding site clusters. They are brought in proximity to the promoter via loop formation between both chromatin regions – a mechanism called 'looping out'. There, enhancers exhibit similar functions as the promoter itself by recruitment of activating TF. Silencers act similarly to enhancers but bind repressive proteins that block transcription.[42] Insulators prevent distal regulatory elements from acting on other nearby genes by interfering with the looping mechanism of enhancers. Furthermore, they mark chromatin barriers that prevent transcriptionally inactive heterochromatin from spreading into transcriptional active euchromatin.[46] Locus control elements contain multiple regulatory elements that orchestrate the spatiotemporal transcription regulation of multiple adjacent genes.[42]

The binding of activators or repressors to these elements as well as the recruitment of coactivators or corepressors via protein-protein interactions (PPI) synergistically controls the expression of the underlying gene. As already mentioned, epigenetic chromatin modifications represent another level of complexity in transcriptional regulation and often involve modifications of CpG islands overlapping with the regulatory regions.[47] The epigenetic modifications and their effects will be illustrated in section 8.2.

8.2 Epigenetic marks in the eukaryotic genome

As introduced in section 8.1.5.3, the binding of activating or repressive TFs to genetic control elements, such as promoters or enhancers, regulates gene transcription and therefore the spatiotemporal protein expression within a cell or organism. Surprisingly, it was found that transcriptional regulation via TF is not reversible for all phenotypic characteristics of a cell. This implicated the existence of an additional regulation layer that maintains the cellular differentiation state. Today we know that the underlying elements of this layer are chemical modifications of the chromatin. These chemical marks are attached to cytosines and histones and are referred to as epigenetic modifications.[48,49] The modifications shape the chromatin structure and regulate the transcriptional activity of genes during and after cellular differentiation. The maintenance of these modifications throughout replication and cell division allows for an inheritable epigenetic memory that transfers the information of the differentiation state to the daughter cells. The term epigenetic refers to the storage of information on top of the genetic sequence ("epi" ≙ on top) and was originally defined by Riggs, Martienssen, and Russo in 1996 as "The study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence".[50] Although, many new details and roles of epigenetic marks have been unraveled since then, this definition still covers the central basics of the field.

8.2.1 Epigenetic marks of cytosine

In 1975 Riggs[51], Holliday and Pugh[52] independently proposed that cytosine methylation is involved in developmental processes and acts as an epigenetic mark. Since then, further cytosine modifications and epigenetic modifications of other DNA and RNA bases have been discovered to exist in all domains of life. The most frequent DNA modification in eukaryotes is the methylation of cytosine at the fifth position (5mC). 5mC is majorly found in palindromic CpG dyads of which 60-80% are found to be methylated in human somatic cells.[53] In general, gene bodies of highly expressed proteins and repetitive regions like transposons are hypermethylated whereas CpG islands near promoter regions are usually hypomethylated.[54] There, 5mC can reshape the binding affinity of TFs and MBDs, which are components of the chromatin remodeling machinery. Besides the modification of cytosine, the methylation of adenosine at the sixth position (6mA) has first been discovered in bacteria but has been found in the mammalian genomes as well. Studies indicate that 6mA might be involved in nucleosome positioning and transcriptional regulation in eukaryotes.[55] However, 5mC represents a key modification of the epigenome and regulates crucial processes like genomic imprinting (e.g., during X-chromosome inactivation), transcriptional regulation, and chromatin remodeling which will be described in section 8.2.4.

8.2.2 Regulation of epigenetic CpG marks – writer and eraser enzymes

The attachment of the methyl group on cytosine in CpG dyads is catalyzed by DNA methyl transferases (DNMTs). The group is transferred from the methyl donor S-adenosyl methionine (SAM) to the fifth carbon of the nucleobase via a mechanism that involves catalytic cysteine and glutamate residues of the writer enzymes (Figure 10).[56]

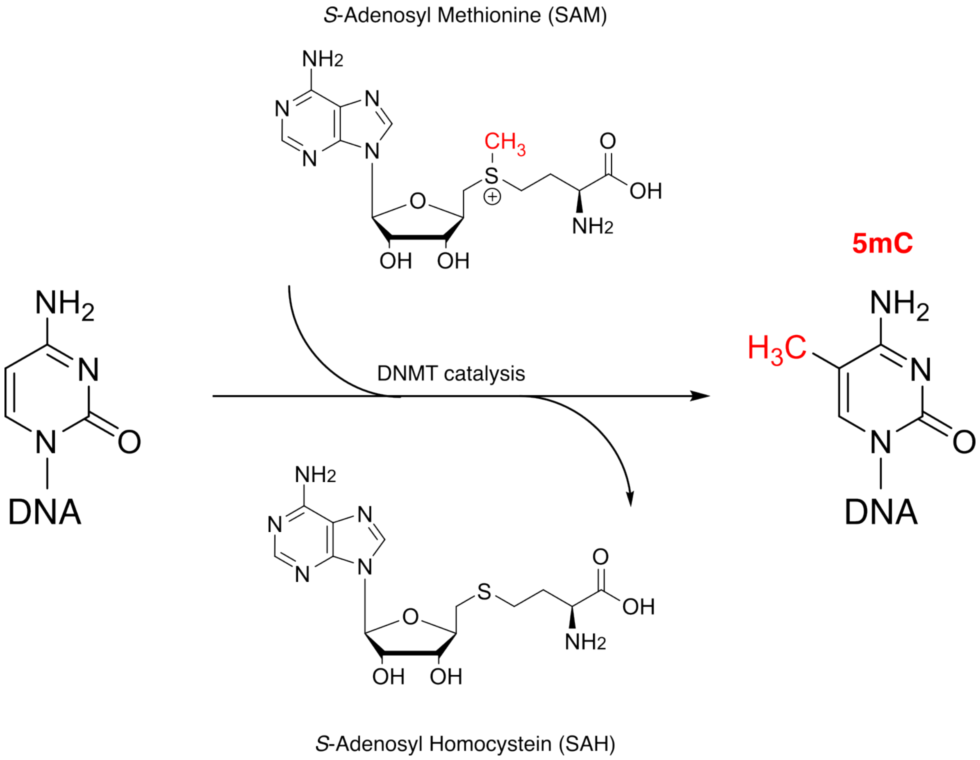

Figure 10: Reaction scheme of cytosine methylation at carbon 5 catalyzed by DNMTs using SAM as a methyl donor. [56]

Human cells express DNMT1, 2, 3A, and 3B, whereby DNMT1 is responsible for the maintenance of 5mCpG patterns after replication, 3A and B mainly introduce novel 5mCpG sites and DNMT2 lacks any DNA methyltransferase activity but functions as tRNA methyltransferase. After replication, the methylation mark only remains on the parental strand of the CpG dyad leading to hemimethylated CpG. DNMT1 preferentially binds to hemimethylated DNA and catalyzes the methylation of the daughter strand which preserves the epigenetic information between generations of differentiated cells. The specific interaction of DNMT1 with hemimethylated CpGs is facilitated by its target-recognition domain (TRD) and further via its recruiting by UHRF1 which also recognizes hemimethylated DNA. DNMT3A and B are mainly expressed in undifferentiated cells where they de novo generate new methylation sites at previously unmethylated CpGs. Misregulation of these enzymes often leads to developmental defects as new epigenetic patterns have to be introduced during the differentiation of the cell.[57]

The embryonal development requires the removal of parental methylation patterns. As already mentioned, methyl CpG marks are removed during the replication process if not maintained by DNMT1. This way of removal is observed in the maternal genome after fertilization – a process called passive dilution. The methylation pattern of the paternal genome is removed actively, catalyzed by eraser enzymes, and is referred to as active demethylation. During embryogenesis novel methylation patterns required for the developmental process are introduced by de novo methylation (Figure 11).

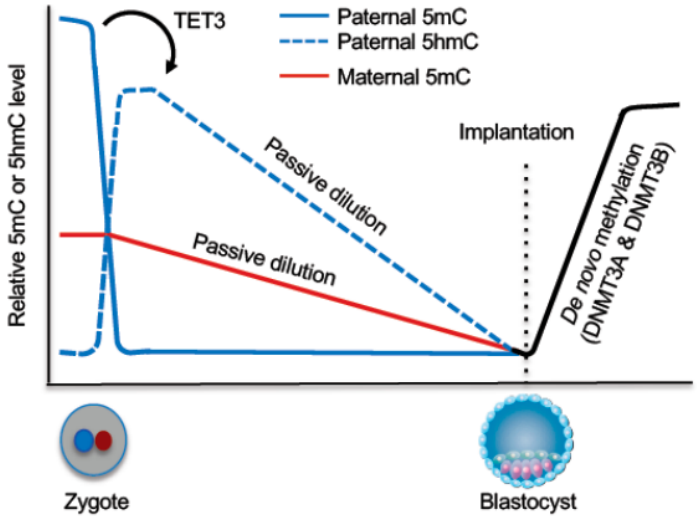

Figure 11: Changes in the genomic methylation level during development. After fertilization, paternal methylation patterns are removed by active methylation whereas maternal methylation patterns undergo passive dilution. Due to active demethylation via oxidation, 5hmC levels increase in the paternal genome. Figure derived from Zeng and Chen.[54]

Active demethylation occurs on a large scale after fertilization but also locus-specifically during further development. The eraser enzymes that catalyze the demethylation reaction are the ten-eleven translocation (TET) enzymes TET1-3 – a protein family of dioxygenases. They oxidize the methyl group of 5mCpG in a non-processive, step-wise manner to 5-hydroxymethyl- (5hmC), 5-formyl- (5fC), and 5-carboxyl cytosine (5caC) by using α-ketoglutarate and Fe(II) as cofactors. The two highest oxidized derivates of 5mC are selectively recognized by thymine DNA glycosylase (TDG) and excised from the DNA strand. The resulting abasic site is then repaired by the base excision repair (BER) machinery under the incorporation of an unmodified cytosine and restoring of the unmodified state (Figure 12a and b). It has to be mentioned that TET enzymes are indirectly involved in passive dilution, as the oxidized derivates of 5mC are not recognized by DNMT1 and therefore escape maintenance methylation after replication.[54]

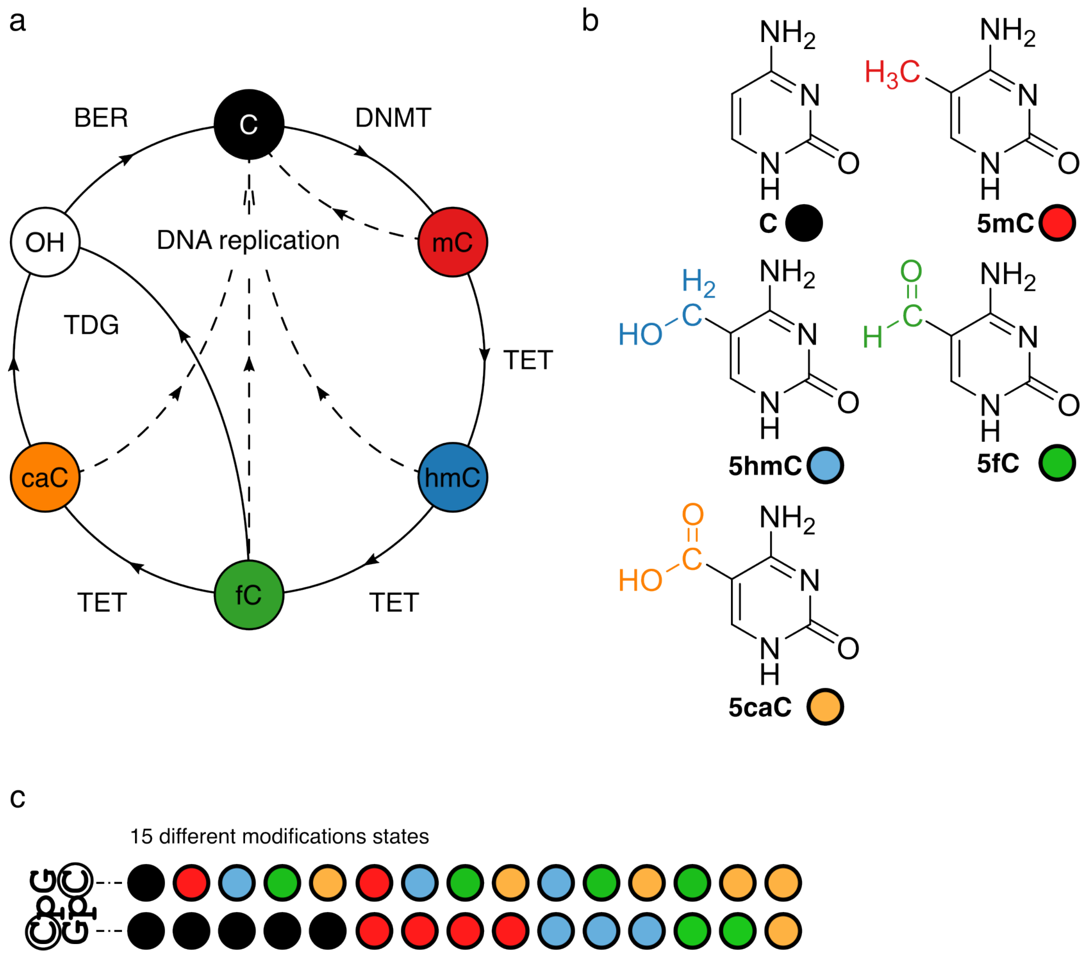

Figure 12: Turnover of epigenetic cytosine modifications. (a) Methylation by DNMTs and TET oxidation-facilitated demethylation at CpG dyads. Dashed lines represent passive dilution during DNA replication. (b) Chemical structures of cytosine and its modified derivates. (c) 15 different modification states at CpG dyads can exist due to the non-processive oxidation via TET enzymes. Figures or figure concepts adapted from Buchmuller et al.[58]

Due to the non-processive manner of the oxidation catalyzed by TET enzymes 15 different modification states, including the unmodified and four hemi-modified ones, can occur in CpG dyads (Figure 12 c). Interestingly, it has been found that the oxidized derivates do not only function as demethylation intermediate states but are rather stable epigenetic marks.[59,60] This makes TET enzymes not only erasers of 5mC but also writers of its oxidized derivates. The abundance of 5mC and other modified cytosine derivates differs between cell types and tissues whereby a general anticorrelation between abundance and oxidation state can be observed. 5hmC has been found to be a prominent mark in parts of the brain, accounting for 40% of all modified cytosine, and 5fC and 5caC are accumulated during the early stages of neuronal cell development.[61,62] These findings further underline the relevance of the oxidized derivates in cellular differentiation and maintenance.

Therefore, techniques for the detection and mapping of the epigenetic cytosine modifications within the genome were established. The first method for genome-wide mapping of 5mC was bisulfite sequencing (BS-seq) which utilizes bisulfite for the deamination of unmodified cytosines to uracil and identification of 5mC by combinatorial readout with an untreated sample after sequencing.[63] Since then, further methods for the combinatorial readout of the oxidized derivates have been developed and alternative approaches utilizing genome enrichment via modification-specific antibodies and reader proteins have been established.[64,65]

Another important way to unravel the function of CpG modifications is the understanding of their selective recognition by reader proteins which exhibits a layer of transcription and organizational regulation. These reader proteins often additionally recognize epigenetic post-translational modifications (PTMs) on histone tails. This crosstalk between the epigenetic modification types is exemplified on the basis of the chromatin remodeling complex NuRD in section 8.2.4. Similar to cytosine modifications, the attachment, and removal of histone PTMs are dynamically regulated by their own writer and eraser enzymes. A typical histone PTM is the acetylation of lysine (Kac) that negates the positive charge of the residue. This results in tighter binding of the histone tail to the DNA backbone which leads to chromatin decompression. Therefore, the acetylation of histone tails is generally considered to be a mark of euchromatin and transcriptional activity. Histone PTMs, including arginine methylation and phosphorylation of serine or threonine residues, occur in complex combinatorial signals that are referred to as 'histone code'. The code is read out and translated into signals for transcriptional regulation and chromatin remodeling by the reader proteins.[66]

In the following section, the modification selectivity and binding modes of the reader proteins of the epigenetic CpG modifications are discussed in more detail.

8.2.3 Reader proteins - mechanisms of the recognition of epigenetic CpG marks

The epigenetic information, stored in CpG dyad modifications, influences the interaction of various DNA binding proteins which facilitate the translation of the information into cellular functions. While so-called reader proteins show increased and selective binding to 5mC or its oxidized derivates, other proteins are repelled or not affected by the presence of the modifications.[67,68] Furthermore, some proteins have an affinity for or are repelled by multiple modifications or are influenced by hemi-modified DNA. These effects are often connected to the sequence context.[69] By now, reader proteins for all modifications of CpGs, whether symmetric or hemi, have been identified and their composition and expression levels seem to vary between cell lines and tissues (Table 1). However, the interactome of asymmetric CpG modifications is still unraveled. Interestingly, more reader proteins have been found for 5fC and 5caC than for 5hmC, although they are less abundant in human cells. The higher polarity or negative charge might cause the increased interaction with many proteins compared to 5hmC. Furthermore, 5fC can be covalently linked to proteins by forming a reversible Schiff base with lysine residues. It has been shown that 5fC can crosslink to histones in vivo and in vitro, suggesting that a covalent bond could further be involved in the interaction with 5fC readers.[70,71] In general, most readers of epigenetically modified CpGs are transcription factors, epigenetic writers or erasers, and further components of the chromatin remodeling machinery.[72,73] Insights into how the binding of these readers allows for maintaining or introducing changes in the epigenome as well as for transcriptional regulation during the developmental stages are given in section 8.2.4.

Table 1: Exemplary list of reader proteins of 5mC, 5hmC, 5fC, and 5caC identified by proteome screenings or in vitro studies. For all proteins the affinity profile for the modifications, if available within the CpG context, the cellular function, and the domain responsible for the binding are given. The listed proteins do not reflect the overall distribution of binders for each epigenetic cytosine modification.

Protein	Affinity profile	Cellular function	Binding domain
MBD1	mC/mC, mC/hmC, mC/fC[58]	Transcriptional repressor[74]	MBD [58]
MBD2	mC/mC[58]	Part of the NuRD complex[75]	MBD [58]

MBD4	mC/mC[58]	DNA glycosylase[76]	MBD [58]
MeCP2	mC/mC > (mC/hmC, mC/fC, mC/caC)[58]	Transcriptional repressor[77]	MBD [58]
Kaiso (Zbtb33)	mC/mC[78]	Pioneering TF[79]	C2H2 zinc finger [78]
CEBP α	mC, fC, caC > hmC[80]	Pioneering TF[67]	Leucine zipper [80]
OCT4	mC[81]	Pioneering TF [67]	Homeodomain [81]
UHRF1	mC/C > hmC/C > hmC/hmC[82,83]	Histone ubiquitin ligase[84]	SRA domain, NKR loop (hemi specificity) [82]
UHRF2	hmC/hmC > hmC/C >mC/C[82]	Ubiquitin ligase[85]	SRA domain [82]
TCF4	hmC[86]	Transcription factor[86]	bHlH domain [86]
MTA2	hmC[72]	Part of the NuRD complex[87]	GATA zinc finger domain[88]
FOXP1	fC[72]	Transcription factor[72]	Forkhead box [72]
FOXP4	fC[72]	Transcription factor[72]	Forkhead box [72]
ZSCAN21	fC[68]	Transcription factor[68]	SCAN domain[68]
TDG	caC > fC[89]	DNA glycosylase [89]	Catalytic domain [89]
MAX protein	caC[90]	Transcription factor [90]	bHlH domain [90]
MBD3	caC/caC [58]	Part of the NuRD complex[75]	MBD [58]

The most studied readers of 5mC are proteins of the MBD family – MBD1, 2, 4, and MeCP2. The MBD of MeCP2 recognizes 5mCpG dyads via a complex network of hydrogen bonds to water molecules within the major groove of the DNA that is shaped by the methyl groups. Additionally, direct hydrogen bonds to the guanines mediate the binding and recognition as shown in Figure 13a.[91] Despite the structural differences between the MBD of MeCP2 and the C2H2 domain of Kaiso, they possess the similarity to form direct hydrogen bonds to the guanines of both strands of the dyad within the major groove (Figure 13b). Furthermore, weak, noncanonical hydrogen bonds between acidic residues and the C5 methyl group are postulated for both proteins.[91,92] Additionally, the C2H2 domain forms a hydrophobic pocket via C505 and T507 in which the methyl groups of two 5mCs protrude. Further interactions of Kaiso with 5mCpG involve water-mediated hydrogen bonds and interaction with the DNA backbone in the minor groove.[78]

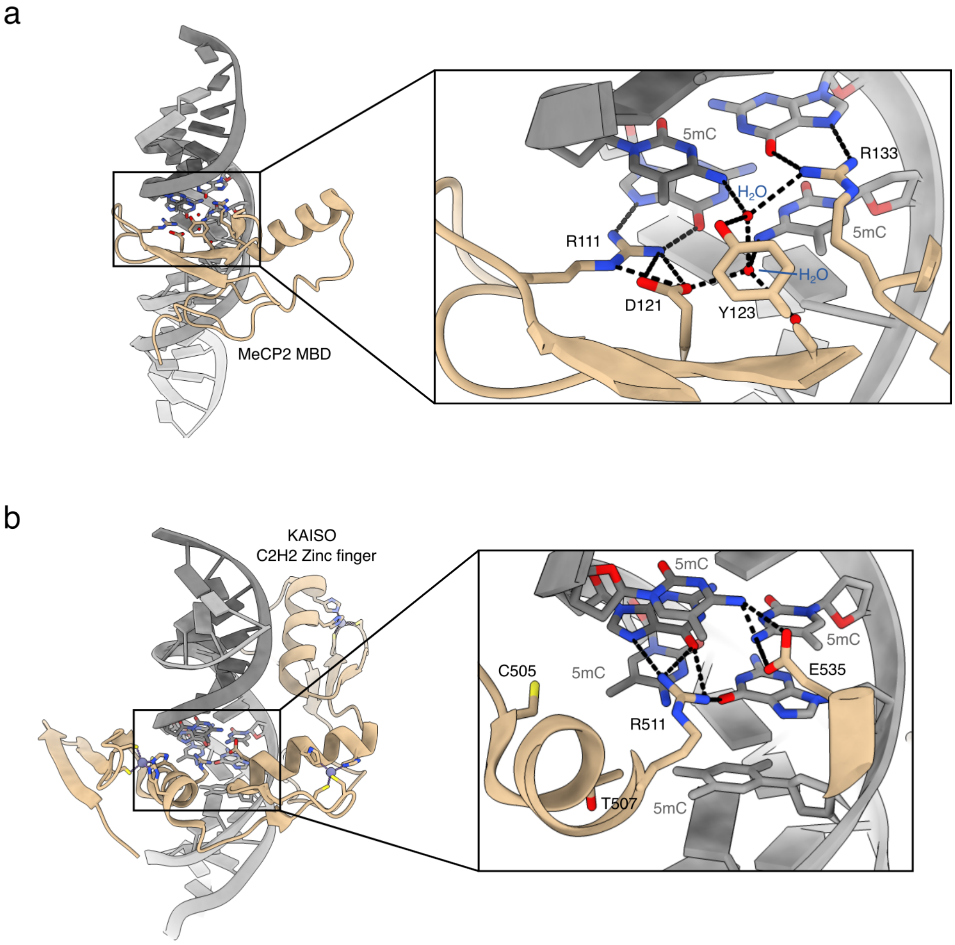

Figure 13: Major groove interactions of MeCP2 MBD and the TF Kaiso C2H2 domain bound to symmetrically methylated CpG. Hydrogen bonds between key residues and water molecules as well as the CpG nucleobases are depicted as dashed lines. All other water molecules, residue side chains, and their hydrogen bonds are omitted for clarity. (a) The MBD of MeCP2 recognizes the 5mCpG via hydrogen bonds between arginine residues and the guanines and by the formation of hydrogen bonds to a network of water molecules that are uniquely shaped by C5 methyl groups. PDB: 3C2I. (b) The Kaiso C2H2 zinc finger recognizes the 5mCpG dyad via hydrogen bonds of R511 and E535 and the methylated cytosines and guanines of both strands. Furthermore, C505 and T507 form a hydrophobic pocket that accommodates the methyl groups of two 5mCs. PDB: 4F6N.

Like Kaiso and the MBDs, many readers recognize epigenetic cytosine modifications via interactions in the major groove, where the nucleobases of the B-DNA helix are accessible. Often additional recognition of the narrowed minor grooved in 5mC is speculated as well.[93] Moreover, readers exist that bind to the modified base via a base-flipping mechanism. This is the case for the SET and ring finger associated (SRA) domain that can be found in UHRF1 and 2. In both proteins, the cytosines of the CpG dyad are flipped out of the DNA helix and inserted into the domain. UHRF1 exhibits affinity for hemimethylated and hydroxymethylated CpGs whereby 5mC is preferred. Variations of some residues in the SRA domain of UHRF2 allow the reversed binding profile compared to UHRF1 and the binding of symmetric 5hmC. These variations increase the binding pocket size and form hydrogen bonds between the hydroxy group of 5hmC and threonine and glutamine residues.[82]

All these examples illustrate that selectivity for certain modifications is often mediated by steric hindrance or a selective hydrogen bond network in the major groove. The change of DNA 3D structure upon methylation introduces an additional shape recognition mechanism that involves minor groove interactions and further contributes to the binding of reader proteins. [93]

8.2.4 Reader proteins mediate epigenetic regulation of genome organization and transcription

The general function of epigenetic modifications is the dynamic but inheritable regulation of transcription and chromatin packing. These fundamental and complex processes are orchestrated in a defined interplay of the cytosine and histone modifications, their writers and erasers, and the reader proteins. Reader proteins, like MBDs, are often part of chromatin remodeling complexes and guide the complexes to regions of certain epigenetic modifications. The complexes alter the chromatin packing and the epigenetic modifications and thereby change the accessibility for the transcriptional machinery. Often the recruiting of remodeling complexes is facilitated by a crosstalk between epigenetic histone and cytosine modifications like in the case of the 'Nucleosome Remodeling Deacetylase complex' (NuRD).

The NuRD complex is composed of readers, erasers, and chromatin remodelers that together facilitate the formation of heterochromatin by "nucleosome sliding", the movement of nucleosomes into closer proximity. In the following, the components of the NuRD complex and their functional interplay are described. Notably, the components of the complex can be substituted with some of their protein family members (Figure 14a).

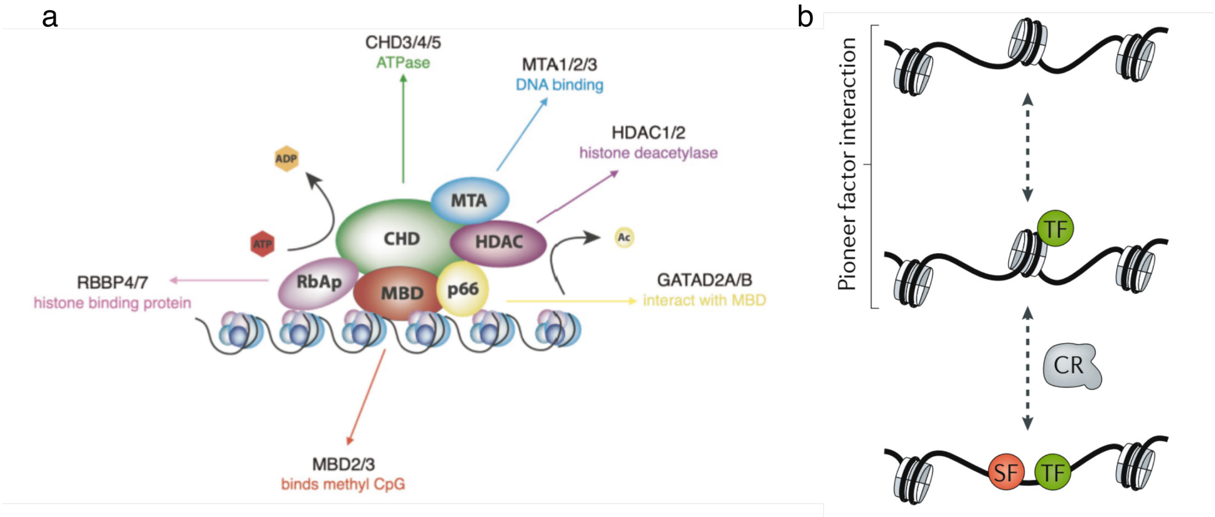

Figure 14: Chromatin remodeling and epigenetic modifications. (a) Composition and constitution of the NuRD DNA remodeling complex. (b) Mechanism of chromatin remodeling introduced by pioneering transcription factors. Figures adapted from Boulasiki et al. and Klemm et al.[94,95]

The core reader of cytosine modifications in NuRD is MBD2 or 3 which directs the complex to 5mC or 5caC marks on CpG islands.[58] Additionally, the MTA protein is also involved in the DNA binding and presumably also in the binding of 5hmC in the case of MTA2.[72] Furthermore, MTA is an important scaffold of the complex as it recruits HDACs, the erasers of histone acetylation. Lysine deacetylation leads to stronger interactions between DNA and histones in the nucleosomes and represses the binding of TFs and the transcriptional machinery. The RBBP proteins exhibit scaffolding function as well but also bind histones and are involved in mediating the deacetylation of histone lysines.[87,94,96] The central protein of the NuRD complex is of the chromodomain/helicase/DNA binding (CHD) subfamily. As intended by their name these proteins exhibit multiple roles: DNA binding, binding of repressive histone methylation marks (H3K9me3), and ATP-dependent helicase functions. The energy released from ATP hydrolysis is used to disrupt histone DNA interactions and to facilitate the "nucleosome sliding".[97] The NuRD complex exemplarily demonstrates the interplay between recognition of the histone code and cytosine modifications by readers, the altering of these modifications by erasers, as well as changing the overall chromatin packing by the chromatin remodelers. Expectedly, mutations or alterations in the expression of components of NuRD and other chromatin remodeling complexes can cause severe developmental diseases, genome instability, and carcinogenesis, e.g. by silencing tumor suppressor genes.[94,98]

In addition to NuRD and other proteins or complexes, the establishment of novel heterochromatin is usually facilitated by forward loops of repressive epigenetic marks that self-propagate to adjacent chromatin. Exemplarily, the histone methyl transferase Clr4 catalyzes the methylation of H3K9 and further binds to H3K9me via its chromodomain leading to the spreading of the heterochromatin mark. Misregulation can cause heterochromatin to spread into neighboring euchromatin which leads to false gene silencing.[99] Maintenance of heterochromatin is facilitated, amongst others, by the protein MeCP2 and UHRF1. UHRF1 selectively binds to hemimethylated DNA and recruits DNMT1 to restore symmetrically methylated CpG dyads after replication.[100] Like other MBDs, MeCP2 shows high affinity for 5mCpG, the mark of transcriptionally silent chromatin. After binding, MeCP2 recruits an HDAC-containing silencing complex that deacetylates nearby histones leading to further chromatin compaction and transcriptional silencing. MeCP2 is of further interest as its mutants exhibit higher selectivity towards 5hmC, which is a mark of active chromatin. This affinity alteration of the reader protein is thought to be one of the causes of the Rett syndrome, a neurodevelopmental disorder.[58,101]

For the establishment of euchromatin and the transcriptional activation of genes, chromatin remodeling complexes are recruited by a small fraction of TFs, the "pioneering TFs". They bind to repressive chromatin by recognizing nucleosomal DNA and methylated cytosines and induce the "opening" of the chromatin. This is achieved by the recruiting of chromatin remodeling complexes or epigenetic erasers and writers that induce epigenetic marks of active chromatin or loosen up the tightly packed nucleosomes of heterochromatin. After the nucleosome repositioning, secondary TFs bind and maintain the euchromatin (Figure 14b) which is accessible for further TFs and the transcriptional machinery.[95,102] However, it has to be mentioned that the detailed mechanisms of the chromatin remodeling mediated by pioneering TFs are poorly understood.

Overall, readers recruit further key players of the epigenetic or transcriptional machinery towards specific chromatin regions. This tight interplay of readers, writers, and erasers controls the development process and cell-fate maintenance by regulating eu- and heterochromatin states. The various developmental disorders and tumor types rising from misregulation and misinterpretation of epigenetic information demonstrate the importance of all these processes and the reader proteins in particular.

8.2.5 Platforms for the identification of epigenetic readers of cytosine modifications

As highlighted before, readers of epigenetic histone and DNA modifications are crucial for shaping the chromatin landscape and regulating transcriptional processes during development. Hence readers that are over- or underexpressed or mutated are often involved in carcinogenesis and other diseases. Therefore, it is of great interest to find novel reader proteins of 5mC and its oxidized derivates to improve understanding of the epigenetic machinery and to identify potential novel drug targets.

The MBD protein family was the first to be identified as 5mCpG readers and the MBDs are therefore the best-studied epigenetic readers so far. In 1989 the MeCP1 complex, which is nowadays referred to as NuRD complex[103], containing MBD2, was identified as a 5mC reader by the Bird group. They used methylated and unmethylated dsDNA probes as bait to identify selective interactors of 5mC from nuclear extracts – a general experimental setup that has been applied in various reader screenings since then.[104] Nowadays, the most common high throughput method for the elucidation of the interactome of modified CpG is the pull-down of reader candidates from nuclear extracts by incubation by binding to the probe and subsequent submission of bound proteins to fragmentation and identification via tandem mass-spectrometry (pulldown-MS/MS).[68,72,73] The data of modified and unmodified probes are compared for readers that were selectively enriched for modified CpG (Figure 15a). MS measurements can further be parallelized when SILAC media are applied.[73] The method offers comprehensive insights into the nuclear interactome of a cell line or tissue. So far, the interactomes of all symmetric CpG modifications have been investigated in mouse embryonic stem cells (mESC) and neuronal progenitor cells (NPC) by Spruijt et al. and Iurlaro et al.[72,73] They used artificial CpG-rich sequences or partial sequences of the Pax6 and Fgf15 promoters as probes, which provided no or low sequence diversity. Hence, certain sequence-specific binders may have been overlooked or might exhibit modification-specific consensus sequences. To reduce the influence of sequence specificity Bai et al.[68] created a single 2 kbp probe covering several TF-consensus motifs. Still, this sequence selection may restrict the binding of certain proteins that consensus sequences have not been covered. However, the analysis of the study solely focused on TFs and showed that the interactome of 5mC, 5hmC, and 5fC differs between cancer cell lines of different tissues. This explains certain variations in the interactome of these studies that used different cell lines and conditions. Notably, the majority of the found reader candidates from pulldown-MS/MS lack validation in orthogonal in vitro and in vivo assays. Orthogonal validations of readers often include the determination of dissociation constants via electrophoretic mobility shift assays (EMSA) or the identification of the cellular bindings sites by chromatin immunoprecipitation followed by sequencing (ChiP-seq). However, these studies have provided an overview of the interactome of 5mC and its oxidized derivates as well as the first insights into the reader affinity profiles.

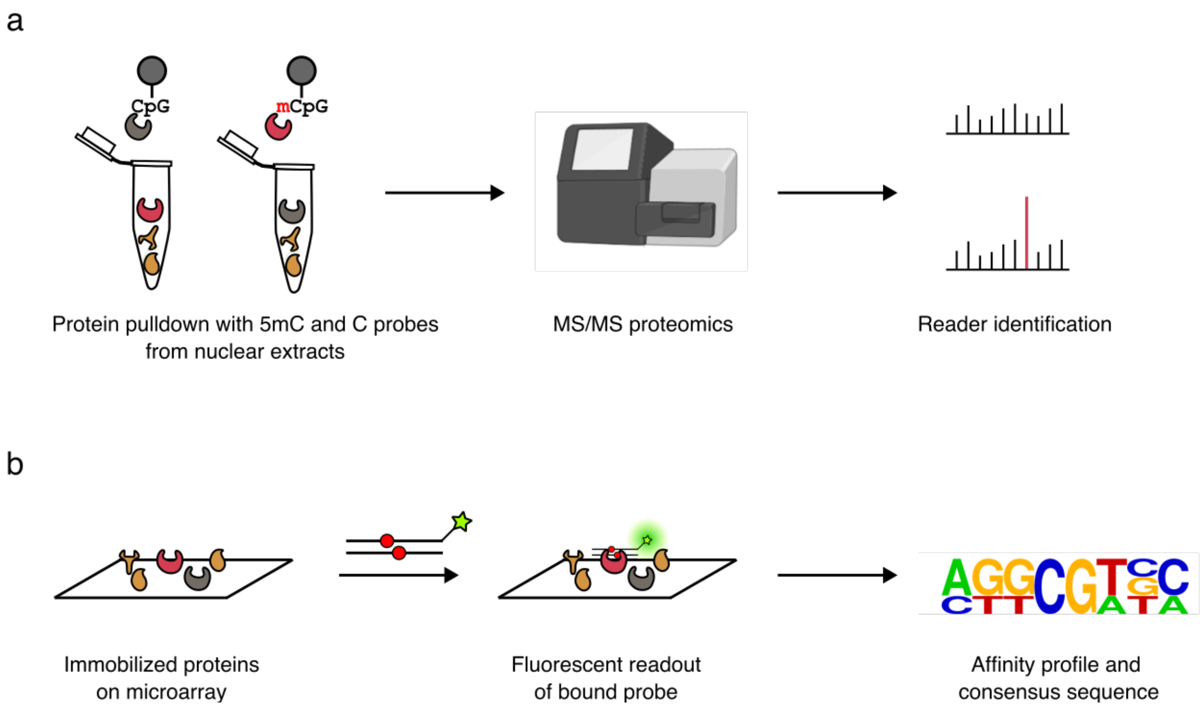
a
CpG
mCpG
Protein pulldown with 5mC and C probes from nuclear extracts
MS/MS proteomics
Reader identification
b
Immobilized proteins on microarray
Fluorescent readout of bound probe
Affinity profile and consensus sequence

Figure 15: Exemplary methods for the identification of epigenetic DNA readers. (a) Pulldown-MS/MS: Nuclear lysate extracts are incubated with the modified DNA probe. Bound proteins are identified via MS and readers are identified via comparison with the binders of the unmodified DNA probe. (b) Protein-microarray: Immobilized proteins are incubated with a fluorescently labeled dsDNA probe. Readout of the fluorescence signal of each probe allows for theidentification of reader proteins and their consensus sequences in the epigenetic context.

Protein microarrays represent an alternative method for high throughput interactome studies based on immobilized proteins on chips. Hu et al.[105] incubated microarrays containing over 1500 immobilized TFs and coactivators with 150 methylated and unmethylated CpG-motifs in a competitive binding assay. The data allowed not only for the generation of affinity profiles but also for first insights into methylated consensus motifs (Figure 15b). Protein microarrays comprise a comprehensive analysis of affinity profiles independently from cellular expression levels and in high throughput but are restricted to the preselection of candidate proteins. DNA microarrays reverse this principle, by incubating single proteins with immobilized DNA covering all sequences and the desired modifications. This requires the careful selection of a small number of reader candidates but allows for more accurate consensus motifs in the epigenetic context.[106]

Further methods for the investigation of the methyl consensus sequences and the affinity profile are based on systematic evolution of ligands by exponential enrichment (SELEX) in which sequences are iteratively enriched based on the protein's selectivity. High throughput methyl SELEX was developed by Yin et al.[107] and applies the regular SELEX setup to hundreds of TFs in parallel. Kribelbauer et al.[108] developed Epi-SELEX-seq in which the competitive binding of methylated and unmethylated probes can be investigated for individual proteins. Recently, dubbed digital affinity profiling via proximity ligation (DAPPL), an exhaustive method for the identification of binding contexts and modification affinities, has been developed. The human TF open reading frame (ORF) library was expressed, barcoded, pooled, and incubated with DNA libraries of random sequences carrying the individual cytosine modifications. By ligating the bound probes to the protein barcode, the consensus sequence in each modification context as well as the affinity profile could be determined for all TFs in a single sequencing step. In this study, Song et al. also investigated the affinity for hemi-modified CpGs that revealed significant differences compared to asymmetrically modified CpGs in the case of some TFs.[69]

Overall, pulldown-MS/MS methods have provided comprehensive insights into the interactome of epigenetically modified DNA of specific cell types and tissues but were restricted to the individual protein expression levels. Additional approaches focusing on TF libraries were able to cover all human TFs but are therefore intrinsically biased and not suitable for the identification of yet unknown dsDNA binders or reader candidates. Although these methods have likely unraveled the majority of reader proteins there are likely still open gaps that have not been covered due to methodological restrictions or experimental choices. So far readers of the asymmetric CpG modification combinations have only been investigated in the hemi context. Bacterial surface display mediated directed evolution of MeCP2 by Buchmuller et al.[109] showed that the chemical space for selective recognition of asymmetry exists in MBDs and it is up to future investigations to examine the existence of asymmetry readers of modified CpG dyads.

8.3 Cellular surface display

The cellular membrane of an organism functions as a semi-permeable layer for compartmentation and the creation of diffusion barriers and gradients. Furthermore, the surface proteome (surfaceome) of all species of life embodies a hub for a variety of functions that are crucial for interaction with the surrounding environment. Thereby, all surface-exposed proteins share the necessity to be anchored in the cellular membrane which is usually facilitated via glycolipid anchors or hydrophobic transmembrane domains.[110–112]

In *Saccharomyces cerevisiae* over 40 cell surface proteins have been identified that function in protein biosynthesis, stress response, cellular and cell wall organization, and carbon metabolism.[113] Moreover, surface-exposed proteins are crucial for cell-surface or cell-cell adhesion, of which the Aga1-Aga2 protein pair is a well-studied example that facilitates the adherence of mating.[114] In the outer membrane of the Gram-negative prokaryote *Escherichia coli* almost 100 proteins were found to be embedded via β-barrel structures and to be (partially) exposed to the cellular surface.[111] Type V secretory proteins, so-called autotransporters, represent a special fraction of these proteins as they contain self-secretory structures that allow transport across the outer membrane.[115] Amongst the autotransporter protein family, intimins, like EaeA from enterohaemorrhagic *E. coli*, and adhesins like AIDA-I from enteropathogenic *E. coli* are requisite for virulence by facilitating the display of the adhesion factor to the human cell on the surface of the pathogen.[116,117]

In addition to the ubiquitous presence of surface-exposed proteins in the domains of life and their variety of functions, protein surface display systems have been adapted to display heterologous proteins for biotechnological purposes.

8.3.1 Biotechnological applications of surface display platforms

The utilization of surface display for biotechnological purposes originated from the 'phage display' in the 1980s where George P. Smith showed the display of heterologous proteins on virion surfaces by expressing them as fusion constructs with a filamentous coat protein. This pioneering work was rewarded with the Nobel prize together with Sir Gregory P. Winter in 2018.[118–120] By now this technique has been further optimized and adapted to coat proteins of various bacteriophages.[119] Furthermore, the good and simple accessibility of surface proteins in assays and the well-established techniques for genetic transformation and editing have motivated the development of surface display systems of microbes like *E. coli* and *S. cerevisiae* in the following years.[121–123] There the heterologous protein is fused to an anchoring motif or carrier protein (e.g. intimin autotransporter domain) that facilitates the display. These techniques have primarily been used for screenings of proteins and peptides from native sources and rationally and randomly designed mutant libraries towards a property of interest (selective overview see Table 2).[119,124]

Table 2: Selective overview of display platforms used in biotechnology showing the sizes of successfully displayed heterologous proteins and exemplary applications in protein library screening.

Organism	Display platform	Het. Passenger sizes	Protein libraries
Phage/*E. coli*	Filamentous phage proteins	7-60 kDa	Nanobodies[125–127] scFV library[128–130] cDNA encoded proteins[131] Mutant library[132–134]
	Lytic phage proteins	Up to 115 kDa	Nanobodies[135,136] cDNA encoded proteins[137–140] Mutant library[132–134,141,142]
Yeast	Aga2p	7aa-96.2 kDa	Nanobodies[143,144] cDNA encoded proteins[145–148] Mutant library[134,149–152]

E. coli	Intimin	3-30 kDa	Nanobodies[153–155]
	AIDA	13 aa – 60 kDa	Nanobodies[153] Cyclic peptides[156] Mutant library[109]
	Others	9 aa – 50 kDa	scFV library[157] Mutant library[158–160]

Surface display techniques generally benefit from the omission of protein isolation and purification as well as the direct link between pheno- and genotype that allows for fast downstream analysis and iterative screening processes. Besides the screening of protein libraries, these technologies have also been successfully utilized as (medical) biosensors as well as in numerous additional biotechnological applications such as whole-cell biocatalysis and biofuel production.[161–164]

While phage, bacterial, and yeast surface display have been the most commonly used techniques, it has to be mentioned that other methods like mammalian cell surface display or cell-free methods like ribosome- or mRNA display have emerged and aim to overcome limitations of the established platforms in terms of PTMs, protein folding or library diversity.[119]

8.3.2 Phage display in biotechnology

Filamentous phage display represents the oldest surface display technique for heterologous proteins based on biological entities and has been used frequently for the screening of peptide and antibody libraries for selective binding. Hereby the heterologous protein is displayed on the virion surface as a fusion protein with a viral coat protein. Typically, the heterologous gene is inserted in-frame and upstream of the coat protein gene in the viral genome or a phagemid vector and is expressed and displayed as an N-terminal fusion protein. The number of displayed proteins is thereby dependent on the fusion partner and can extend up to 2700 copies in the case of the coat protein pVIII.[119] It was shown that this technique allows for the functional display of proteins and enzymes from a size range of 7 kDa (mustard trypsin inhibitor

MTI-2) up to 60 kDa (alkaline phosphatase).[165,166] The later developed bacteriophage display platforms such as the T7Select® (Novagen) developed by Rosenberg et al. are capable of displaying selected proteins of over 115 kDa (*E. coli* β-galactosidase) as a C-terminal fusion protein with capsid protein 10B.[167]

During the screening of displayed peptide or protein libraries, the phages are incubated with the immobilized target molecule, unbound phages are removed by washing, and phages with target affinity are eluted. This process is referred to as biopanning. Phage amplification in *E. coli* allows for multiple rounds of selection before analyzing the enriched genotypes via sequencing (Figure 16).

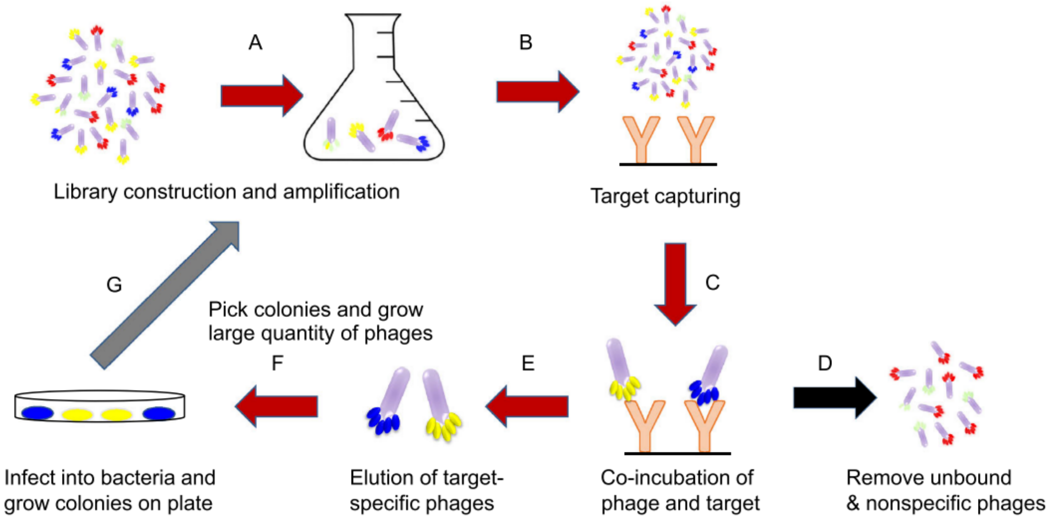

Figure 16: Phage display biopanning workflow. After library construction, the displayed protein library is isolated from *E.* coli and subjected to several rounds of enrichment for binders of the target molecule while unbound phages are removed by washing. Figure adapted from Saw and Song.[168]

Screening for high-affinity binders often requires a reduced copy number of displayed proteins to prevent compensation of lower affinity by increased valency. In bacteriophage-based display platforms, the strength of the promoter controls the valency of the displayed protein or peptide.[169] In filamentous phages, lowering the copy number is enabled by controlling the expression of the fusion protein by either substituting the coat protein gene or additional insertion of the fusion-protein gene in the phage genome. A way to ensure monovalent display is the transformation of bacteria with dsDNA plasmid vectors, so-called phagemids. Phagemids contain a secretion signal, the fusion protein sequence, as well as phage, and bacterial origins of replication, but require further genes of helper phages for single-strand replication and packaging in the bacterial host cell. Furthermore, phagemids can be generated using the repertoire of standard techniques for dsDNA editing. The high transformation efficiency of *E. coli* and the increased genetic stability make them the most common genetic system for phage display.[119,170]

The general advantages of the phage display are its great library diversity (>10^{10} variants) and fast viral reproduction which make it a powerful and rapid screening tool for selective peptide and protein binders for medicinal, biotechnological, and chemical biology purposes. One of the widest applications of phage display is the screening of diverse peptide libraries for target-specific interactors. It has been successfully applied for the identification of autoantigens, the selection of G4-quadruplex peptide ligands, as well as peptides targeting disease drivers such as the SARS-CoV-2 receptor binding domain and various genes involved in carcinogenesis like WDR5, which is part of the chromatin remodeling complex NuRD.[168,171–176] The fast screening cycles and the straightforward generation of diverse libraries have made phage display a useful tool for the directed evolution of proteins. These range from the engineering of single domains, like chromo domains, to the evolution of whole enzymes toward higher affinity or altered substrate specificity.[132,133,177,178] Furthermore phage display has made meaningful contributions to medicinal research, as it allows for fast selection of antibodies from natural IgM or IgG sources or synthetic antibody libraries with randomized variable chains and successfully yielded 14 FDA-approved antibodies for cancer and non-cancer disease treatment.[179–181] Phage display is not restricted to synthetically designed mutant libraries and can be utilized for the parallel display of a variety of proteins encoded in eukaryotic complementary DNA (cDNA) libraries. Due to the cDNA structure, this requires the C-terminal fusion to the coat proteins, such as filamentous phages pVI and T7 bacteriophages 10B, or the indirect display via the Jun-Fos dimerization system.[119,182,183] Amongst others, this helped to identify human autoantibodies and tumor antigens, as well as RNA and DNA binders.[131,137,140,184]

Despite the successful and effective application of phage display in a variety of fields, the platform presents disadvantages that lead to limitations in practice. The majority of these disadvantages apply to the original filamentous phage display where the disulfide bridges of heterologous proteins are reduced in the bacterial periplasm which can lead to misfolding. Moreover, the protein-decorated phages need to penetrate *E. coli*'s outer membrane which can fail for certain heterologous proteins. Both issues have been overcome with the display via lysogenic bacteriophages which fold the proteins in the bacterial cytoplasm but still face potential misfolding, lack of PTMs, size limitations of the heterologous proteins remain as well as the need for reinfection after each selection cycle.[119,169]

8.3.3 Yeast surface display in biotechnology

In yeast surface display, heterologous proteins, so-called passenger proteins, are displayed as fusion proteins with surface anchored proteins, so-called carrier proteins, on the cell surface. The fusion proteins are expressed at the endoplasmic reticulum (ER) and transported to the cellular surface by the vesicle trafficking machinery of the Golgi apparatus. Since the development of the method in 1997 by Boder and Wittrup, who used an Aga1-Aga2-based display platform, several surface proteins have been utilized for surface anchoring (Figure 17), although the original system is still the most commonly used. In the Aga2p system, the passenger can be fused N- or C-terminally to the Aga2 protein which is covalently attached to the surface-embedded Aga1 via two disulfide bridges. Aga1 itself is covalently anchored to the cell wall glycan via its C-terminal glycosylphosphatidylinositol (GPI) anchor.[144,185,186] The requirement of disulfide bonds between Aga1 and 2 has been shown to lower the display efficiency although still up to 10^5 molecules can be found on the yeast surface, depending on the heterologous protein or peptide.[185] Moreover, it has been shown that direct fusion to Aga1 can double the efficiency when spaced with an appropriate linker sequence and other GPI-anchored carriers like Sed1 or Flo1 have shown higher display efficiency compared to the Aga2p platform.[187,188] Further regulation of the display quantity can be achieved by optimizing the signal peptide sequence and the choice of the promoter.[164,189]

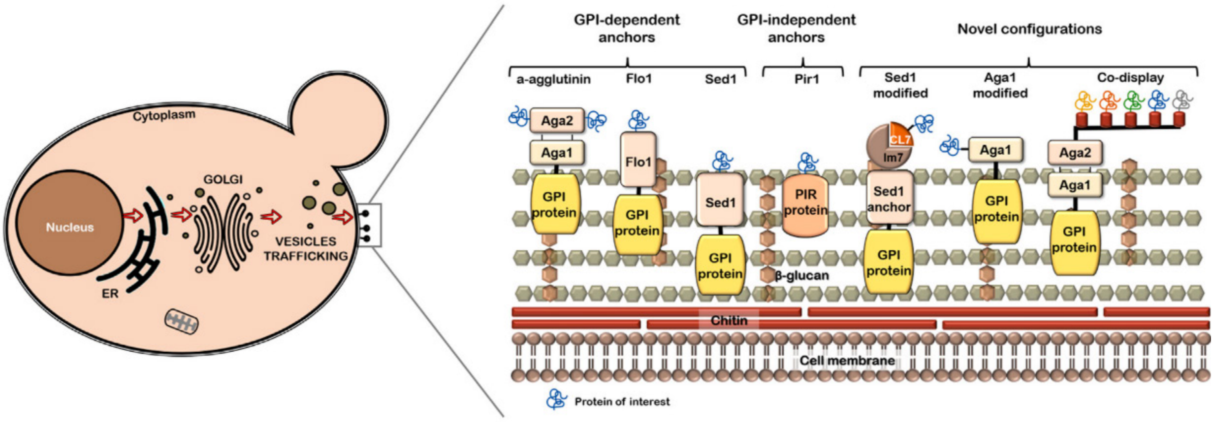

Figure 17: Mechanism of yeast surface display. Proteins are expressed in the cytoplasm and transported to the cellular surface via vesicle trafficking. A variety of yeast surface proteins has been utilized for surface anchoring of heterologous proteins. Figure adapted from Teymennet-Ramírez et. al.[164]

Typically, the fusion protein is encoded in an extrachromosomal plasmid vector which offers the benefit of employing standard and efficient cloning techniques. Taking this into account, the major limiter of the size of screening libraries is the transformation efficiency of yeast which results in library diversities a few magnitudes lower than with phage or bacterial display platforms and is usually limited to 10^7 individual clones.[189,190] Benefits of yeast display are efficient cDNA surface display, fluorescence-activated cell sorting (FACS) compatibility, and the capability of displaying proteins of up to 136 kDa (β-glucosidase) as has been shown in *Aspergillus aculeatus*.[145,146,191,192] Furthermore, the eukaryotic PTM machinery of yeast contributes to native protein folding and makes yeast surface display a favorable tool for many screening purposes.[164,193]

Screenings make primary use of FACS selection as it provides quantitative information and precise selection paired with online monitoring. There have been additional attempts to optimize the screening for low-abundant target proteins by combining FACS with magnetic-activated cell sorting (MACS).[194] Yeast surface display has been extensively used for the identification and affinity maturation of full-length IgG, Fab, Fv, and Fc fragments as well as nanobodies.[191,193] Further use includes the enrichment of binders from linear and cyclic peptide libraries for various targets. Moreover, yeast surface display has been applied to the screening of cysteine knots (knottin peptides) which offer rigid tertiary structures, thereby making use of yeast's disulfide isomerization capability.[195–199] Another field that has benefitted from yeast surface display is the engineering of whole proteins or domains by directed evolution, including the evolution of proteins for increased affinity or altered target selectivity, enzymatic enantioselectivity, as well as thermostability.[185,200] Apart from the selection of novel antibodies, the screening of full-length and random-primed, cDNA-derived libraries has found widespread application in the identification of antigens, readers of PTMs, and small molecule binders.[145–148]

The advantages of yeast surface display regarding efficient expression, folding, and the posttranslational modification of mammalian proteins, have made it a widely used tool in biotechnology.[119] Nevertheless, drawbacks include the prolonged culturing and expression times, as well as smaller library diversities compared to *E. coli*.[201,202] Furthermore, the introduced PTMs may not necessarily reflect the native mammalian modifications and could lead to altered functionality.[203]

8.3.4 Bacterial surface display in biotechnology

Bacteria were the pioneering microbes utilized as hosts for surface-displayed heterologous proteins. Like yeast surface display, they offer the advantage over phage display of being self-replicative and FACS compatible. Originally, it was shown that the phosphate-limitation-inducible outer-membrane protein (PhoE) of *E. coli* allowed the display of a 58 amino acid-long viral repetitive peptide on the cellular surface by inserting it in surface-exposed regions.[204] Since then many bacterial surface proteins have proven to be suitable and further optimized as anchoring motifs for heterologous passenger proteins in various bacterial hosts.[205–207] In Gram-positive bacteria, the passenger is either anchored to the cell membrane or the cell wall. Where the first utilizes fusion proteins with transmembrane domains or lipid anchor carriers, the latter e.g. requires a Sortase A recognition motif which is then cleaved, and the passenger protein is subsequently transferred and bound to the cell wall via an isopeptide bond.[124,208] In general, the surface display on Gram-positive bacteria has been used much less in biotechnology than on Gram-negative bacteria. There, heterologous proteins have been fused to more than 20 different carrier proteins which are translocated through the inner and the outer membrane of the bacterium and anchored in the latter. The translocation process is facilitated by proteins of the secretory machinery of the cell (e.g., the Sec translocon) which are located within the inner and outer membrane.[209]

The passenger protein can be fused N- or C-terminally, as well as inserted internally in the carrier protein.[124] Exemplarily, in the case of *E. coli* peptidoglycan-associated lipoprotein (PAL), the protein of interest is fused N-terminally to the anchor. Here, the amount of surface-displayed protein is limited by the fact that 70-90% of PAL is not facing the extracellular space which is why other N-terminal display techniques, such as AIDA-I, are presumably more efficient.[124,210,211] A prominent example of a C-terminal display platform is the *E. coli* Lpp-OmpA chimera. The trimeric chimera consists of an N-terminal single peptide derived from major lipoprotein (Lpp) fused to a fragment of the outer membrane protein A (OmpA) and the C-terminal passenger protein (Figure 18a).

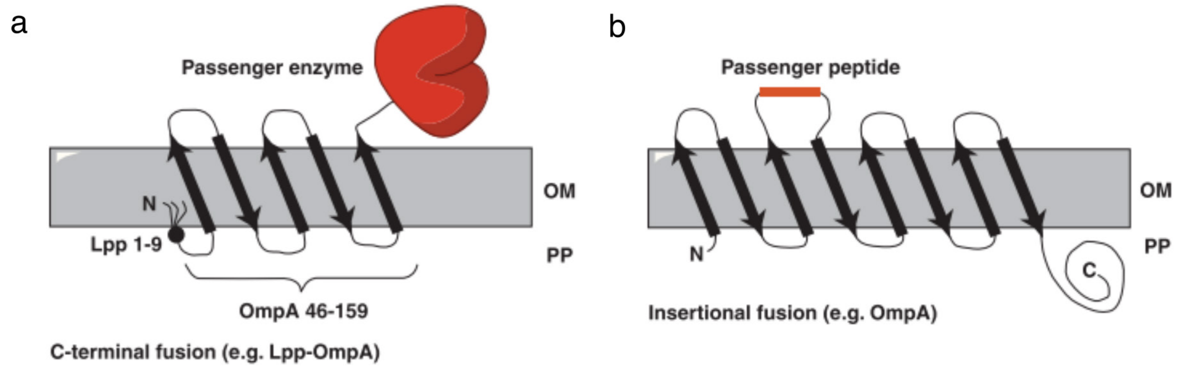

Figure 18: Exemplary systems of Gram-negative surface display in *E. coli*. (a) Display of the passenger by C-terminal fusion to the Lpp-OmpA chimera carrier. (b) Sandwich display via OmpA in which the passenger is inserted in a surface-facing loop between the β-barrel sheets. OM: outer membrane, PP: periplasm. Figure adapted from van Bloois et al.[206]

The signal peptide is required for the outer membrane localization of the chimera whereby the OmpA part integrates into the outer membrane by adapting a transmembrane β-barrel fold. Furthermore, OmpA is responsible for the translocation of the passenger protein.[212] Apart from N- and C-terminal display, sandwich-type systems, like the PhoE, OmpA, and OmpC LamB display, exist in which the passenger is inserted in a surface-exposed loop between β-barrel sheets of the carrier protein (Figure 18b). With only a few exceptions, sandwich-type systems are strongly limited in displaying passengers greater than 70 amino acids and have disruptive effects on the outer membrane.[124,213]

These and other display platforms of Gram-negative and positive bacteria are capable of displaying heterologous proteins and peptides such as epitopes, that have been found useful for immune response stimulation in antibody development.[214–218] Furthermore, displayed metallothioneins and other chelating proteins have been used successfully for the accumulation of toxic heavy metal ions from the environment.[219–222] Moreover, full-length enzymes up to the size of 50 kDa have been displayed to obtain further insights into the respective display mechanisms, as well as for whole-cell biocatalysis (e.g., β-lactamase, carboxymethyl cellulase, and hydrolases).[223–228] Besides single proteins, bacterial display platforms have mainly been used for the screening of diverse peptide libraries containing up to 10^{10} random peptides. Peptides are known to be displayed with high efficiency by these platforms and were successfully enriched for finding disease-specific antigens, recognition sequence profiling of kinases and SH2 domains, antimicrobial activity, and binding of zinc oxide.[229–233] Like the display on yeast and phages, the bacterial surface display has also been used for the directed evolution of enzymes by screening mutant libraries of carboxymethyl cellulases, lipases, and hydrolases for increased affinity, as well as for the affinity maturation of affibody, single chain variable element (scFv) and nanobody libraries towards various targets.[158–160] It should be mentioned that, in contrast to yeast and phage display, the display of full IgG antibodies has not been achieved with these techniques yet.

8.3.5 (Inverse) autotransporter-mediated bacterial surface display in biotechnology

Besides the discussed methodologies of bacterial surface display, the display in Gram-negative bacteria via autotransporters (AT), has emerged and is nowadays one of the most commonly used bacterial surface display platforms. More than 15 different ATs have proven to be capable of the display of heterologous proteins.[205,234] The basis for the popularity of ATs are the high display efficiency of over 10^5 molecules per cell, the straightforward genetic manipulation, and the transferability between Gram-negative hosts.[205,235] ATs consist of an N-terminal signal peptide, the natural passenger protein, like an adherence factor in the case of AIDA-I, a linker, and the C-terminal β-barrel fold domain (Figure 19a and b). The natural passengers can be replaced by the peptide or heterologous protein of interest. Like in almost all outer membrane proteins, the β-barrel structure serves as the transmembrane anchor and bears a central 1.2 nm wide pore.[236,237]

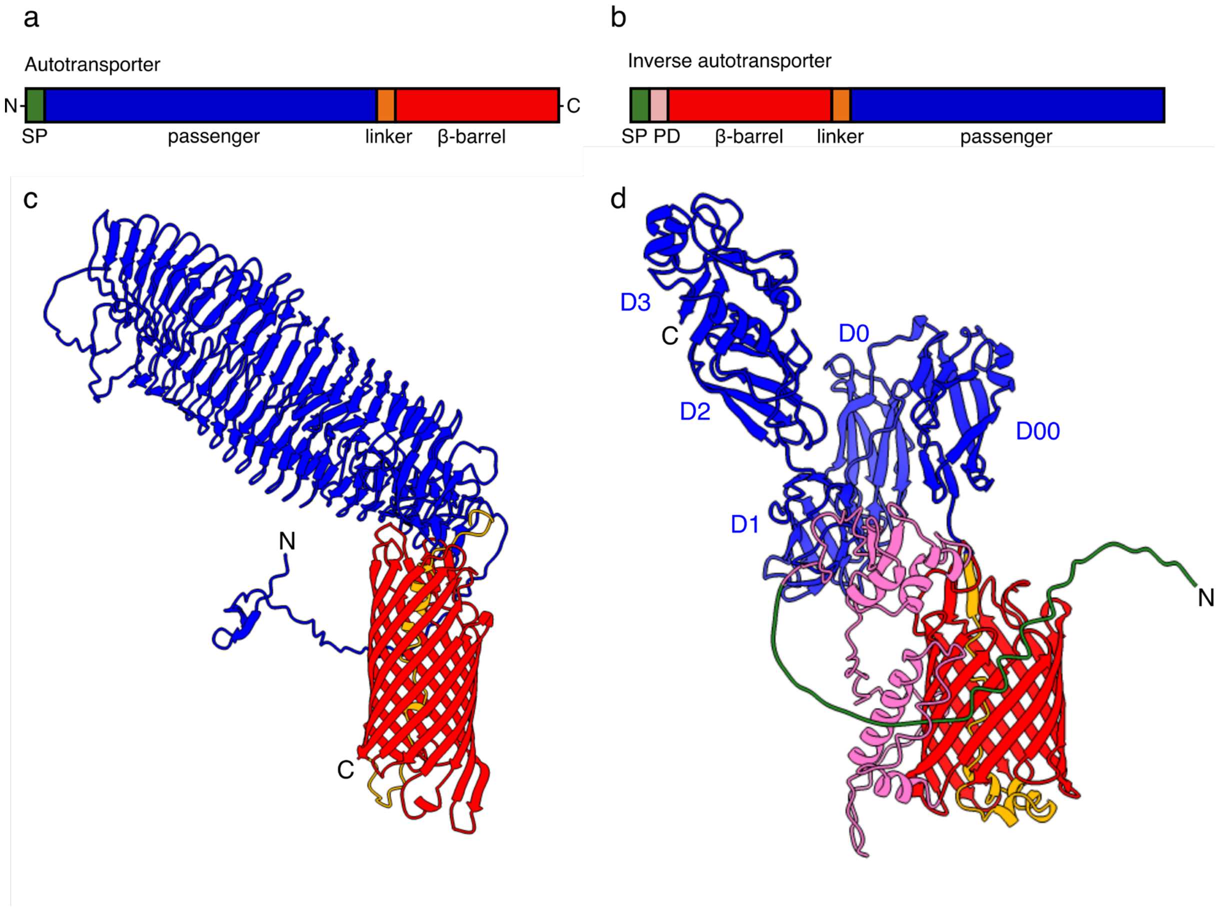

Figure 19: Domain structure of autotransporters and inverse autotransporters. (a) and (b) They contain the signal peptide (SP), a periplasmic domain (PD), a β-barrel domain, linker, and the passenger protein. (c) Alpha-Fold-predicted structure of AIDA-I with adherence factor (residues 1-954), linker (residues 955-1001) and β-barrel (residues 1002-1286).[238] ID: AF-Q03155-F1. (d) Alpha-Fold-predicted structure of EHEC EaeA intimin. I*n vivo*, the SP (residues 1-39) is cleaved off during periplasmatic translocation and the PD remains in the periplasm (residues 40-209). The β-barrel (residues 210-411) is embedded in the outer membrane and holds the linker (residues 412-450), extracellular immunoglobulin-like domains D00-D2, and the lectin-like domain D3.[239–242] ID: AF-P43261-F1.

ATs are expressed as single polypeptides and are targeted to the inner membrane by their N-terminal signal peptide sequence. There the Sec translocon enables the translocation of the unfolded protein into the periplasm and the signal peptide is cleaved off (Figure 20a).[236,243,244] Interactions with the β-barrel assembly machinery (Bam) complex facilitate the β-barrel fold of the C-terminal domain, its insertion into the outer membrane, and the translocation of the passenger. The β-barrel of BamA compresses the membrane and assists together with BamB-E in the folding of the AT domain. It has long been assumed that the unfolded passenger is translocated through the central pore of the ATs β-barrel, but there is evidence that the ATs C-terminal β-barrel domain forms an extended widened hybrid channel of 1.7-2.0 nm with BamA through which the passenger is transported (Figure 20b). This allows passengers to hold small structural elements such as disulfide-bridged loops and single α-helices during translocation. However, low structural complexity is still favored.[236,237,245–248]

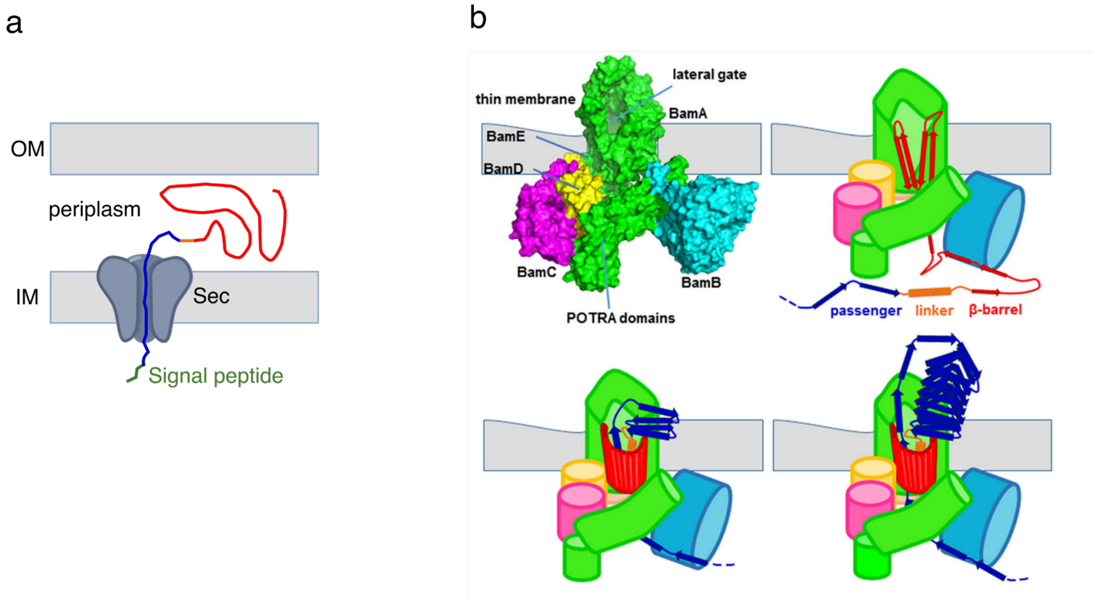

Figure 20: Secretory mechanism of autotransporters. (a) Translocation of the unfolded autotransporter into the periplasm via inner membrane (IM)-embedded Sec translocon. (b) Schematic representation of the structure of the Bam complex and the mechanism of autotransporter β-barrel assembly and passenger translocation through the outer membrane (OM). Figure (b) is adapted from van Ulsen et al.[236]

Additional involvement of translocation and assembly module (Tam) in the folding and membrane-embedding of the β-barrel is proposed, but it so far remains unclear whether it acts simultaneously with Bam or serves as a backup mechanism. It has also been shown that the importance and role of Tam varies upon culturing conditions, the investigated AT, and the bacterial strain.[237] The passenger secretion of ATs takes place in the C- to N-terminal direction and the extracellular folding is the central driving force for further of the rest of the protein. Furthermore, it has been shown that acidic residues promote the translocation of N-terminal protein parts which could explain the successful display of intrinsically disordered proteins.[236,247,249] In addition to the periplasmatic translocation, the outer membrane anchoring and the translocation to the surface represent crucial steps in AT-mediated surface display and are potential bottlenecks. The periplasmatic quality control machinery prevents misfolding of the β-barrel and promotes the translocation of the passenger via the chaperones SurA and Skp. Furthermore, nascent stalled ATs are removed by protease DegP degradation to prevent aggregation and leakage of the cellular membrane.[236,237]

Like the other bacterial display platforms, autotransporters have been used in various fields of biotechnological research such as immune response stimulation with surface-displayed epitopes for antibody generation.[250,251] AIDA-I and IgA1 display systems have primarily been used in whole-cell biocatalysis of lipids, via hydrolases, and dehydrogenases, in the development of assays for antibody diagnostics and inhibitor screenings, as well as cadmium ion accumulation. This proved the capability of ATs to display eukaryotic proteins of almost 60 kDa in their active conformation.[252–258] Due to their high display quantities, ATs have been a widely used system in protein and peptide library screenings.[115] Cathepsin G inhibitors have been enriched from an AIDA-I displayed random peptide library. Furthermore, nanobody and anticalin-mutant libraries of over 10^9 variants have been successfully subjected to affinity maturation against certain antigens.[153,259–261] Moreover, ATs have been utilized for the directed evolution of esterase mutants for altered substrate preferences, and MBD mutants for binding selectivity towards asymmetric CpG dyad modifications.[109,262] Besides the display on the cellular surface, ATs are also used for the secretion of expressed heterologous proteins and subsequent purification which has proven to improve folding, solubility, and stability of certain proteins as it avoids aggregation and degradation.[263,264]

Since the maintenance of an un- or less folded state of the passenger in the periplasm is crucial for translocation, the expression, and display in periplasmatic thiol-disulfide oxidoreductase (dsbA)-deficient strains has been proven advantageous in cases where a disulfide-bond-stabilized tertiary structure prevents successful secretion through the β-barrel pore.[265] Alternatively, the addition of reducing agents like 2-mercaptoethanol can prevent disulfide-bridge formation in the periplasm as well.[265,266] Furthermore, the utilization of outer membrane protease T (ompT)-negative strains was shown to be crucial for successful display via autotransporters, as it cleaves surface displayed proteins followed by the secretion into the medium. As described before, the secretion can also be desired for extracellular expression.[263,267] For passengers who are hardly or not displayed via ATs, the process of translocation has been circumvented by displaying an affinity moiety, e.g. avidin, and anchoring the heterologous protein to the surface via an e.g. biotin-affinity tag. However, via this technique, genotype and phenotype are unlinked which makes it unsuitable for the screening of protein or peptide libraries.[268]

Following the display, the extracellular folding of the heterologous passenger is essential for its activity. In the case of lipases, the co-display of foldases on the surface has been reported to be requisite for the correct folding and activity of the enzyme.[269,270] In general, the success of the display of heterologous proteins heavily depends on the combination of AT, passenger protein, promoter, signal peptide, and host and has to be further optimized based on empirical data. Furthermore, it should be mentioned that the display is rather limited by the structural complexity of the passenger than by its size.[205,245,271,272]

Inverse ATs, such as the EHEC and EPEC EaeA intimin, contain the same set of components as AT but in reverse order, except for their signal peptide which is alike located at the N-terminus (Figure 19b). Furthermore, their secretory and surface display mechanisms are similar to the regular ATs. Inverse AT passenger translocation occurs in N- to C-terminal direction and the role of TAM in the β-barrel assembly appears to play a more important role.[237] In the case of intimin surface display the natural C-terminal immunoglobulin-like and the lectin-like passenger domains D1- D3 (residues 659 to 934) are replaced by the heterologous protein (Figure 19d). At its first utilization by Wenzel et al., it was believed that the residues 450 to 550 were part of OM-embedded β-barrel but have later been found by Fairman et al. to form an additional extracellular immune-globulin-like domain D00 which is secreted before D0 and the heterologous passenger.[240,242,247] Like in all ATs and inverse ATs, the linker (residues 955-1001) spans the β-barrel cavity of the intimin AT in between the periplasmatic space and the surface located displayed passenger protein. The linker further widens the pore which may benefit the translocation, and its periplasmic part is assumed to block the cavity after successful translocation.[236,240]

Compared to ATs, like AIDA-I, the inverse AT intimin has been used less frequently for the display of single proteins and the screening of protein libraries. However, it has been shown that proteins within the range of 3 kDa (*Ecballium elaterium* trypsin inhibitor II) up to the size of almost 30 kDa (β-lactamase) can be successfully displayed in active conformation in quantities of up to 36,000 proteins per cell.[242,247] It has to be noted that for some passengers, including the β-lactamase, the periplasmatic folding and disulfide bridge formation had to be prevented by adding reducing or chelating agents as well as mutating cysteine residues to facilitate or improve the display.[247] Interestingly, the C-terminal fusion of the nanobodies by the intimin display system is beneficial compared to regular ATs in terms of surface display level stability and protein accessibility. This might be based on the additional D00 and D0 domains which increase the distance between the heterologous passenger and the cellular surface.[153] So far, affinity maturation of nanobody libraries, human fibrinogen, and the extracellular domain of human EGFR have been the only reported cases in which the intimin surface display has been used for library screening. However, the C-terminal fusion of the passenger would generally allow for cDNA surface display.[153,155,273]

In summary, AT and inverse AT surface display platforms are frequently used techniques for library screening and other biotechnological applications. They benefit from the general advantages of bacterial surface display including rapid cell division, high cell concentration in liquid culture, and increased library size. Furthermore, these systems allow for a high number of displayed proteins per cell which increases the fluorescence signals in FACS-based library sortings.[242] The ubiquitous presence of ATs and inverse ATs in bacteria, the simple and modular domain structure, and the variety of displayed structures simplify the adaption of the systems to new hosts and passengers. Although the effective library sizes are smaller than in phage display and the expression and posttranslational modification of eukaryotic proteins is less efficient than in yeast, (inverse) AT platforms combine some of the major advantages of phage display (high library diversities) and yeast display (self-replicative and FACS accessible) which make them a versatile tool in library screening.[154]

9 Aim of the Work

The primary objective of this study was the development and evaluation of an alternative screening tool for identifying readers of epigenetically modified CpG dyads. Current in vitro interactome studies utilize protein microarrays and ORF expression libraries, enabling the affinity profiling of the individual proteins.[69,105] However, the main method for proteome-wide reader identification is the pulldown of nuclear proteins followed by MS analysis.[68,274,275] This approach has provided valuable insights into the interactome of modified CpGs but bears its own limitations. It is constrained by the endogenous expression level of the proteins, the enrichment of indirect binders forming complexes with the readers, and might overlook readers of lower affinity due to competition for the DNA bait.

To address these limitations, our method should be based on the bacterial surface display of cDNA-encoded protein libraries and should serve as a platform to identify yet unknown reader and anti-reader proteins, thereby complementing the current approaches. The method should provide access to proteins that are typically lowly expressed in their originating cells without the requirement for purification. It should further allow for the enrichment of direct readers in a competition-free manner by displaying proteins individually on the cell surfaces. For this, surface display libraries originating from human cDNA, encoding full-length or fragmented proteins, should be created. These libraries should be subjected to selection by a FACS-based two-color assay, allowing the enrichment of proteins with defined on-target and off-target selectivity. Moreover, by screening protein fragments, we aim to obtain further information on the domains responsible for the interaction.

Initially, we will assess the general functionality of the intimin bacterial surface display platform for displaying human proteins encoded in cDNA. Ultimately, we aim to demonstrate the system's feasibility in enriching known and novel reader candidates of symmetrically methylated CpGs.

10 Results & Discussion

10.1 Initial decisions and prospects

We aim to develop a screening system for identifying protein readers of specific epigenetic CpG modification combinations. Therefore, the underlying key factors for an effective screening system are the library's quality and the selection methodology.

We opted for a FACS-based selection methodology that allows precise, simultaneous selection for on-target and against off-target probes by using distinct fluorophores. This strategy could prove particularly advantageous since it accounts for the affinity profile of readers towards multiple modifications in a single run, thereby enhancing the system's efficiency. Importantly, FACS offers real-time monitoring of the screening process which assists in the collection and evaluation of information for further optimization of the screening parameters.

Furthermore, the cellular surface display system grants access to protein libraries without the need for purification and provides a direct linkage between pheno- and genotype (Figure 21). We preferred bacterial surface display over yeast display, the most commonly employed display system for cDNA-encoded protein libraries, due to its similar compatibility for FACS, its additional rapid growth and the potentially higher library diversities. To the author's knowledge, bacterial surface display has not been applied to express cDNA-encoded protein libraries, but studies from phage display suggested that the heterologous expression of human protein libraries with the *E. coli* translation machinery is feasible.[140,184]

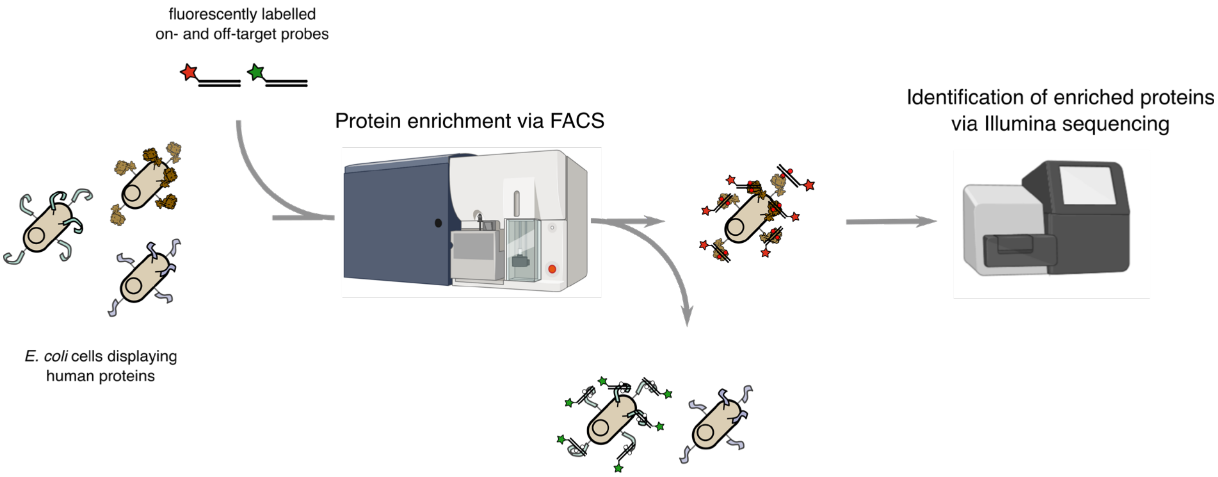
fluorescently labelled
on- and off-target probes
Protein enrichment via FACS
Identification of enriched proteins
via Illumina sequencing
E. coli cells displaying
human proteins

Figure 21: Schematic representation of the FACS screening workflow of bacterial surface-displayed proteins with on- and off-target selectivity. Illumina sequencing allows for the detection of the enrichment of low-abundant proteins.

We chose the intimin system from the family of inverse autotransporters as the display platform, which has originally been adapted for biotechnological usage by Wentzel et al.[242] The platform offers the genetic framework for C-terminal fusion of the passenger proteins, similar to the Aga2p system in yeast. In contrast to regular AT and other N-terminal display systems, this system facilitates the translation of the β-barrel anchor prior to the downstream cDNA-encoded passenger protein. As a result, the display will not be influenced by frameshifts or contain stop codons in cDNA inserts. While the intimin system in generally considered to disfavor the disulfide-bridge-containing passengers, similar to the ATs and inverse ATs, literature has shown that it actually supports the display of nanobodies that contain rigid disulfide-based structures. Hence it could potentially facilitate the translocation of an extended portfolio of disulfide bond-containing proteins.[154]

Following the selection system, the library quality is key to success and here relies on the diversity of displayable proteins. As discussed before, not all proteins are expected to be displayed. Therefore, maximizing genetic diversity within the library will statistically result in a broader spectrum of displayed proteins. For library construction, human cDNA libraries were chosen as genetic sources due to their commercial availability and extensive genetic diversity, comprising more than 80,000 transcripts encoding even more different proteins.[276] Comprehensive screenings necessitate the coverage of a significant portion of the transcriptome, for which the easily adaptable and up-scalable cloning protocols of the intimin system are valuable.

Overall, we aim to make use of the rapidity of bacterial expression and cultivation, coupled with FACS to provide precisely and adjustable selection. Further, the genetic structure and simplicity of the intimin system could facilitate the effective display of a wide array of proteins encoded within cDNA libraries and allow for the screening of readers of epigenetically modified CpGs.

10.2 Cloning of intimin-based display vectors

First, we created two display vectors harboring the truncated intimin display cassette developed by Wentzel et al. (residues 1-659)[242]. These vectors also include antigen tags for display monitoring, restriction enzyme recognition sites for cloning of cDNA inserts, and Tobacco Etch Virus (TEV) protease-digestion sites (Figure 22 a).

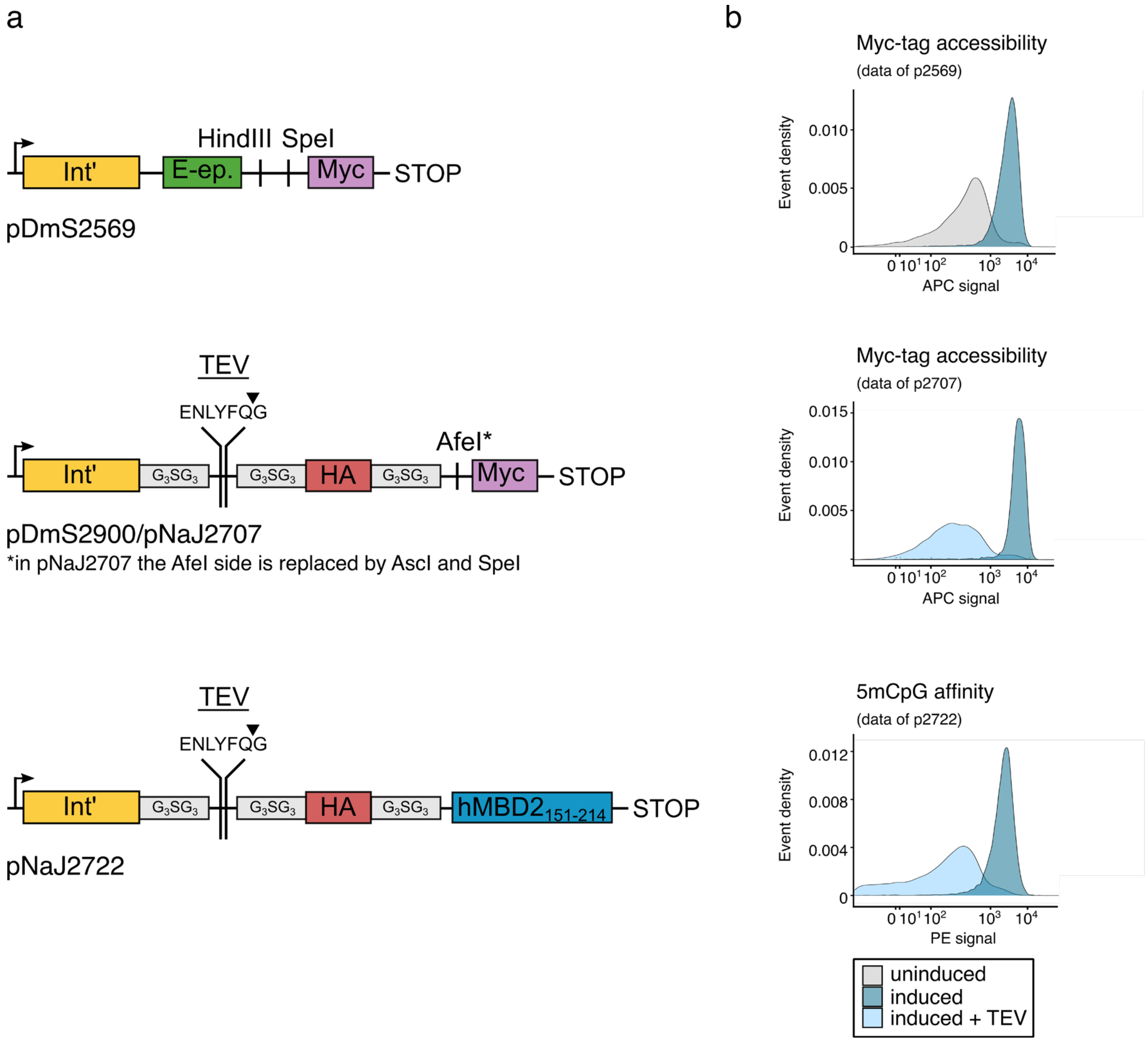

Figure 22: Genetic structure of the created vectors for intimin surface display of proteins and FACS monitoring of the surface display. (a) The full-length cDNA display vector pDmS2569 contains truncated intimin (Int') E-epitope (E-ep.), HindIII, and SpeI recognition sites for cDNA insertion and a C-terminal Myc-tag. The structure of pDmS2900 an additional G_3SG_3-flanked TEV recognition site and HA tag, an out-of-frame C-terminal Myc-tag. pNaJ2707 only differs in the removed AfeI site and the in-fram shifted myc-tag. pDmS2722 contains the hMBD2 DNA binding domain ORF (residues 151 - 214). (b) The Myc-tag display of *E. coli* BL21(DE3) Tuner cells expressing pDmS2569 and pNaJ2707 was determined via FACS with an anti-Myc antibody. The display of $hMBD2_{151-214}$ was determined via incubation with a fluorescently labeled dsDNA probe containing a 5mCpG dyad.

Both vectors were based on the pBAD33.1 vector (Addgene) allowing for tight expression control via the araBAD promoter. The vector pDmS2569 contained two restriction enzyme recognition sites (HindIII, SpeI) to allow directional and in-frame cloning of cDNA inserts via Gibson assembly. pDmS2900 contains a blunt-end restriction enzyme recognition site for cloning randomly fragmented cDNA. Notably, the Myc-tag of pDmS2900 is out of frame to the Int' sequence to avoid its expression when lacking a cDNA insert. Vectors pNaJ2707 and pDmS2900 have identical sequences except for that the blunt-end recognition site of pDmS2900 is replaced by AscI and SpeI and the Myc-tag is shifted in-frame to the intimin sequence in pNaJ2707. In addition, multiple Glycine-serine linkers were included in pDmS2900 and similar vectors to increase the flexibility of the passenger and its distance to the cellular surface, which was suggested by Salema et al. to be beneficial for its surface accessibility.[153]

We evaluated the expression and the intimin-mediated display of pDmS2569 and pNaJ2707 in *E. coli* BL21(DE3) Tuner cells by immunostaining the Myc-tag with an APC-labeled anti-Myc antibody (ab72580, Abcam, Figure 22b). A clear increase in APC signal was observed in induced cells expressing pDmS2569, indicating successful expression of the fusion protein and the display of the Myc-tag. Cells containing pNaJ2707 showed a slightly higher APC signal, but it remains to be elucidated if this observation is based on the higher display level or the improved tag accessibility due to linker incorporation. In addition to the display of the Myc-tag, we further demonstrated that a functional protein domain can also be displayed via our designed vectors. The DNA binding domain of human MBD2 (hMBD2$_{151\text{-}214}$) was displayed via the intimin system and showed binding to a 5mCpG-containing probe designed by Buchmuller et al., which was used for validating the display of human MBD DNA binding domains via the AIDA system.[109] Furthermore, the successful display is underlined by the accessibility of both passengers for the TEV protease, which resulted in an almost removal of the Myc-tag and hMBD2 after digestion. Cloning of pNaJ2707 and pNaJ2722 as well as the associated FACS experiments were executed by Natalie Jácobo Goebbels.[277]

Summary

We successfully constructed two entry vectors for displaying proteins encoded in randomly fragmented cDNA or directionally-inserted full-length cDNA. The initial FACS experiments demonstrated the functionality of the display vectors by the successful display of the Myc-tag and a functional human protein domain. This suggests that the bacterial surface display via the intimin system can be transferred to other human proteins or protein domains encoded in cDNA libraries.

10.3 Cloning of cDNA surface display libraries

To assess the potential of the intimin system in displaying human cDNA-encoded protein libraries, we utilized commercially available cDNA libraries originating from human thyroid and human prostate tissues. Our objective was to make use of the mRNA-derived, protein-coding cDNA to display protein libraries on the surface and to further compare libraries encoding full-length proteins and randomly fragmented protein segments (Figure 23).

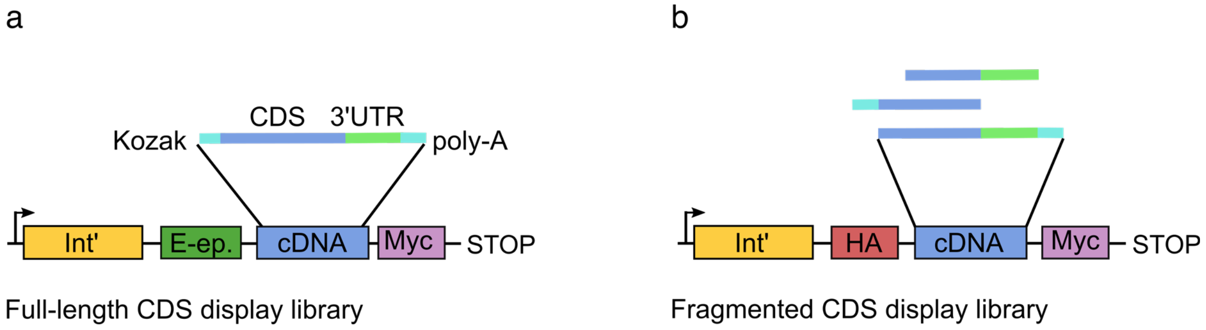
a
Kozak
CDS
3'UTR
poly-A
Int'
E-ep.
cDNA
Myc
STOP
Full-length CDS display library
b
Int'
HA
cDNA
Myc
STOP
Fragmented CDS display library

Figure 23: Overview of the surface display libraries used in this study. Simplified genetic structure of (a) the 5'-UTR-removed full-length CDS display library and (b) the fragmented CDS display library.

The 5′ UTR and 3′ UTR of mRNA exhibit sequence and size variations among transcripts (mean 5'UTR: 259 nt, mean 3'UTR: 1,470 nt)[278] and serve essential transcriptional regulatory functions. For instance, upstream ORFs (uORF), found in the 5′ UTR of over 50% of mammalian mRNAs, contain functional upstream start and stop codons that encode microproteins forming regulatory complexes with the downstream ORF.[279] Unlike the stop codons in the downstream 3′ UTR which do not affect the passenger display (Figure 23), stop codons in the 5′ UTR would terminate translation before reaching the CDS, thereby preventing passenger expression and display. Additionally, in-frame cloning of the CDS is complicated due to the variable length of the 5'UTR, resulting in out-of-frame constructs in two-thirds of the cases. Therefore, we removed the 5′ UTR prior to cloning, as depicted in Figure 23a, to enhance the display of cDNA-encoded ORFs. This was achieved by amplification of the human cDNA with a primer mix designed for vertebrate Kozak sequences (Figure 25b). These sequences are consensus sequences positioned upstream of the start codon and that play a crucial role in initiating eukaryotic ribosome translation.[280] Lee et al. reported the removal of the 5′ UTR from human urothelial cell mRNA with the primer mix in 86% of cases.[281] We employed single and dinucleotide 3′ overhangs, referred to as anchors, in the reverse poly-dT primers which targeted the 5′-end of the poly-A tail of mRNA (Figure 25b). These anchors, initially developed by Wang et al., have demonstrated effectiveness in minimizing internal poly-A stretch priming thereby ensuring the amplification of full-length CDS.[282,283] Additionally, anchored poly-dT primers have been shown to reduce the poly-A tail length in amplicons, thus eliminating additional non-coding regions of mRNA.[283] This amplicon size reduction could improve the cloning of cDNA, as the insert size is inversely correlated with DNA assembly efficiency.[284]

10.3.1 Bias in cDNA library amplification by anti-Kozak primers

First, we investigated the influence of the anti-Kozak primer mix on the amplicon composition. The human prostate cDNA (10108-A, Biocat), inserted into the pExpress plasmid vector by the manufacturer, was amplified with the anti-Kozak primer mix and M13 reverse primer (Kozak-M13) or the M13 forward and reverse primer pair (M13-only) as shown in Figure 24a (gel electrophoresis results see Figure S1). Subsequently, we subjected these amplicons to Nanopore sequencing, and the processed data was mapped to the human transcriptome using minimap2[285]. We obtained 283,169 mapped reads for Kozak-M13 and 262,269 mapped reads for M13-only amplicons. The datasets were then compared based on their composition of RNA types as well as transcripts per genomic locus (HGNC annotations).

The impact of the anti-Kozak primer mix on the RNA-type composition appears to be minor (Figure 24b). Interestingly, there is a slight reduction in the fraction of protein-coding transcripts in Kozak-M13 amplicons, which was not expected since the targeted Kozak sequence is present in all coding transcripts. However, there is an increase in the fraction of retained introns, that are incompletely spliced mRNA also containing the Kozak sequence, to a similar extent.

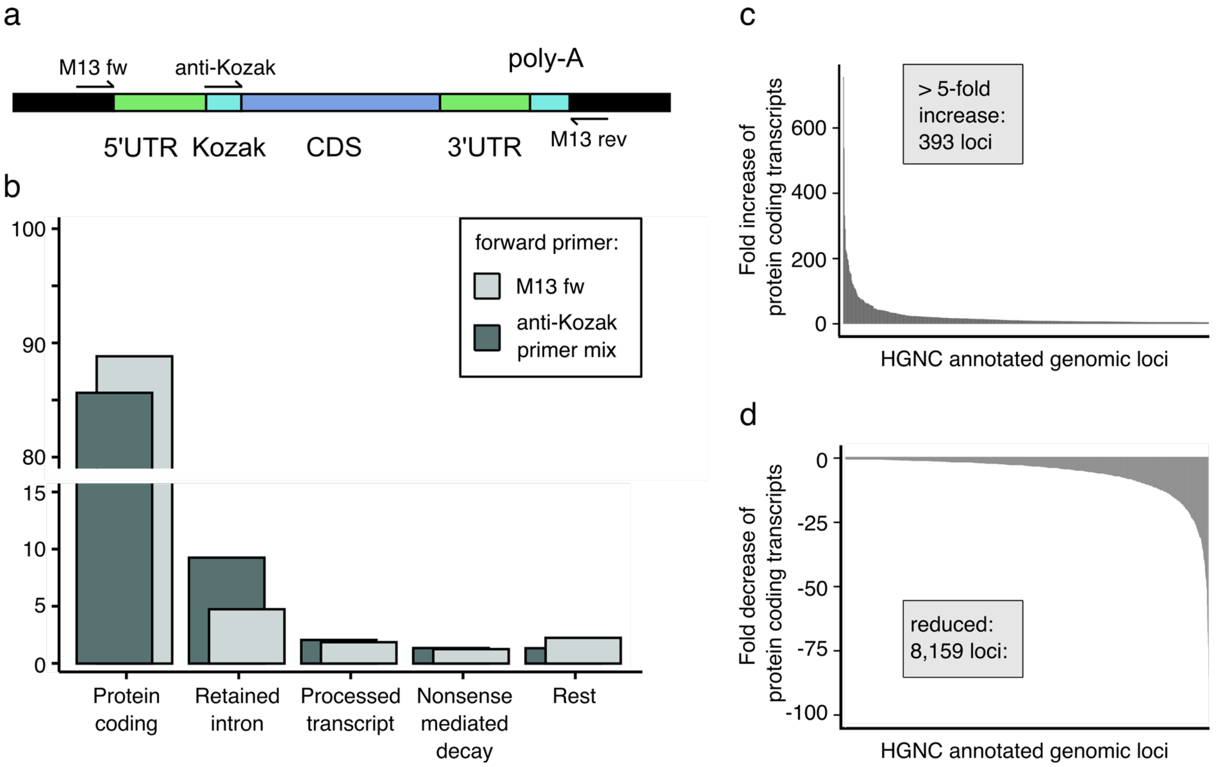

Figure 24: Impact of the anti-Kozak primer mix developed by Lee et al. on the cDNA amplicon composition[281]. (a) Schematic binding sites of the M13 primers to the vector sequence (black) and of the anti-Kozak primers to the Kozak sequence. (b) Bar chart comparing the proportion of each RNA type covered by the cDNA amplicons generated with the M13-only or anti-Kozak-M13 primers. Histogram of the coverage of protein-coding transcripts per genomic loci that were found to be more than fivefold increased (c) or underrepresented (d) in Kozak-M13-amplified cDNA.

In our analysis, we observed that protein-coding transcripts within Kozak-M13 amplicons covered approximately one-third fewer genomic loci (9,524 vs. 5,957) compared to M13-only amplicons. More precisely, M13-only amplicons covered 4,430 genomic loci exclusively, while Kozak-M13 amplicons exclusively covered only 863 loci. Additionally, the use of the anti-Kozak primer mix significantly increased the transcript proportion of a a minority of loci, accompanied by a reduction of over 8,000 loci (Figure 24c and d). These changes exhibited varying intensities between the genomic loci, ranging from a 600-fold increase to nearly complete depletion.

Summary

Amplification with anti-Kozak primers reduced the genetic diversity of protein-coding transcripts, affecting the overall transcriptomic coverage and the abundance of transcripts of specific genes. Nonetheless, targeting the Kozak sequence offers the advantage of cloning displayable, in-frame ORF libraries from cDNA. It's important to note that generating unbiased libraries is generally unattainable here since non-normalized cDNA library are used as templates. In this context, the induced bias by the anti-Kozak primers remains acceptable for cloning initial display libraries and should enhance the surface display of ORFs.

10.3.2 Creation and characterization of a full-length CDS display library from human thyroid cDNA

As evaluated in section 10.3.1, the bias in the cDNA amplicons introduced by the anti-Kozak primers was considered tolerable considering the improvements for the cloning of ORFs by the 5′ UTR removal. Full-length CDS were amplified from human thyroid cDNA (HD-503, Zyagen) with anti-Kozak primers and anchored poly-dT primers (Figure 25b). The primers contained 5′-overhangs allowing for Gibson assembly. The resulting amplicons showed a uniformly distributed size range of 500 to 3,000 bp (Figure 25c), which was slightly lower than the reported average size of human mRNA (median 2,787 nt, mean 3,392 nt).[278] This indicated the successful primer-derived removal of the 5′UTR and parts of the poly-A tail. The observed intense bands could be attributed to highly abundant transcripts within the non-normalized cDNA template or could rely on introduced PCR bias by the anti-Kozak primer mix. The amplicons were assembled with the display vector pDmS2569, yielding a total diversity of 1,022,500 ± 77,000 transformants, as determined through cfu counting (Table S1). Sanger sequencing revealed that four out of seven analyzed transformants contained in-frame, full-length CDS with successfully removed 5′ UTRs (Table S2). These values agree with the findings on the anti-Kozak primers by Lee et al.[281]

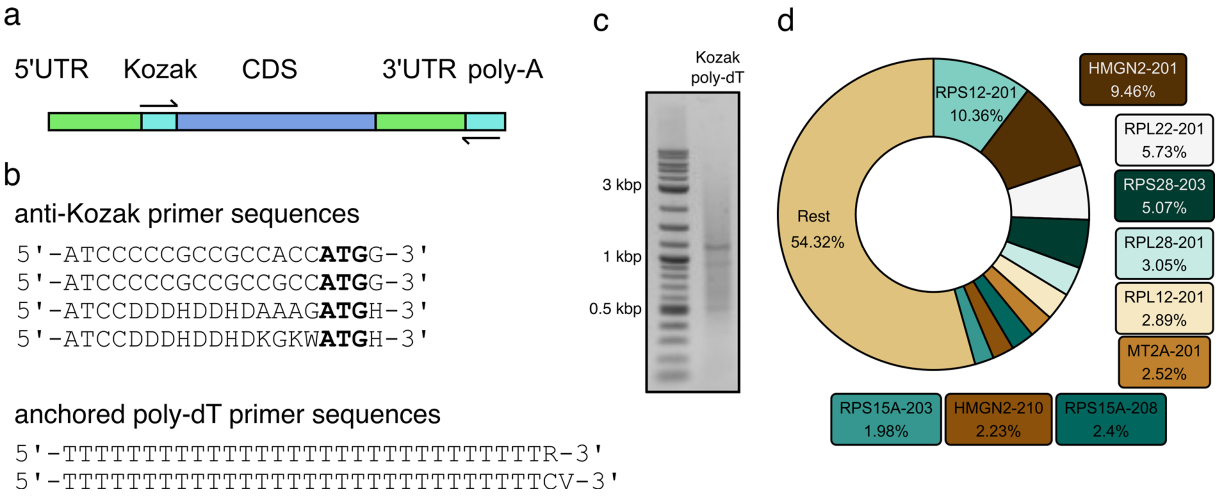

Figure 25: Creation of a full-length CDS display library from human thyroid cDNA. (a) Schematic structure of mRNA-derived cDNA and the binding sites of the forward anti-Kozak primer mix developed by Lee et al.[281] and the reverse anchored poly-dT primer mix based on Wang et al.[286] (b) Sequences of the anti-Kozak primers and the anchored poly-dT primers. The nucleotides annealing to the start codon are highlighted in bold. The 5'-end overhangs for directional cloning via Gibson assembly are not shown. (c) Product of the PCR of human thyroid cDNA with anti-Kozak and anchored poly-dT primer mixes. (d) Library composition as determined by the mapping of nanopore sequencing reads against the human transcriptome with minimap2[285].

The composition of cDNA inserts of the display library was investigated by next-generation sequencing (NGS). For this purpose, we chose Nanopore sequencing since it allows for efficient sequencing of the large cDNA inserts within of library.[287] Moreover, we avoided to introduce PCR bias to the sequencing by linearizing the plasmid library with a restriction enzyme instead of PCR amplification. Mapping of the reads against the human transcriptome unveiled a library bias, characterized by a higher proportion of ribosomal protein transcripts and HMGN2 transcripts (Figure 25d, Table S3). Despite the limited sequencing depth (33,569 mapped reads a total of 3,958 distinct transcripts were identified. Moreover, the majority (2,295) were classified as protein-coding transcripts. This suggests an adequate genetic diversity for the development and evaluation of the screening system for 5mC-specific reader proteins.

10.3.3 Creation and characterization of a fragmented CDS display library from human prostate cDNA

A common method for enhancing ORF display in phage and yeast systems is the utilization of randomly fragmented cDNA. This approach facilitates the accessibility of structurally complex proteins for surface display by displaying protein parts or domains (see section 10.3, Figure 23b).

To create a fragmented cDNA surface display library, the commercially available prostate cDNA library (10108-A, Biocat) was used. The library was originally inserted into the plasmid vector pExpress and served as the genomic template for the creation of the display library by Nadeshda Kataev.[288] Since amplification of the plasmid-inserted library with the anti-Kozak and anchored poly-dT primer mixes was not successful (Figure S2), the full-length cDNA was first amplified with the M13 primer that bound directly upstream and downstream of the insertion site. The sizes of the resulting amplicons ranged from 500 bp to more than 10,000 bp (Figure 27a, left), which exceeded the values for human mRNA in literature. We did not further investigate whether this is attributed to tissue-specific RNA composition or PCR artifacts. However, these amplicons served as templates for the PCR with the anti-Kozak and anchored poly-dT primer mixes. The product showed a reduced size of 300-1500 bp, which is similar to the PCR product obtained from human thyroid cDNA (Figure 27a, right). The presence of multiple intense bands suggested a significant bias, which was likely introduced by the two consecutive PCR steps.

Next, the 5′ UTR-removed amplicons were subjected to random shearing by sonication. It is essential to control the size of the fragmented cDNA to ensure the display of protein fragments of typical domain size. In general, protein domain sizes peak at around 100 residues, as determined by a sequence-dissimilar protein dataset of Wheelan et al.[289] While MBD domains match these values, the SRA domains of human UHRF1 and UHRF2 emphasize the relevance of domains, exceeding 200 residues, in recognition of epigenetically modified DNA (Figure 26).

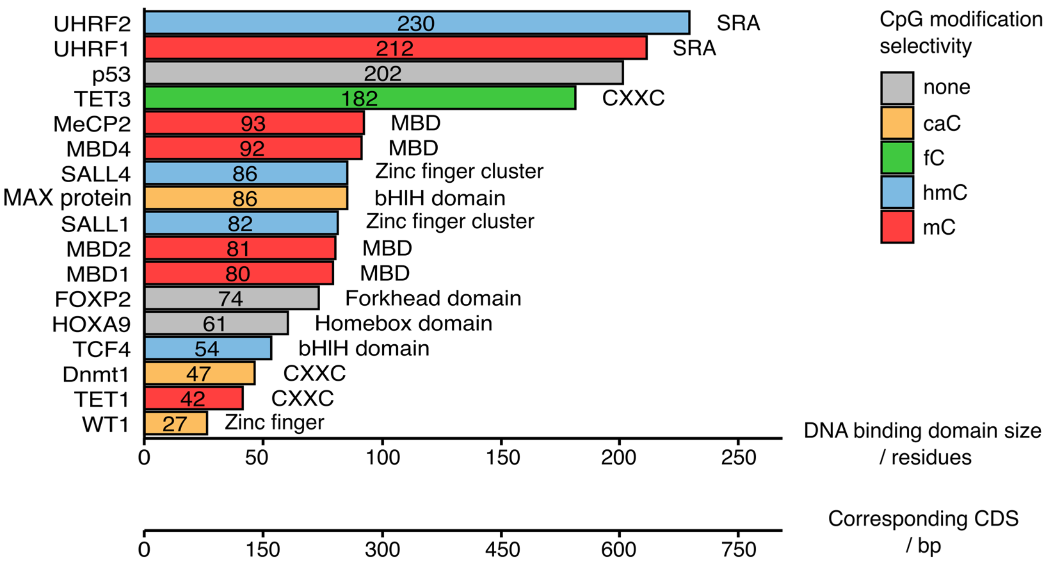

Figure 26: DNA binding domain (DBD) size of a selection of human epigenetic readers and transcription factors. Information is obtained from literature or the UniProt database.[58,82,90,290–294] In cases where multiple DBDs are present, the domain with the highest DNA affinity and/or CpG modification selectivity is shown. The predominant CpG modification selectivity of the domains is depicted by color.[58,73,82,86,90,274,290,291,293,295,296]

To ensure the coverage of domains up to the size of SRA domains, the 5′ UTR-removed prostate cDNA amplicons were sheared to a size of 200-900 bp (Figure S3). Notably, some highly abundant cDNA amplicons were not fully fragmented. The fragments were subsequently end-repaired, phosphorylated, and ligated with pDmS2900, yielding a diversity of 350,000 ± 20,000, as determined via cfu counting (Table S4).

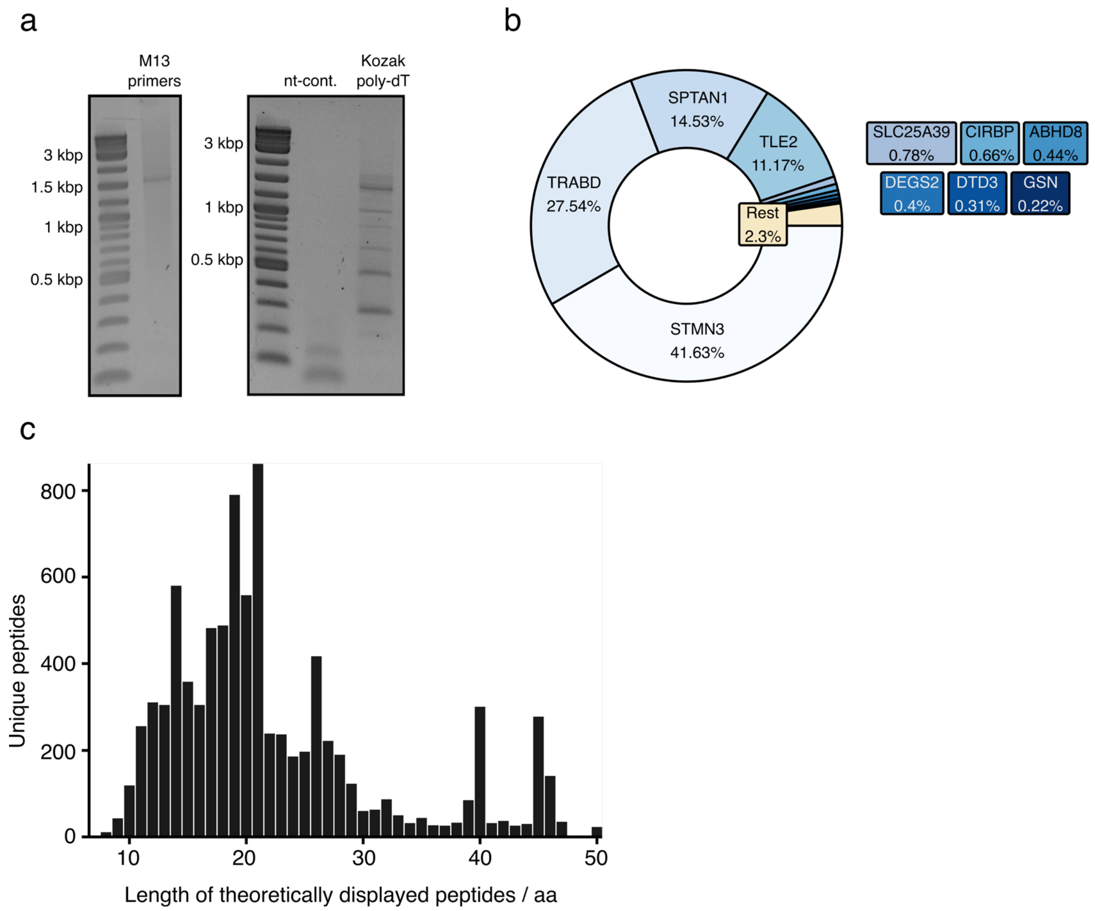

Figure 27: **Creation of a fragmented display library from human prostate cDNA**. (a) Human prostate cDNA was amplified with M13 primers (left). The amplicons were used as templates for the amplification with anti-Kozak and anchored poly-dT primer mixes (right). nt-cont: no-template control. (b) Library composition as determined by mapping of Illumina sequencing reads against the human transcriptome. (c) Length distribution of in silico-translated unique peptides encoded in ORFs that are in-frame with the intimin display cassette. The peptides of the seven highest abundant proteins were omitted for clarity. The length determination was limited to 50 residues by the Illumina read length.

Illumina deep sequencing of the library and mapping against the human transcriptome revealed a pronounced bias towards transcripts of certain genes (Figure 27b) which is likely resulting from the bias introduced during cDNA amplification. Sanger sequencing of the library confirmed the integration of cDNA inserts in 92% of plasmids. These exhibited a length of 400 bp ± 269 bp and could potentially encode peptides of 133 residues. Notably, typical inserts harbored ORFs of only 55 amino acids (Table S5). Furthermore, 23% of inserts showed the correct orientation but all were out of frame with the intimin sequence. Studies of C-terminal cDNA fragment display in yeast and phage showed that only a fraction of 2-10% of the libraries contains displayable ORFs, being limited by non-directional insertion, UTRs, and frameshifts.[194,297,298] To overcome these limitations, directional cloning of random-primed fragments and frame-shift PCR have been utilized in some studies. Both techniques enhance the probability of correct ORF insertion but were not employed in this study.[299] Instead, it is likely that the removal of the 5′ UTR improved the insertion of protein-coding ORFs in frame with the intimin sequence, which was not detected here due to the limited sequencing depth (n = 13).

Consequently, to obtain more comprehensive insights on the library, the Illumina sequencing reads were subjected to in silico translation of the ORFs that are in frame with the intimin sequence. The resulting peptides were mapped against the human proteome. This allowed for the identification of ORFs coding for protein sequences. The dataset revealed that 5.45% (748,411 reads out of 13,735,080 total reads) of the library inserts coded for in-frame ORFs, corresponding to 1,333 genetic loci (Table S6). Notably, the peptides of the majority of proteins, were relatively short, typically around 20 residues or less (Figure 27c). However, the peptides of the few highly abundant proteins were of increased length and shifted the total peptide size distribution of the library toward 50 residues (Figure S6), which is the upper limit due to the Illumina read length. This connection between the number of fragments and fragment size indicates that sufficient fragment coverage per protein is required to reliably encode longer ORFs in the library.

Summary

Library type	Full-length CDS	Fragmented CDS
Organism	Human	Human
Tissue	Thyroid	Prostate
Library diversity	$1.023 \pm 0.077 \times 10^6$	$3.5 \pm 0.2 \times 10^5$

In this study, two cDNA libraries, suitable for the bacterial surface display via the intimin platform, were created. The libraries contain full-length or fragmented CDS from human cDNA. By removing the 5′ UTR we likely improved the number of ORFs that are in frame with the intimin sequence and are theoretically displayable. Both display libraries show a bias towards certain transcripts and genes and exhibit moderate diversity in comparison to typical ORF screenings in phage and yeast systems (~10^7)[138,300]. Still, their quality was found sufficient for assessing the potential of the protein surface display system and for developing a selection system for epigenetically modified CpG readers. To maintain practical timeframes during FACS, we decided against expanding the library diversity, since diversities exceeding 10^8 are impractical in most setups.[301]

10.4 Establishment of the surface display of cDNA libraries

10.4.1 Influence of the cDNA library surface display and the selection process on cell viability

The survival of the selected cells is crucial for the selection of modification-specific readers via FACS in our system (see Figure 21). Heterologous protein expression, particularly of intimin-fusion products, can significantly lower the viability of bacterial cells due to increased cellular stress.[242] With the full-length and fragmented CDS display libraries in hand, we assessed the effect of human protein expression and surface display on the cellular survival of *E. coli* cells.

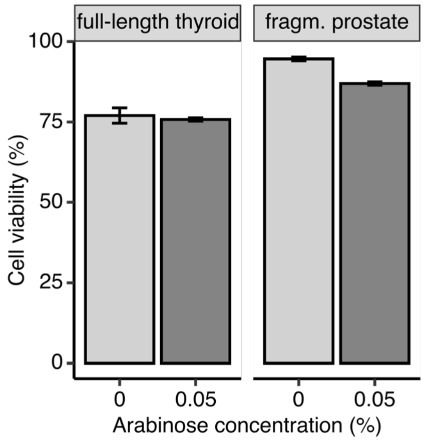

Figure 28: Survival rates of *E. coli* BL21 (DE3) Tuner cells after FACS, with and without induced expression of the display libraries. The means and standard deviations are calculated based on biological duplicates.

Studies of the Lpp-OmpA and PhoE display systems showed improved cell viability and fusion protein expression at 30 C in *E. coli*, which could further be applicable to the intimin system.[204,223] Consequently, we employed mild expression conditions (0.05% arabinose, 30 min, 30 °C) to minimize cellular stress. Furthermore, these conditions ensured sufficient display rates in the functional display of hMBD2$_{151-214}$ (see section 10.2). To assess the impact of the protein expression and display on cell integrity, *E. coli* BL21(DE3) Tuner cells with and without induced expression of the full-length and fragmented CDS display plasmid libraries were utilized. After harvesting and washing, the cells were incubated with a staining mix similar to the one for monitoring of 5mCpG-binding by hMBD2$_{151-214}$. Afterward, single cells were sorted on LB agar plates.

In contrast to the findings of Wentzel et al., who only observed sufficient cell viability upon expression reduction via amber suppression, we did not see a significant difference in cell viability of induced and uninduced cells for full-length and fragmented CDS libraries (Figure 28, raw values see Table S7).[242] Given the tight expressional regulation through the employed araBAD promoter[302], the lower viability in the case of the full-length library is unlikely due to background expression. Considering that the survival rates of cells expressing library fusion products fell within the suitable range of 75% to 90%, we chose not to further investigate this phenomenon.

Summary

The expression of the intimin surface display libraries did not significantly impact the viability of *E. coli* (DE3) Tuner cells. Considering the full experimental process, as employed in later protein screenings, the cellular survival rates were found to be between 75% (full-length CDS library) and 90% (fragmented CDS library). These survival rates are suitable for exploring the display scope of the intimin system and for establishing a screening assay for reader proteins of epigenetically modified CpG dyads. Therefore, these expression and incubation conditions were kept constant in all further experiments.

10.4.2 Display scope of human protein fragments by the intimin platform

The expression of the display libraries, along with FACS procedures have proven not to impact cellular survival. Therefore, we proceeded to investigate the scope of the intimin system for displaying human protein fragments. For this purpose, we selected cells displaying ORFs via FACS. This was achieved by immunostaining of the passenger's C-terminal Myc-tag by Nadeshda Kataev (Figure 29a).[288] Similar approaches for ORF selection via selection for C-terminal tags have already demonstrated functionality tags in phage display systems.[138,299] FACS gating was based on the signal of negative control cells, expressing the empty entry vector pDmS2900, whose Myc-tag is out of frame and not expressed (Figure S4a). The selection of cells carrying the fragmented human prostate cDNA library resulted in an overall increase in Myc-tag display after the second selection round. This demonstrated the successful enrichment of displayed ORFs with C-terminally fused Myc-tag (Figure 29b). Additionally, a slight reduction in the overall signal was observed after the final selection round. Assuming a uniform display of the human protein fragments with comparable levels to those reported for the E-epitope by Wentzel et al.[242], this phenomenon can be attributed to the saturation of free antibodies by the displayed Myc-tags. The epitopes potentially exceeded the available antibodies by more than six-fold (Equation S1). Considering the increase in fluorescence of the entire population and the antibody saturation, we concluded that the library almost exclusively contained displayed ORFs after the final selection round.

To gain a more comprehensive understanding of the surface display capabilities of the library, we employed Illumina sequencing for analyzing the cDNA inserts from the selection rounds and the parent library (Figure S5a). Our primary objective was to assess the range of biologically relevant peptides and protein fragments that were displayed by the intimin platform. To address this, we conducted in silico translation of ORFs that were in frame with the display cassette. The peptides were mapped against the human proteome, thereby focusing the analysis on biologically relevant peptides and protein fragments. This approach allowed us to avoid the common challenge of analyzing unnatural peptides based on out-of-frame cDNA inserts.[194,303,304] The fraction of mapped reads increased from 5.5% to 28.5% during the selection rounds (Table S8) demonstrating the enrichment of protein-coding ORFs.

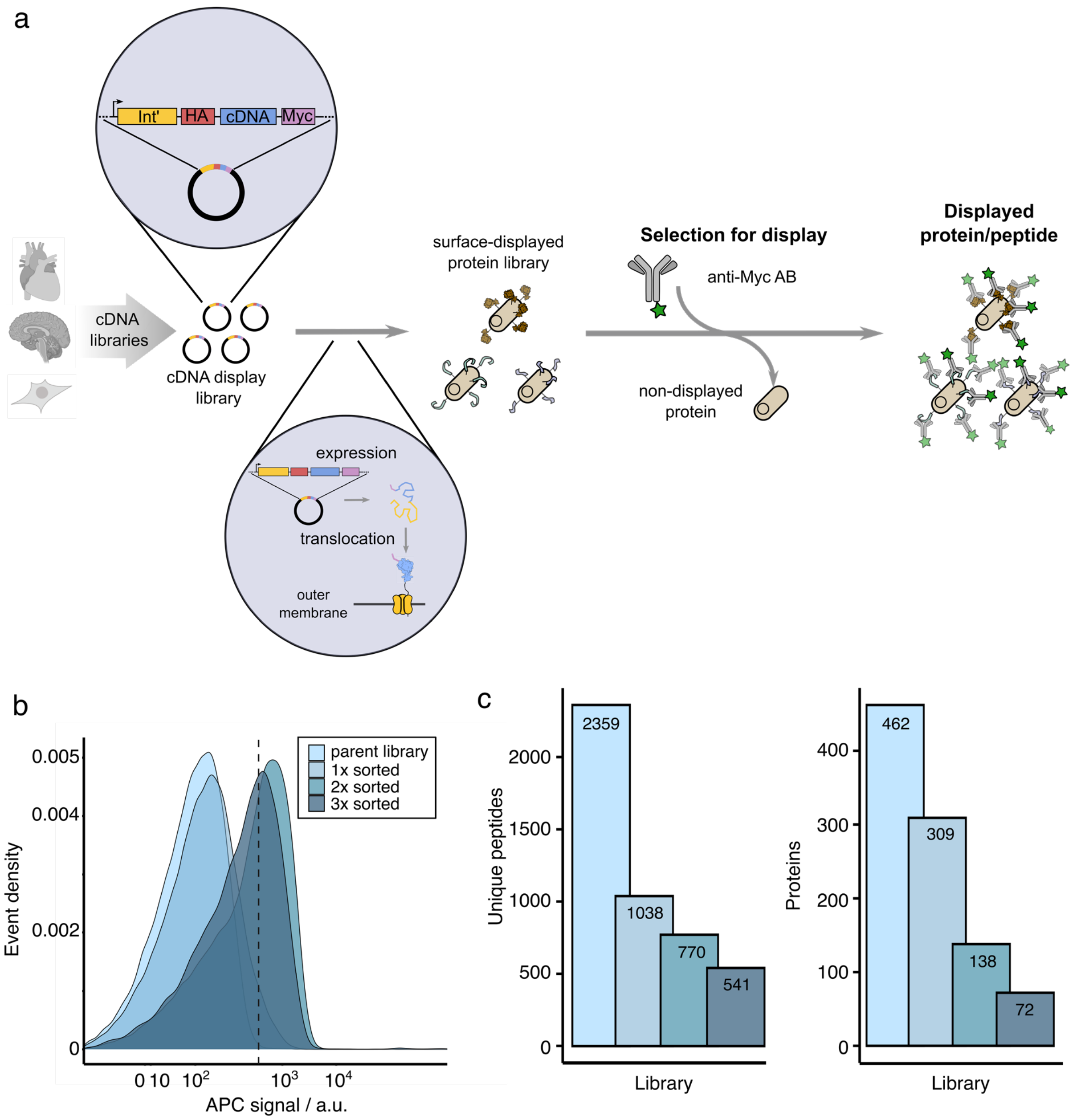

Figure 29: Enrichment of displayed ORFs. (a) Scheme of the screening workflow for the identification of displayable proteins using the C-terminally fused Myc-tag. (b) FACS enrichment of the ORFs from the fragmented prostate cDNA library with anti-Myc antibody. Density plot of the APC-signal throughout the screening process, representing the surface exposed Myc-tag. The area right of the dashed line was chosen for gating. (c) Diversity of the library after each selection round regarding their biologically relevant peptides and the corresponding proteins. Proteins of high abundance (> 1% of the library) were omitted for clarity. For comparability, the same number of mapped peptides of the parent library and each selection round was analyzed.

We compared equal amounts of mapped peptides (n = 150,097) from each selection round regarding their diversity on the peptide and protein level. To ensure that fragments from highly abundant proteins (>1% of the library) did not distort the results, we excluded them from the analysis. We observed a consistent reduction in both peptide and protein diversity throughout the ORF selection process (Figure 29c). This matched the expectations, since statistically the Myc-tag is out of frame in two-thirds of the inserts due to frameshifts and translation termination via stop codons within the 3' UTR.[299] For these reasons and the moderate depth of our analysis, 541 displayable peptides from 72 proteins (excluding abundant proteins) are likely an underestimate of the total displayable library fraction.

The length distribution of the displayed peptides roughly matched the expectations of the theoretical estimations of the parent library that we determined in section 10.3.3. However, most of the longer peptides (>30 residues) that were displayed corresponded to highly abundant genes. Genes of lower abundance showed displayed peptide sizes of 27.7 residues on average (compare section Figure S6 and Figure S7). This demonstrated that sufficient fragment coverage per transcript/protein is necessary to reliably display longer ORFs. The coverage can be improved by increasing the library diversity and reducing the bias regarding certain transcripts. It is important to note that the length of many peptides is underestimated because the analysis is limited to 50 residues due to the Illumina sequencing read length. Additionally, residues from the linker and Kozak sequence (4-8 residues) were trimmed before analysis, further reducing the detectable peptide size.

However, the sizes of displayable peptides are within the size range of Zinc finger, CXXC, and bHlH domains of known epigenetic readers and transcription factors (27-54 residues, compare section 10.3.3 Figure 26). In contrast, the displayed fragments of known nucleic acid binders or epigenetic readers only partially covered binding domains (CIRBP, RBM, MBD2) or belonged to spacing and disordered regions of the protein (Table 3). This suggests a display preference for passengers of low structural complexity. Whether this represents a minor favor or a strict limitation by the intimin platform cannot be determined with this limited dataset. On the one hand, the successful translocation of disulfide-bridge-containing nanobodies suggests that there is no general limitation in displaying more complex passengers via the intimin system. On the other hand, studies by Adams et al.[247] demonstrated hindered display of passengers that contain tertiary structures.[153,154]

Table 3: Displayed protein fragments of epigenetic readers and nucleic acid binding proteins. Molecular function and structural information were received from the UniProt database.[294] The values correspond to the mean and standard deviation of the displayed peptides per protein. The ORF length corresponds to the alignment length determined via BLAST+. *Real ORF might be extended due to detection limitations by sequencing length and trimming.

Protein	Molecular function	Mapped residues	ORF length / aa	Structural annotation
Brd3	AcLys reader	187 - 206	20.0±0.0	Disordered region
CIRBP	mRNA binding	1 - 26±6	26.0±6.1	N-terminus of RRM
MBD2	5mCpG binding	151 - 176	26.0±0.0	N-terminus of MBD
Myc	DNA binding	423 - 438±0.1	19.0±0.1	Myc-tag sequence
NCOR2	DNA binding	1739.6±6.4 - 1768±0.2	29.4±6.4	Basic & acidic residues
RBM3	RNA binding	1±0.15 - 39.9±1.3	39.9±1.3*	N-terminus of RRM
ZC3H3	DNA binding	585 - 600	16.0±0.0	No annotation
Znf195	DNA binding	97 - 119	23.0±0.0	Spacer region
Znf318	No annotation	67 - 89.5±0.7	24.0±1.4	Pro- and Arg-rich

The removal of the 5′ UTR likely facilitated the display of CIRBP and RBM3 starting from the first residue and included the RNA recognition motifs (RRM). However, most ORFs do not cover the complete N-terminus of proteins (Table S9). It has to be noted that the displayed Myc fragment exclusively originated from the utilized Myc-tag, which was likely shifted in frame by short inserts that are not attributed to biological meaningful peptides.

The data demonstrates the successful translocation of ORFs encoded in fragmented human cDNA by the intimin platform. Furthermore, it is probable that a variety of displayed fragments were not detected due to limitations in sequencing depth and C-terminal Myc-tag expression.

Previous studies of the display of cDNA-encoded proteins by yeast and phage platforms faced the challenge that the majority of clones displayed non-ORF peptides. This was based on inverse or out-of-frame insertions or premature stop codons. Addressing this, Caberoy et al.[299] utilized ORF enrichment prior to the actual screening. Following ORF pre-selection, they successfully enriched functional ORFs, some up to 200 residues in size (with an average of 100), when screening for different biological targets.[138,139,299] These ORF sizes are significantly larger than the average displayed peptide in our study, even when considering the limitation by the Illumina read length. The difference in ORF size might be attributed to the library quality improvements implemented in those studies, which included frame-shift PCR and directional cloning, along with a higher library diversity of 1×10^8 (compared to our study's 3.45×10^5). However, drawing a direct parallel to their study is difficult as they additionally employed functional selection after ORF enrichment. This has presumably increased the average ORF size due to the selection of more complex, functional elements.

Summary

For the first time, we could show the display of protein fragments encoded in a human cDNA library on the bacterial surface. Notably, these fragments covered parts of DNA binding proteins and epigenetic reader domains, albeit smaller in size compared to published yeast and phage display studies. This invites further investigation of the display capability of the intimin platform and its application for displaying functional proteins or protein domains.

10.4.3 Functional display of proteins: Enrichment of DNA binders

The successful translocation of proteins and protein fragments to the bacterial surface, as demonstrated in section 10.4.2, is a critical step in developing a bacterial surface display selection system. In our effort to assess the intimin system's capability for the display of functional protein fragments, we opted to enrich DNA binding domains from the fragmented prostate CDS display library via FACS. Simultaneously, we aim to obtain insights into the system's effectiveness for the selection of specific protein properties that can be transferred to the enrichment of 5mC-specific reader proteins. Cells displaying protein fragments were incubated with a biotinylated random 30mer probe (Figure 30a), which was fluorescently labeled with Streptavidin-PE. We opted for a random probe to eliminate sequence-based constraints during the enrichment process.

Throughout the course of three selection rounds, we consistently observed a shift of the cell population towards increased PE signals (Figure 30b). Similar to the ORF selection via anti-Myc antibody in section 10.4.2, the overall fluorescence signal decreased after the final selection round. This observation can be attributed to the saturation of probes by displayed DNA binders (Equation S2). Thus, the probe saturation indicates the successful enrichment of dsDNA binding protein fragments.

To gain deeper insights into the enriched protein fragments, we employed Illumina deep sequencing of the cDNA inserts of each selection round (Figure S5b). The inserts displayed significantly larger sizes compared to the ORF selection (compare Figure S5a). This was expected since molecular function requires a sufficient size of the polypeptide. We adjusted the sequencing depth according to the estimated diversity of each sub-library. Identical to the Myc-enrichment data, we conducted in silico translation of ORFs, performed trimming, and then mapped the ORFs against the human proteome using the BLAST+ suite. With this workflow biologically irrelevant peptides that might have interacted with the probe, e.g. through electrostatic interactions with the DNA backbone, were omitted from the analysis.

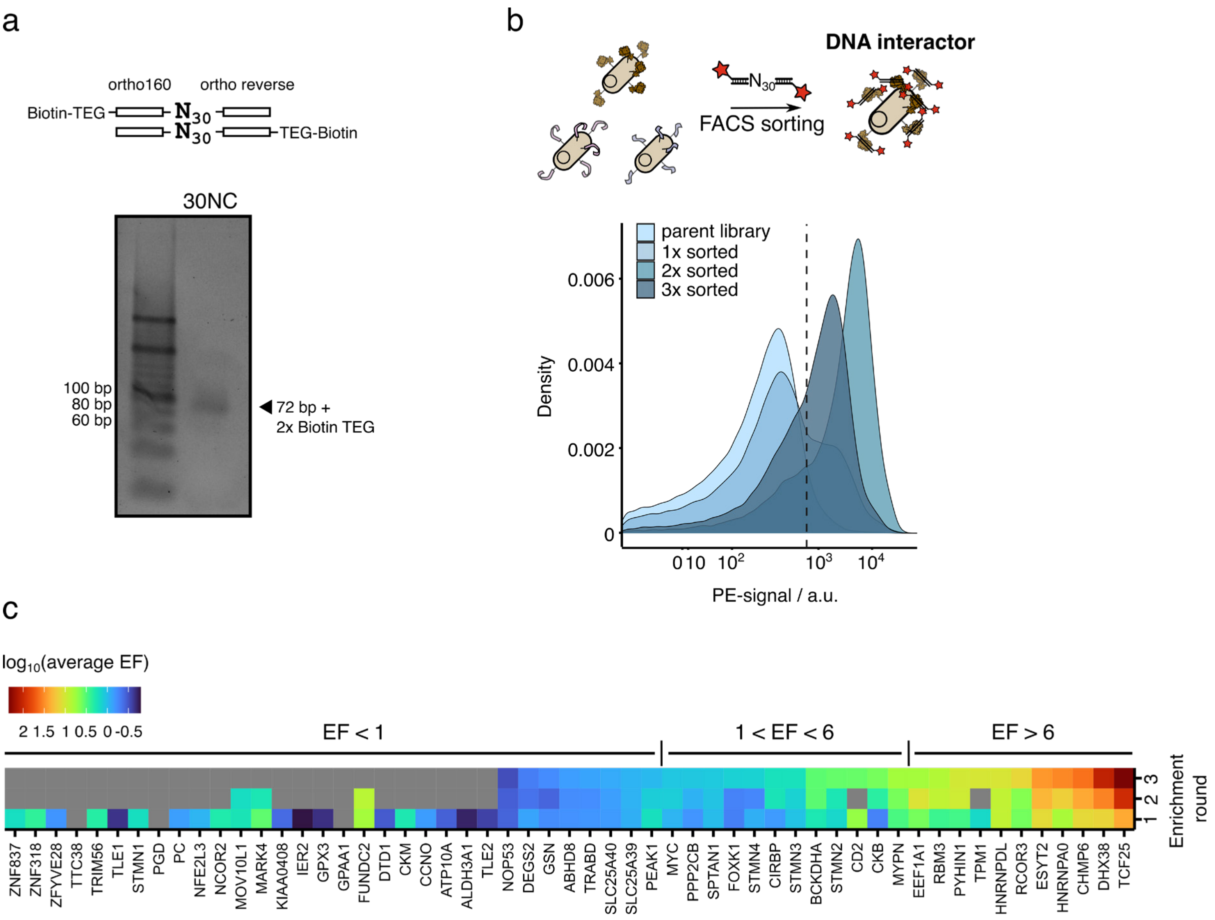

Figure 30: Enrichment of functional DNA binding proteins and their fragments via bacterial surface display of human cDNA libraries. (a) Schematic structure and agarose gel electrophoresis of the PCR-generated, biotinylated random 30mer FACS probe (30NC). (b) Top: Schematic process of the DNA binder enrichment. Bottom: PE signal of the three FACS selection rounds of the fragmented human prostate CDS display library. (c) Heatmap plot of the enrichment process of protein fragments throughout the three FACS selection rounds. Illumina reads were in silico translated and the resulting peptides were mapped against the human proteome. Total enrichment factors (EF) were calculated from the normalized counts of all peptides of each protein. The EF of grey tiles could not be calculated due to low read numbers or is not shown as it is below 0.1.

Notably, a substantial portion of the enriched, mapped peptides exhibited sizes exceeding 40 residues (Figure S8). This could indicate the enrichment of functional, more complex protein fragments compared to the ORF enrichment of section 10.4.2. As mentioned before, peptides of 40 residues might be attributed to prolonged ORFs that remained undetected due to read length limitations and peptide trimming. The mapped peptides were summed per protein and the read counts were normalized on the number of reads of the individual sub-libraries. This normalization allowed for the calculation of enrichment factors (EF) of each protein in the selection rounds. To ensure the reliability of the calculated EFs, proteins with fewer than five reads were excluded from the analysis (Figure 31).

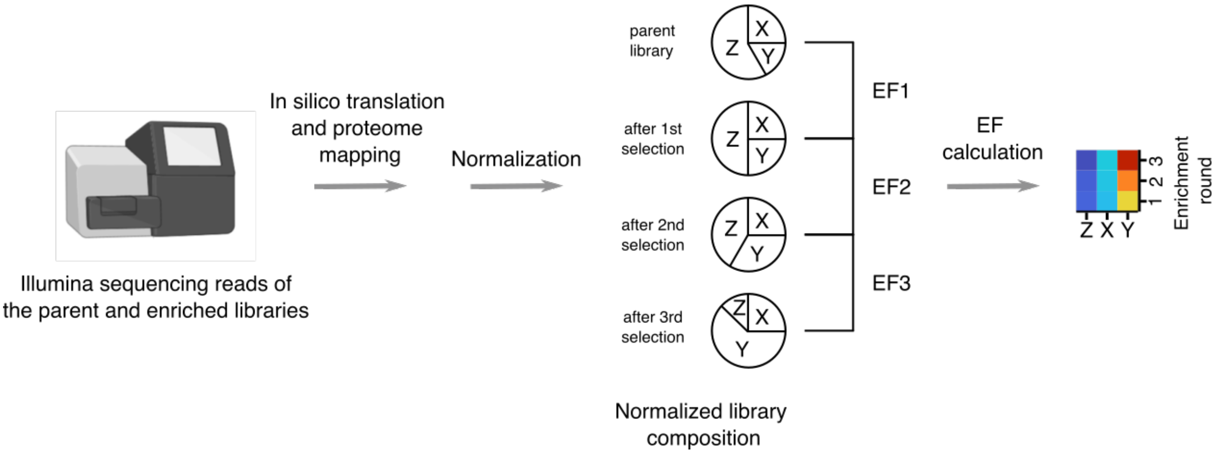
Illumina sequencing reads of the parent and enriched libraries
In silico translation and proteome mapping
Normalization
parent library
after 1st selection
after 2nd selection
after 3rd selection
Z
X
Y
EF1
EF2
EF3
Normalized library composition
EF calculation
Enrichment round
1
2
3
Z X Y

Figure 31: Schematic workflow of the EF calculation of proteins and protein fragments in the library screening. The numbers of in silico translated and mapped peptides of the proteins are normalized on the Illumina read number of the respective library. These normalized values are then used to monitor the enrichment of each protein throughout the selection process by the calculation of EF.

We closely monitored the EFs of all adequately covered proteins, thereby tracking their enrichment and depletion throughout the entire selection process (Figure 30c, Table S10). Fragments of 25 proteins exhibited an EF > 1 after the final enrichment round and were therefore considered to be enriched for general dsDNA binding. Notably, fragments of certain proteins like the transcription factor TFC25 and the RNA helicase DHX38 were enriched more than 100-fold. Fragments of proteins that were highly abundant in the parent library, such as SPTAN1, CIRBP, and STMN3, displayed relatively moderate enrichment factors between 1.2 to 1.7. Consequently, fragments of these proteins were not considered to possess significant DNA binding capabilities. The other highly abundant proteins in the parent library were depleted or were to levels below the threshold criteria (Table S11). These observations demonstrate that the selection system effectively discriminates against peptides of non-DNA binding proteins, even when they are highly abundant in the input library. Additionally, the functionality of the selection process is underlined by the consistent EF trends observed in the majority of either enriched or depleted proteins. It is worth noting that the depletion of zinc-finger-containing proteins Znf318 and Znf837 does not interfere with these findings, as the displayed peptides do not cover the DNA binding domains.

The peptides of Myc that were found to be slightly enriched (EF = 1.1) were consistently mapped to the Myc-tag sequence and were detected in negligible quantities (read count = 5).

The majority of proteins that were enriched more than six-fold were nuclear proteins and are involved in nucleic acid binding, which aligns with our expectations and further supports the effectiveness of the enrichment system (Table 4). Surprisingly, the corresponding enriched peptides of these proteins did not cover full nucleic acid interaction domains. RRMs of HNRNPA0 and RBM3 were partially covered. At this point, it should be noted that various RRMs are known to bind to double-stranded DNA (dsDNA).[305–308] To determine the plausibility of the enrichment of the peptides we subjected them to DNA binding prediction via *DP-bind*[309] and *DRNApred*[310]. *DP-bind* predicted dsDNA binding for the RRM fragments of RBM3 which stays in contrast to the calculations of *DRNApred*[310] that primarily predicted RNA interactions. Both DNA and RNA prediction tools were employed to assess the nucleic acid interaction potential of all enriched proteins (Table S12). As previously mentioned, the limitations of Illumina read length prevented us from determining the complete size of certain protein fragments. In the case of HNRNPA0, it is unclear if the fragments extend toward the predicted DNA-binding, C-terminal regions. Evidence from the literature suggests HNRNPA0 may be involved in binding immunostimulatory (CpG-rich) dsDNA, providing a possible explanation for the enrichment.[311]

Table 4: The proteins enriched for dsDNA binding (EF ≥ 6), their cellular function and localization as well as the properties of the enriched protein fragments. If not mentioned otherwise, information on the proteins was obtained from the UniProt database.[294] Residue and length values represent the mean and standard deviation of all mapped peptides of the respective protein. *Real ORF might be extended due to detection limitations by sequencing length and trimming.

Protein	Cellular function	Cellular localization	Mapped residues	Region function or property	ORF length	Final EF
TCF25	DNA binding	Nucleus	513.7±0.5 - 542.3±8.9*	Part of DUF654 domain[312]	29.6±9.4*	220.3

DHX38	RNA binding	Nucleus	1178.0±0.3 - 1211.0±0.1	Helical, Basic and acidic residues	34.0±0.3	115.1
CHMP6	Cargo sorting	Cytoplasm, Nucleus[313]	1.0±0.1 - 32.9±7.8*	Helical, Basic residues	32.9±7.8*	41.7
HNRNPA0	RNA binding	Nucleus	63.0±0.5 - 92.1±9.8*	C-terminus of RRM	30.1±9.8*	33.6
ESYT2	Lipid binding	Cytoplasm, ER	76.3±1.3 - 102.0±0.5	None	32.7±1.1	29.1
RCOR3	Transcript. regulation	Nucleoplasm	424.1±0.5 - 446.1±0.5	Prolin rich	26.2±1.6	13.4
HNRNPDL	DNA & RNA binding	Nucleoplasm	17.0 - 36.0	Disordered region	32.0±0.0	10.6
TPM1	Actin binding	Cytosol, PM	1.1±0.3 - 34.3±8.2*	Disordered region	34.2±8.3*	10.5
PYHIN1	DNA binding	Nucleolus	166.0±0.4 - 181.4±3.5	Disordered region Polar residues	18.5±3.9	10.0
RBM3	RNA binding	Nucleoplasm	1.0±0.2 - 39.7±2.4*	N-terminus of RRM	39.7±2.6*	7.1
EEF1A1	RNA binding	Nucleolus	360.0 - 404.0*	Central part of domain III[314,315]	45.0±0.0*	6.1
MYPN	Actin binding	Nucleoplasm	841.8±0.6 - 862.0±4.9	Disordered region	21.3±4.7	6.0

Whereas structural, compositional, and functional data, as well as binding predictions of RCOR3, PYHIN1, and MYPN fragments, do not sufficiently explain their enrichment, *DP-bind* or *DRNApred* predicted interactions with dsDNA for the fragments of DHX38, ESYT2, HNRNPDL, TPM1, EEF1A1 (Table S12). Interestingly, in the case of the RNA helicase DHX38 and the elongation factor EEF1A1, the enriched fragments did not correspond to previously published protein domains or regions known to be involved in nucleic acid interactions.[294,315,316] This observation raises the possibility of novel interaction sites. Notably, the fragments of DHX38 contained basic residues that could have interacted with the dsDNA probe in our screening but are not directly involved in the molecular function of the full-length protein.

Similarly, fragments of the domain of unknown function 654 (DUF645) of TCF25 could have exhibited artificial dsDNA binding properties arising from fragmentation that might have contributed to the enrichment. So far, this domain is only known to play a role in transcriptional repression through histone deacetylase recruitment upon TCF25 chromatin binding.[317] The ORF of the enriched fragments (513.7±0.5 - 542.3±8.9) of DUF654 could potentially extend further, potentially covering the C-terminus of the domain and the rest of the protein. The C-terminal region is not known to be involved in transcriptional repressor activity but is predicted to hold an alpha helix that includes basic residues (AF-ID: AF-Q9BQ70-F1).[312] Given that the fragment of TCF25 exhibited the highest enrichment factor in this study, this helix might have bound to the DNA backbone via electrostatic interactions. It must be noted that these properties are not predicted by the *DP-bind* and *DRNApred* tools.

The enriched protein CHMP6 is mostly known for its cytoplasmic localization and involvement in the ESCRTIII mediated vesicle trafficking.[318–320] Strikingly, a recent study has shown that CHMP6 is also located in the nucleus[313], similar to its family member CHMP1 which is known to bind to ssDNA and dsDNA in vitro and chromatin in vivo.[321,322] Moreover, the enriched fragments start from the protein's first residue, hold basic residues, and are predicted to bind to dsDNA by *DP-bind*. Whether the dsDNA interaction of CHMP6 is based on this basic stretch, which is known to be involved in a charged-based activation mechanism[318], or on C-terminal protein regions remains unclear. Nonetheless, our findings support the idea that CHMP6 could be involved in DNA binding upon nuclear localization.

In a phage display-based screening for dsDNA binders, Cicchini et al.[184] showed the enrichment of binders from a fragmented human cDNA library (300-1200 bp) which was of two magnitudes higher in diversity than the one utilized in this study (3.4×10^5 vs. 5×10^7). Interestingly they reported the enrichment of protein fragments of similar size (27-100 aa) compared to our intimin-based study. This comparison considers the potential elongation of open reading frames (ORFs) for some protein fragments due to the limitations of sequencing read length in our study. While our Illumina deep sequencing analysis has constraints on read length, it allowed for precise tracking of enriched fragments throughout the selection process. Consequently, we could differentiate the genuinely enriched fragments from carried-over fragments originating from a bias in the parent library. This phenomenon might have influenced the enrichment results in the phage display study of Cicchini et al. However, if the authors had employed a similar sequencing depth, they would have likely been able to identify more DNA-binding proteins. Unfortunately, Cicchini et al. did not comment on the coverage of dsDNA binding domains by the enriched fragments. We did not observe the enrichment of whole domains in our screening. Enhancing the diversity of the display library could have further facilitated the enrichment of known DNA binding domains as it would have benefitted the statistical coverage of full domains.

The need for a diversity of 10^7 or 10^8 clones to adequately cover the human ORFeome in fragmented libraries[304], could be overcome by displaying full-length proteins. In this case, each protein could ideally be represented by the cDNA insert of a single clone, significantly reducing the library diversity required for comprehensive ORFeome coverage. Aligned with this, preliminary evidence indicates that the intimin platform can adeptly display full-length proteins, as evidenced by the enrichment of presumably extended ORFs of CHMP6, TPM1, and RBM3.

Summary

In the first test for the enrichment of functional protein fragments, we were able to select fragments of cDNA-encoded DNA-binding proteins displayed on the bacterial surface. The majority of enriched fragments were from nucleic proteins that are involved in DNA and RNA binding and transcriptional regulation. Despite this striking evidence of the functionality of our selection system, the identified enriched fragments do not, or only partially cover domains known for nucleic acid interactions. The absence of completely covered domains is either based on the low diversity of the fragmented cDNA library or potential limitations of the displaying system as it has also been assumed in section 10.4.2. In contrast, extended ORFs potentially covering full-length proteins, have been enriched, thereby demonstrating the display potential. One of these proteins, CHMP6, has already been identified to be localized in the area of chromatin. Thus, our data provides further evidence for its DNA interaction properties.

10.5 Screening for 5mC-specific readers encoded in human cDNA

With proof for the functionality of the selection system in hand, we next assessed its potential for discovering DNA-binding proteins with 5mCpG affinity. For this, we designed biotinylated, partially randomized probes containing a central CpG dyad (Figure 32a) via PCR. We created on-target and off-target probes, containing either 5mC (8NmC probe) or the unmodified cytosine (8NC probe). These probes allow simultaneous selection for increased 5mC affinity compared to C in a two-color FACS setup. The random tetramer sequences, ecompassing over 65,000 CpG sequence contexts, were designed to circumvent limitations in reader binding due to sequence specificity. It is worth noting that the two random tetramers within the binding region contain additional 5mCs (or Cs), adding 0.375 CpGs on average to the central binding part. To prevent the enrichment of non-CpG-related binding proteins, we focused the selection on the central CpG by adding an excess of unlabeled background dsDNA, containing the same sequence but having the central CpG substituted with any other dinucleotide (DpH). A similar approach has proven effective in eliminating non-CpG-related DNA binding of MeCP2 mutants in FACS-based directed evolution.[109]

After enzymatic purification and desalting of the PCR product, the probe concentrations were adjusted to a final stock concentration of 142 nM. In agarose gel electrophoresis, 8NmC and 8NC probes shifted to higher molecular weight than the expected 52 bp due to the incorporation of two additional biotin-TEG handles at their 5'-ends (Figure 32b). Furthermore, we observed a faint side product of higher-molecular weight in the 8NmC probe. However, since the DNA concentration was determined by absorption at 260 nm, the total DNA mass concentration of both probes remained the same. The slightly reduced molar concentration of the 8NmC probe, caused by the side product, was considered to not interfere significantly with selection stringency.

To confirm the accurate incorporation of 5mC and unmodified cytosine and to validate the functionality of the two-color FACS system, we incubated 6 nM of each probe with *E. coli* (BL21) Tuner cells displaying hMBD2$_{151\text{-}214}$. We investigated probe binding in both combinations with the streptavidin-bound fluorophores (SAv-PE, SAv-FITC). The FACS data demonstrated the functionality of the probes and the dual-color system by the 5mC-specific binding of hMBD2$_{151-214}$ in both fluorophore-probe combinations (Figure 32c).

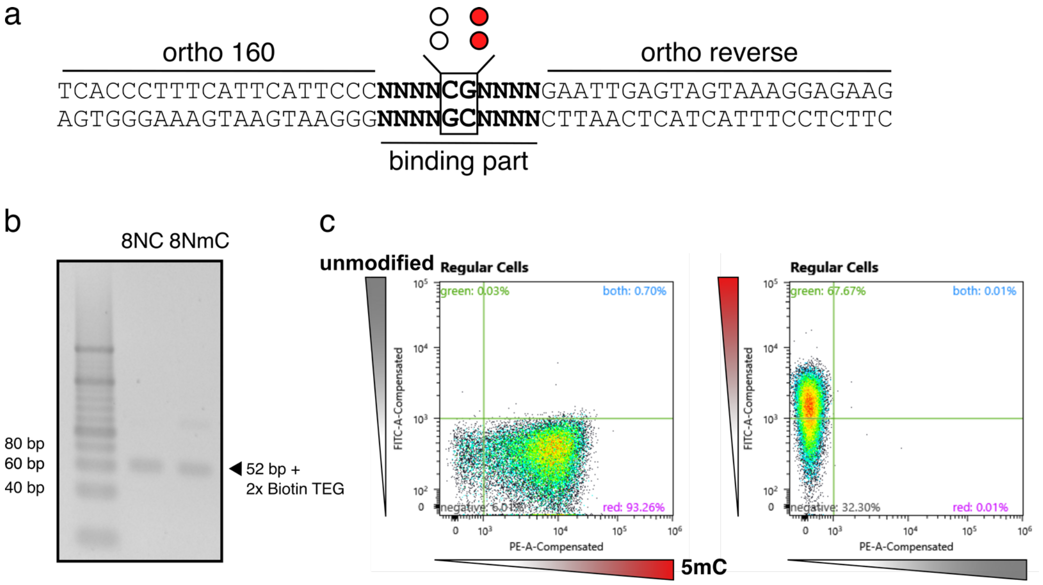

Figure 32: Structure, generation, and functionality of randomized probes (8NxC) used for 5mCpG binder screening. (a) 8NxC probe sequence. The central CpG contains either unmodified cytosine or 5mC. Note that the random flanking tetramers of the 8NmC probe contain 5mC as well. Ortho 160 and ortho reverse sequences were optimized to not incorporate cytosine during PCR. (b) Agarose gel electrophoresis of the purified biotinylated 8NmCC and 8NC probe stocks (stock concentration 142 nM). (c) 5mC-specific binding of *E. coli* (BL21) Tuner cells displaying hMBD2$_{151-214}$ after staining with SAv-FITC and SAv-PE labeled 8NmC and 8NC probes (6 nM each) in a two-color FACS assay.

In the previous section, we showed the first evidence of the intimin platform to facilitate the display of protein fragments larger than 50 residues and even full-length proteins (section 10.4.3 Table 4). Consequently, we proceeded with screening a full-length CDS library derived from human thyroid cDNA for reader candidates of 5mC. In the two-color FACS selection, we incubated *E. coli* (BL21) Tuner cells displaying the full-length proteins with a mixture containing 6 nM of the 8NmC probe (on-target) and the 8NC probe (off-target), along with 90 nM unlabeled DpH-background oligos. The gating parameters were based on the signal of cells expressing the empty entry vector pDmS2659 (Figure S4c) and were set strictly to achieve a clear differentiation between 8NmCpG and 8NCpG binding cells. To avoid interference of fluorophore-protein interactions with the selection process, we alternated the fluorophores of the two probes after each selection round. The initial selection round encompassed 1,466,611 individual clones, statistically representing a substantial portion of the library (79.8%, Table S15), as determined by Equation (1).

$$F = 1 - e^{-\frac{L}{V}}$$

Equation (1): Determination of the library coverage. F: fractional library coverage, L: number of screened clones, V: library diversity. The Equation derived from Patrick et al.[323]

During the screening process, a notable shift toward 8NmC-binding was observed, implicating the successful enrichment of DNA-binding proteins with a selectivity for 5mCpG over CpG (Figure 33a). However, an increase in binding of the unmethylated probe was observed after the second selection round. Assuming display levels similar to what Wentzel et al.[242] observed for the E-epitope, the methylated probe is expected to be fully bound by the enriched reader candidates (Equation S3). Consequently, the increased binding of the unmethylated probe can be attributed to a lack of competition for the binding sites with the methylated probe resulting in binding of certain readers to the unmodified CpG despite their preference for 5mCpG. However, this did not appear to impact the constant enrichment of reader candidates, as shown in Figure 33b. In future experiments, it is recommended to increase the probe concentrations to ensure a consistent saturation of binding sites in all selection rounds.

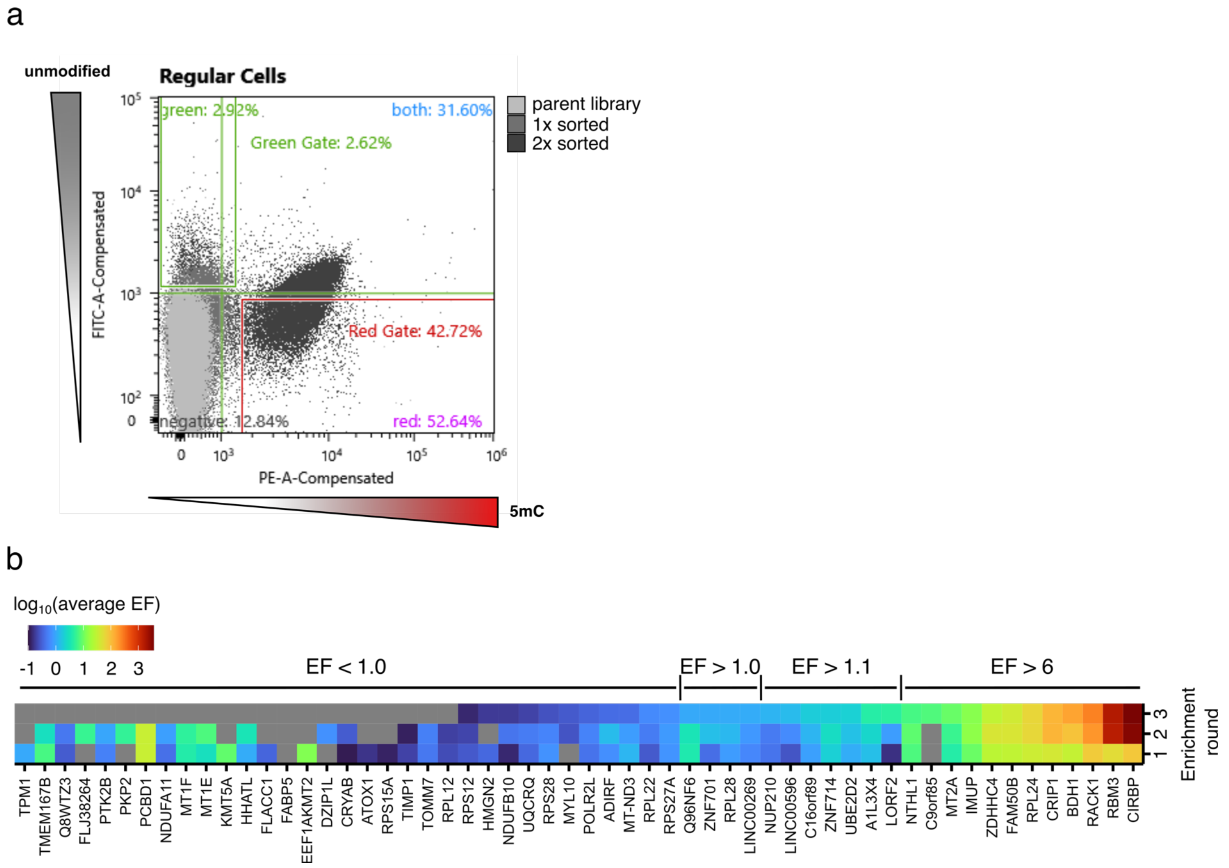

Figure 33: Enrichment of proteins of the bacterial surface display of the full-length human thyroid CDS library for 5mCpG selectivity. (a) Three rounds of FACS enrichment for on-target (8NmC) binding and off-target (8NC) discrimination. Cells were incubated with 6 nM of each probe and 90 nM of unlabeled DpH-background dsDNA oligo. The fluorophores of the on-target probe (PE) and the off-target probe (FITC) were reversed in the second selection round. The gating parameters were kept the same in each selection round. (b) Heatmap plot of the protein enrichment process throughout the three FACS selection rounds. Illumina reads were in silico translated and the resulting peptides were mapped against the human proteome. Enrichment factors (EF) of proteins were calculated from the normalized peptide count of each sub-library. The EF of grey tiles was not calculated due to low read numbers.

For a comprehensive analysis of the selection process, we subjected the cDNA inserts from each selection round and from the parent library to Illumina sequencing. The datasets were processed and EFs were calculated identically to section 10.4.3. It is important to note that the EF does not necessarily provide information about the degree of selectivity and may also encompass effects related to the display level and correct folding.

Many of the amplified cDNA inserts exceeded the recommended sizes for Illumina sequencing (Figure S5a), likely resulting in less efficient clustering before sequencing.

This impedes precise determination of the library composition due to an underestimation of these sequences. However, this size-related effect is expected be consistent for each cDNA insert across all sub-libraries, allowing for the calculation of EFs of the individual selection rounds. Nevertheless, we observed a continuous reduction in quality-control-passed reads (**Table S13**). This suggests an underlying layer of bias in the EFs introduced by the sequencing quality must be considered during evaluation of the results. We chose not to sequence fragmented amplicons to allow determination of the enriched protein parts, as these are expected to predominantly represent full-length proteins.

Similarly, to the enrichment of DNA binding proteins in section 10.4.3, we observed a major fraction of depleted proteins (EF < 1.0), as well as populations of rather unaffected (EF 1.0-4.6) and highly enriched proteins (EF > 6.0) after three selection rounds (Figure 33b). As expected, 50% of the highly enriched proteins were considered as full-length proteins, and two-thirds of the protein fragments exhibited extended ORFs exceeding 40 residues (Table **5**). The display of full-length proteins can be attributed to the 5′ UTR removal and in-frame cloning facilitated by targeting the Kozak sequence during library generation. However, the presence of protein fragments suggests that the anti-Kozak primers also annealed to other regions of the transcripts. Furthermore, the enrichment of protein fragments may indicate a preferred display of passengers of lower complexity as discussed in sections 10.4.2 and 10.4.3. This is further supported by the relatively small size of the displayed full-length proteins (Figure 34). However, the majority of the enriched fragments lacked corresponding full-length proteins in the parent library, which could have also been displayable with the intimin platform.

The majority of proteins enriched for 5mC selectivity are typically located within the cell nucleus and/or are involved in nucleic acid binding or chromatin organization. This strongly suggests the successful enrichment of 5mC reader candidates. Noticeably, we did not observe the enrichment of known 5mC readers, such as MBDs, in our screening. This can be attributed to the absence of hMBD1, hMBD2, and hMBD4 in the parent library (Table S14). The parent library only contains hMBD3, which lacks affinity for 5mCpG, and an N-terminal fragment of MeCP2 (residues 448-484), which lacks the methyl-binding domain.[58,324] Interestingly we found the RNA-binding proteins CIRBP, RBM, RACK1, and RPL24 to be enriched. While the observation of RNA-binding proteins being enriched for 5mCpG may seem unusual, it aligns with the findings from Spruijt et al., where a substantial proportion (8.5%) of 5mCpG interactors in neuronal progenitor cells (NPC) were found to be RNA binders.[73]

Some known DNA-binding proteins, such as HMGN2 (residues 1-29) or ZNF714 (residues 530-555), were either depleted or rather unaffected by the selection, as their fragments inadequately covered DNA binding domains (see Table S16 and Table S17).[294,325] Others, like full-length POLR2L, may not have been enriched due to their inability to bind 5mCpG, insufficient display, or issues related to surface folding. Unfortunately, these factors cannot be distinguished based on this dataset. Future screenings that inverse the selection towards the preferred binding of the unmodified CpG, can identify anti-reader candidates encoded in the library.

Table 5: Proteins enriched for 5mCpG affinity (EF > 6), their cellular function and localization as well as the identified or potentially covered domains and structures. If not mentioned otherwise, information on the proteins was obtained from the UniProt database.[294] Mapped residue values represent the mean and standard deviation of all mapped ORFs of the respective protein. *Real ORF might be extended due to detection limitations by sequencing length and trimming.

Protein	(Main) Cellular function	Cellular localization	Mapped residues	Potentially full-length	Potentially covered domains	Final EF
CIRBP	RNA binding	Nucleus, cytoplasm	1.0±0.1 - 39.3±3.0*	Yes	RRM, RGG domain[326]	3642.4
RBM3	RNA binding	Nucleus, cytoplasm	1.0±0.1 - 39.9±0.8	Yes	RRM, RGG domain[327]	1727.4
RACK1	RNA binding	Nucleus, cytoplasm, PM	1.0±0.1 - 23±0.1	No	First WD repeat (partially)	256.8
BDH1	Dehydrogenase	Mitochondrial membrane	308.4±0.9 - 336.3±2.3	No	None	157.3

CRIP1	Zinc absorption and transport, transcriptional regulation	Nucleus[328], Cytoplasm	1.0 - 40.0±0.6*	Yes	LIM domain	117.4
RPL24	rRNA binding	Cytoplasm	1.0 - 40.0*	Yes	Basic and acidic residues	62.3
FAM50B	Chromatin organization	Nucleoplasm	280.0±0.2 - 320.9±1.7*	No	None	39.1
ZDHHC4	Lipidation	Golgi membrane, ER[329],(Outer) nuclear envelope[330]	281.0 - 322.7±2.5*	No	Di-lysine motif	30.3
IMUP	Unknown	Nucleus	1.0 - 40.0*	Yes	Basic and acidic residues (NLS)	13.3
MT2A	Zinc binding	Nucleus[331], Cytoplasm	1.0 - 27.5±9.6*	Yes	Metal ion cluster A & B	9.4
C9orf85	Unknown	Nucleus[332]	133±19.9 - 165±15.6*	No	Unknown	7.3
NTHL1	DNA N-glycosylase	Nucleus, Mitochondrion	256.0±0.1 - 305.0±0.1*	No	Iron-sulfur cluster[333]	6.5

A total of 12 proteins have been enriched for 5mCpG selectivity (Table 5). These proteins showed wide-ranging EF between 6.5-fold (NTHL1) and over 3,600-fold (CIRBP). Notably, the RNA-binding and stabilizing protein CIRBP and the DNA glycosylase NTHL1 both demonstrated a preferred binding to 5mC over unmodified cytosine in NPC (CIRBP) and mouse embryonic stem cells (mESC, NTHL1) in the pulldown study of Spruijt et al.[73] Moreover, CIRBP has been shown to exhibit 5mC preference in mRNA binding.[334] It should be noted that CIRBP also bound 5mC in the pulldown study of Iurlaro et al.[72] but showed higher affinity for 5fC. Overall, these related findings provide strong evidence of the functionality of the screening system for selecting 5mC reader candidates, although the enrichment of MBDs and other known 5mC readers was not observed. The following paragraphs are dedicated to discussing the potential significance of 5mC selectivity exhibited by enriched proteins in the context of their known molecular and cellular functions.

CIRBP was enriched as a full-length protein, suggesting that the N-terminal RRM might be responsible for the 5mC preference. However, it cannot be excluded that the centrally located RRG motif is also involved in this interaction (Figure 34), as it has been found to mediate topology-selective binding of CIRBP to G-quadruplex structures and to be involved in binding the 5′ UTR and 3′ UTR of mRNA.[326,327] An in vitro study of the chick homologue of CIRBP, APBP-1, revealed dsDNA binding and repressive transcription factor activity. This suggests that similar functions of human CIRBP might exist as well and could be related to potential binding of genomic 5mC sites.[335] It has to be mentioned that the peptides mapped to RBM3, are likely attributed to CIRBP as well, as both proteins share a highly conserved N-terminus which was mapped against the human proteome. Furthermore, RBM3 was not detected in Sanger sequencing analysis of enriched single clones, which exclusively revealed CIRBP ORFs (Table S18). Nevertheless, considering the structural similarities of both proteins, RBM3 might also exhibit 5mCpG affinity and has further demonstrated binding of 5mC-containing mRNA.[334]

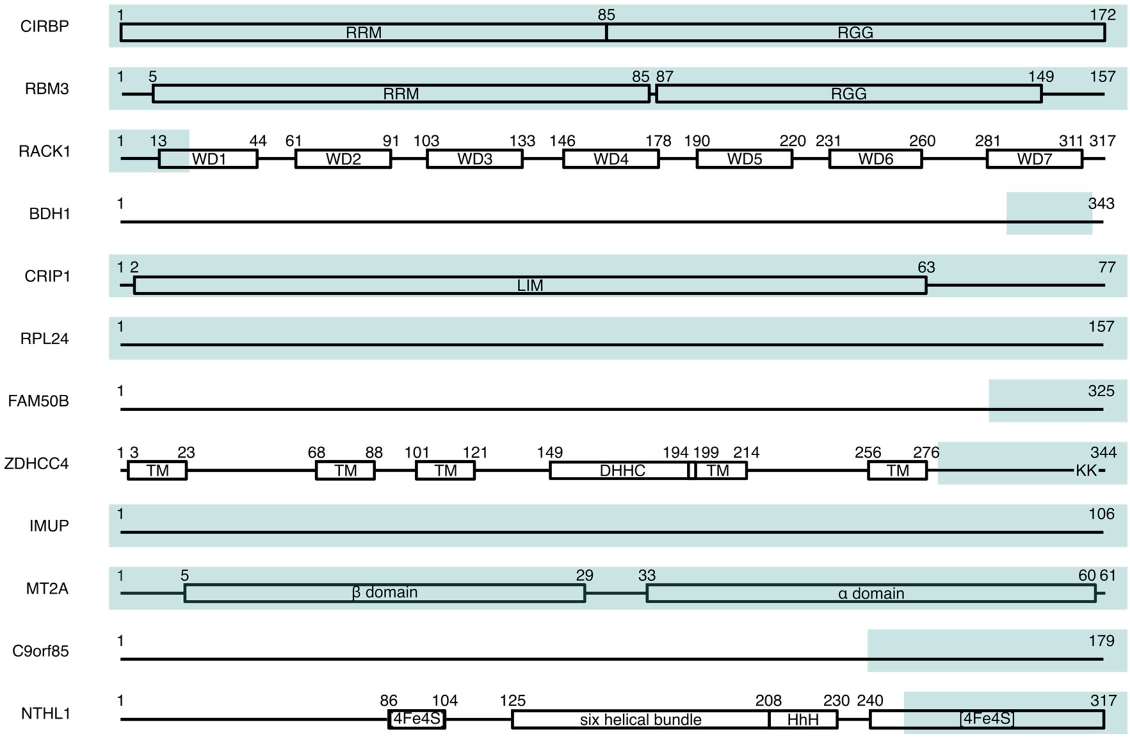

Figure 34: Schematic representation of the domain structure of the enriched proteins. Boxes represent the respective domains. Proteins that are only depicted by a line have no domain information available. Light blue boxes represent the enriched part of the proteins, assuming the extension of Illumina sequencing-identified ORFs. Structural information was obtained from the UniProt database[294] and publications of Kim and Hong[336], Hu et al[337], Jayawardene et al.[338] and Carroll et al.[339] RRM: RNA recognition motif, RGG: RGG motif, WD: Try-Asp repeat, LIM: LIM domain, TM: transmembrane domain, DHHC: Asp-His-His-Cys domain, [4Fe4S]: Iron-sulfur cluster domain, HhH: Helix-hairpin-helix motif.

In contrast to CIRBP, only a C-terminal fragment of NTHL1, a DNA glycosylase that recognizes lesions containing oxidized pyrimidine bases[339,340], was enriched in the screening. Reviewing the Illumina deep sequencing data revealed that the full-length ORF was not encoded in the parent library (Table S14). The enriched fragment covers the C-terminal part of the [4Fe4S] domain and contains all residues that form the iron-sulfur-cluster itself (Figure S9).[333,340] Notably, the HhH DNA binding motif is not included. The [4Fe4S] domain interacts with the DNA backbone via the residue Q287 and is rather involved in scaffolding functions than in DNA lesion interaction and detection.[339,341] It is up to further investigations to determine if the relatively low EF (EF = 6.5) is based on a preferred binding of the [4Fe4S] domain fragment to the slightly narrowed backbone in DNA harboring 5mC.[342,343] Similar shape recognition mechanisms of the DNA backbone have been speculated to contribute to the 5mC preference of other reader proteins as well.[93]

Whereas the fragment covering the N-terminus of BDH1 cannot be attributed to any structural or functional feature, the 23-residue long fragment of RACK1 covered the N-terminal part of the first tryptophan-aspartate (WD) repeat (Figure 34, Figure S9). Notably, the parent library contained further, longer ORFs of RACK1 (Table S14), implicating that the enriched fragment was selected on its intrinsic properties. Nonetheless, BDH1 and RACK1 are neither known to obtain any nucleic acid binding properties nor do the size and sequence of the corresponding enriched peptides suggest any novel characteristics that would allow for the preferential binding of modified DNA.[344] Additionally, BDH1 had one of the lowest overall read count among the enriched proteins (n = 25) after selection. Although RACK1 is involved in the regulation of chromatin remodeling by binding to repressed promoter regions, this process is known to be mediated by interactions with histones instead of binding to modified genomic DNA.[344] Similarly, the enrichment of the ribosomal protein RPL24 and the lipid transferase ZDHHC4 was not expected. Although RPL24 exhibits RNA-binding properties and ZDHHC4 has been unfoundedly referred to as transcription factor[345] there is no evidence that these proteins are located within the nucleus and could therefore exhibit 5mC selectivity. The covered di-lysine motif of ZDHHC4 is solely involved in the ER localization of the protein and does not participate in nucleic acid interactions.[329] Notably, RACK1, BDH1, ZDHHC4 fragments, and full-length RPL24 were not co-enriched due to interactions with reader proteins, since we selected for single cells, only expressing a single protein, during FACS. So far, the exact reasons for the enrichments remain unclear.

The enriched protein FAM50B is involved in transcriptional regulation and might further contain chromatin regulatory function like its *Chlamydomonas reinhardtii* and *Schizosaccharomyces pombe* orthologue XAP5.[346,347] FAM50B and XAP5 proteins are highly conserved in human and mouse, including the here enriched C-terminus of the protein.[348] The AlphaFold model of FAM50B suggests the enriched fragment to be part of a C-terminal domain. However, folding of the fragment in a biological functional structure is rather unlikely (Figure S9). Furthermore, ORFs including the entire protein or the entire domain were not present in the parent library (Table S14). In general, the involvement of FAM50B and the putative domain in 5mC-specific binding is plausible in the context of chromatin regulation, so further investigation of our findings could provide valuable insights into the regularity functions of the protein.

Although it is primarily localized in the cytoplasm, MT2A was also found to be located in the nucleus during early development.[331,349] The metal ion-chelating protein is mainly responsible for ion transport, but there is further evidence that it is involved in transcription factor activation and mitochondrial genome preservation via free-radical scavenging.[350,351] Our data suggests, that MT2A could exhibit additional dsDNA binding properties in the context of 5mC. The two cysteine-rich metal ion binding domains (α and β) of the enriched, full-length protein likely contribute to this interaction. Whether the finding is biologically meaningful in comparison to the other functions of MT2A needs to be investigated further. Notably, we did not enrich the alternative ORF of MT2A which has been found to show transcriptional activator properties.[352]

CRIP1 was enriched as a full-length protein including its LIM domain which is involved in PPI-mediated transcriptional regulation.[353] Early studies of Pérez-Alvarado et al.[328] suggested potential dsDNA binding properties of CRIP1 due to structural similarities to LMO2 and GATA-1. Since 2014 a CRIP1 diagnostic assay, developed by Xie et al.[354], makes use of the dsDNA binding properties of the protein. These properties, along with its differential expression in various cancer types and the potential 5mC-preference observed in our study, suggest that CRIP1 may have additional functions beyond PPI-mediated transcriptional regulation and the regulation of DNA damage repair by homologous recombination.[328] Whether the binding of genomic 5mCpG represents an additional level of transcriptional regulation by CRIP1 or contributes to other functions in the nucleus is up for further investigation.

There is very limited information on IMUP (immortality upregulated protein) and C9orf85 available. Both proteins lack comprehensive structural information, except for the presence of a C-terminal NLS (nuclear localization signal) of IMUP. [355,356] In this study, IMUP has been enriched as a full-length protein (Figure 34). Additional to its upregulation in immortal cells, such as cancer cell lines, a recent study has found IMUP to be involved in cell cycle regulation by stabilizing the epigenetic regulator NPM1.[357,358] Although in vitro binding to nucleic acids has been thought to be limited to poly (rG) RNA[358] our finding suggests selective interaction of IMUP with methylated dsDNA which could contribute to yet unknown regulatory functions of the protein. A C-terminal fragment of C9orf85 was found to be enriched for 5mC affinity (Figure S9), although it should be noted that the counts of mapped peptides barely matched the minimal criteria of our analysis. Interaction studies suggested PPI of C9orf85 with nucleoli-associated proteins in response to cellular stress.[332,359] The potential involvement in genome binding at 5mCpG sites could serve as a good starting point to further elucidate its functional roles within the nucleus.

For all the enriched reader candidates further in vivo and in vitro experiments, such as EMSA and ChiP-seq, are indispensable to confirm their interaction profiles and elucidate their cellular roles. Structural information, such as that obtained from NMR studies or protein crystallography, could provide insides into the mechanism of methylated DNA recognition. Furthermore, since our randomized probes were designed to minimize influences by sequence preferences, it would be of interest to investigate methyl consensus sequences of the identified reader candidates.

In conclusion, the initial functional test of the bacterial surface selection system successfully demonstrated its capability to discover 5mC reader candidates from human cDNA. Our evaluation is based on pulldown studies of Spruit et al.[73], which likewise identified CIRBP and NTHL1 as reader candidates of 5mC. Additionally, we identified novel 5mC reader candidates like CRIP1 and IMUP, whose nuclear localization and known functions align with this potential new function. Nevertheless, we failed to identify known 5mC reader proteins, of which the MBDs were not covered by our prototype display library. A more comprehensive interactome coverage could be achieved by increasing the diversity of the display library and reducing its bias. Furthermore, the diversity of the displayed proteome depends on efficient translocation and surface folding. Screening fragmented cDNA display libraries can provide access to additional reader protein domains. Moreover, reducing disulfide bridges of passenger proteins, e.g., by adding reducing agents like 2-mercaptoethanol or by using dsbA-deficient strains, is known to improve surface translocation.[265] Supplementing metal ions to assist in protein folding can further improve the functionality of displayed proteins. However, it is important to notethe affinity of certain 5mC reader proteins, such as MBD2 and MeCP2, is known to be regulated by PTMs.[360,361] Since mammalian PTM patterns cannot be introduced by *E. coli*, readers whose activity relies on these modifications might be inaccessible in our system.

This study demonstrates the capability of our selection tool to enrich 5mC-reader candidates. It thereby benefits from the bacterial surface display of cDNA-encoded protein libraries, enabling fast and iterative selections without the requirement of protein purification. Furthermore, it provides access to the ORFs of reader candidates for downstream analysis purposes. In future, improved display libraries and conditions can offer insights into the interactome of modified cytosine, complementing existing pull-down-MS studies.[68,72,73]

Summary

Our study aimed to establish a bacterial surface display platform to discover 5mC readers from cDNA libraries. We subjected a prototype cDNA display library to iterative selection via a two-color FACS system, enabling on- and off-target selectivity. The platform's functionality was validated by successfully enriching the 5mC reader candidates CIRBP and NTHL1, consistent with the pulldown-MS results of Spruijt et al.[73] Additionally, we identified novel reader candidates, such as CRIP1 and IMUP. However, it should be noted that certain enriched proteins and fragments lacked an apparent connection to DNA binding. Furthermore, we were unable to enrich well-studied 5mC readers such as MBDs, which was based on their absence in the display library or on potential limitations in their surface display by the intimin platform. This suggests the need for future improvements regarding the library diversity and display performance to increase the surface-accessible proteome and provide more comprehensible insights. Furthermore, the probe concentration should be increased in future screenings to ensure sufficient accessibility of free probes in all selection rounds.

In conclusion, the first test screening for 5mC-selective reader candidates validated the functionality of our selection system and demonstrated its potential to complement existing studies on readers of 5mC and its oxidized derivates. It thereby benefits from the iterative and fast screening process and allows for the detection of direct reader candidates independently from their original expression levels.

11 Conclusion and Outlook

We successfully established a bacterial surface display selection platform to enrich known and novel reader candidates of symmetrically methylated CpGs encoded in human cDNA libraries. In this course, we were the first to demonstrate the compatibility of a bacterial platform for the surface display of protein libraries encoded in human cDNA.

For this purpose, we created prototype libraries from human prostate and thyroid cDNA that allowed for the display of full-length and fragmented proteins via the intimin reverse autotransporter platform[242]. Initially, we could show that the heterologous expression and display of human proteins or protein fragments had no significant effect on the cell's viability. Assessment of the systems display scope, demonstrated the feasibility of displaying protein fragments and functional full-length proteins, but also suggested improvements regarding the library diversity and display performance to increase the surface-accessible proteome.

Building on these results, the function test showed the successful enrichment of reader candidates of symmetrically methylated CpGs. Thereby, we made use of the rapid and iterative selection process enabled by the fast bacterial growth and protein expression. By labeling the methylated and unmethylated probes with distinct fluorophores, we could precisely select for on-target and off-target selectivity of binding proteins via FACS. The functionality of the selection system was demonstrated by the enrichment of the 5mC reader candidates CIRBP and NTHL1, which have also been found in previous proteome-wide pulldown studies.[73] However, our results lacked the enrichment of known 5mC readers, such as MBDs, which was based on their absence in the display library and on potential limitations in their surface display by the intimin platform. It should be noted that mammalian PTM patterns are not introduced by the *E. coli* expression machinery so readers that activities rely on these modifications are likely missed by our system. Nevertheless, we identified novel reader candidates like CRIP1 and IMUP. This suggests that our system could have addressed the limitations associated with pulldown approaches, such as competition among proteins for the probe and low endogenous expression of certain proteins. Since every protein was displayed on the surface of an individual cell, the enrichment can be attributed to the direct interaction of the proteins with the methylated DNA. However, it is essential to further characterize all identified reader candidates in vitro and in vivo to unravel their complete affinity profiles and cellular roles.

In summary, we have developed a bacterial surface display selection system for the screening of cDNA-encoded human protein libraries for readers of epigenetically modified CpG dyads. Additionally, anti-readers, which are repelled by the CpG modification, can be targeted by reversing the selection criteria. At the same time, our system is independent of the endogenous expression levels in the cellular origin, reduces competition between proteins, solely selects direct binding events, and does not require protein purification. It further offers ready access to the ORFs of reader candidates for downstream analysis. In general, future screenings for 5mC reader candidates would benefit from an increased surface-accessible proteome. This can readily be achieved by utilizing more diverse display libraries than the ones employed for the function test in this work. In addition, the usage of dsbA-deficient strains or reducing agents has already been proven to enhance the translocation efficiency of disulfide bridge-containing proteins, and metal ion additives could improve folding on the surface.[265] Furthermore, we expect that our approach can be extended to identify novel reader candidates of the oxidized 5mC species, thereby complementing existing interactome studies[68,69,72,73,105], as well as of the various modification combinations in CpG dyads that have not been investigated so far.

12 Material and Methods

12.1 Materials

Lab equipment and instruments

Type	Model	Manufacturer/Company
Agarose Gel Electrophoresis System	Midicell primo EC330	ThermoEC
Analytical balance	M-pact AX224	Sartorius
Autoclave	Varioklav	HP Labortechnik
Autoclave	VX-150	Systec
Balance	Science Education	VWR
Bunsen burner	1010	Usbeck
Centrifugal vaccum concentrator	Concentrator plus	Eppendorf
Cooling tabletop microcentrifuge	5427 R	Eppendorf
Cooling water chiller	Minichiller 300	Huber
Cooling tabletop centrifuge	5810 R	Eppendorf
Dewar flask	Type G-C	KGW Isotherm
Electroporator	Eporator	Eppendorf
Ethidium bromide staining bath	Steel chamber 18/10	Bochem
Flake ice maker	UFP 0399 A	Manitowoc
Fluorescence-activated cell sorter	SH800 SGP	Sony biotechnology
Freezer (-20 °C)	Profi line GG4010	Liebherr
Fridge	Profi line FKU 1800	Liebherr
Incubating shaker	I26	New Brunswick Scientific
Incubator	INCU-Line	VWR
Laboratory water purifier	PURELAB flex 2	Veolia Water Systems
Magnetic stirrer	Hei-Mix S	Heidolph Instruments
Magnetic stirrer	MR Hei-Standard	Heidolph Instruments
Magnetic stirrer	CB161	Stuart
Microliter pipette (10 µL, 100 µL, 200 µL, 1000 µL)	Research Plus	Eppendorf
Microwave	ED 8525.3S	Exquisit
Minicentrifuge	Sprout plus	Heathrow Scientific

Multi-dispenser pipette	Multipette plus	Eppendorf
Nanopore sequencing device	MinIon	Oxford Nanopore Technologies
PCR cleanhood	PCR workstation pro	Peqlab
PCR thermocycler	TOne	Biometra
PCR thermocycler	T-Personal	Biometra
PCR thermocycler	MyCycler	Bio-Rad
PCR thermocycler	SimpliAmp	Applied biosystems
pH meter	Five easy	Mettler Toledo
Photometer	BioPhotometer plus	Eppendorf
Pipette controller	Accu-jet pro	Brand
Power Supply for gel electrophoresis	PowerPac Basic	Bio-Rad
Power Supply for gel electrophoresis	EV233	Consort
Scanner	CanoScan 9000F	Canon
Sonication device	Biorupter pico	Diagenode
Spectrophotometer	Nanodrop 2000	Thermofisher scientific
Tabletop microcentrifuge	MiniStar	VWR
Tabletop microcentrifuge	5424	Eppendorf
Thermomixer	ThermoStat plus	Eppendorf
Thermomixer	ThermoMixer C	Eppendorf
Ultra deep freezer (-80 °C)	U725-G Innova	New Brunswick Scientific
UV Imager	BDAdigital compact	Biometra
UV Transilluminator	UVstar 312 nm	Biometra
Vortex mixer	Vortex-Genie 2	Scientific Industries
Camera	PowerShot G10	Canon

Services

Service	Company
Illumina Sequencing	Novogene (UK, Cambridge)
Oligonucleotide synthesis	Sigma Aldrich (Merck, Darmstadt)
Sanger Sequencing	Microsynth Seqlab (Germany, Göttingen)

Software

Name	Purpose	Developer/Distributor
BioDoc Analyze 2.1	UV light imaging	Jena Bioscience
biomaRt 2.52.0	R package for conversion of database nomenclature	Durinck and Huber
Biostrings 2.64.1	R package for data analysis	Pagès et al.
BLAST+ suite 2.14.0	proteome mapping via blastp	National Center for Biotechnology Information, U.S.
BlastN and BlastP	Sequence alignment to databases	National Center for Biotechnology Information
Bowtie2 2.5.1	Illumina read transcriptome mapping	Langmead et al.
Cell Sorter Software 2.1.5	FACS analysis and sorting	Sony biotechnology
ChemDraw 22.2.0	Drawing of chemical structures	Perkin Elmer
data.table 1.14.8	R package for data analysis	Dowle and Srinivasan
DP-bind as of 07/2023	predicition of nucleic acid interactions	Hwang, Gou and Kuznetsov
dplyr 1.1.0	R package for data analysis	Posit, PBC
DRNApred as of 07/2023	predicition of nucleic acid interactions	Yan and Kurgan
Excel	Data analysis	Microsoft
fastp 0.20.1	Illumina read quality control and trimming	OpenGene
GenSmart Optimization as of 09/2020	Codon optimization for heterologous expression	GenScript
ggbreak 0.1.2	R package for plotting	Guangchuang Yu
ggplot2 3.4.2	R package for plotting	Posit, PBC

ImageJ 1.52p	image analysis	National Institutes of Health, U.S.
Inkscape 1.3	Figure design	The Inkscape Project
Mendeley Reference Manager 2.103.0	Citation Manager	Elsevier
microseq 2.1.5	R package for data analysis	Snipen and Liland
minimap2 2.12-r828-dirty	Nanopore read transcriptome mapping	Dana-Farber Cancer Institute
MinKnow 19.12.2	Base calling of nanopore reads	Oxford Nanopore Technologies
NanoDrop 2000 Software 1.6.198	Analysis of DNA absorption data	Thermo Fisher Scientific
NEBuilder Assembly tool 2.2.5 et seq.	Design of Gibson assembly primers	NEB
porechop 0.2.3	Nanopore read trimming	Ryan Wick
R 4.2.1	Data analysis and plotting	*R Core Team*
Rsamtools 2.12.0	R package for data analysis	Morgan et al.
RStudio 4.3.0	Data analysis and plotting	Posit
samtools 1.6	data conversion	Genome Research Limited
SnapGene 4.3.11	Sanger sequencing data analysis, Plasmid map design	GSL Biotech LLC
stringr 1.5.0	R package for data analysis	Wickham, RStudio
svglite 2.1.1	R package for plotting	Posit, PBC
tidyverse 1.3.0	R package for data analysis	RStudio
Tm Calculator 1.12.0 et seq.	Annealing temperature calculation	NEB
UCSF ChimeraX 1.6	Protein structure visualization	RBVI
Virtual Ribosome 2.0	in silico translation of DNA sequences	DTU Health Tech
Word	Manuscript writing	Microsoft

Disposables and glassware

Name	Manufacturer/Distributor
1.5 mL Bioruptor Pico Microtubes	Diagenode
Bottletop filter 0.2 µm, 500 mL (Filtropur)	Sarstedt
Combitips advanced (2.5 mL)	Eppendorf
Conical plastic tubes (15 mL)	Sarstedt
Conical plastic tubes (50 mL)	Sarstedt
Cuvettes, standard	Sarstedt
DNA LowBind tubes (1.5 mL)	Sarstedt
Electroporation cuvettes (1 mm)	Carl Roth
Erlenmayer flasks (100 mL)	Schott
Glass beads	VWR
MµultiGuard barrier tips (10 µL, 100 µL)	Sorenson Bioscience
Measuring cylinder (100 mL, 1 L)	Hirschmann
Multiply-Pro 0.2 mL reaction tube	Sarstedt
Multiply-Pro 0.2 mL reaction tube 8-strip	Sarstedt
Nitrile gloves	VWR & Arnowa
Onewell plate (127.8/85.5 mM) for FACS	Greiner bio-one
Petri dish (150x20 mm)	Sarstedt
Petri dish (92x16 mm)	Sarstedt
Pipette tips (10 µL, 200 µL, 1000 µL)	Sarstedt
SafeSeal reaction tubes (1.5 mL)	Sarstedt
SafeSeal Reaction tubes (2 mL)	Sarstedt
Scalpel	B.Braun
Schott flasks (10 mL, 25 mL, 50 mL, 100 mL, 250 mL, 500 mL, 1000 mL)	Duran & VWR
Serological pipette (10 mL, 25 mL)	Sarstedt
Syringe filter (0.2 µM)	Sarstedt

Consumables

Name	Manufacturer/Distributor
1 kb plus DNA ladder	New England Biolabs (NEB)
Automatic setup beads for Cell Sorter SH800	Sony biotechnology
CutSmart buffer 10 x	NEB
dNTP mix (10 mM with dCTP substituted with 5mdCTP)	Zymo Research
dNTP mix (10 mM)	NEB
Flonge Flow cell	Oxford Nanopore Technologies
Human prostate cDNA library 10108-A	BioCat
Human thyroid cDNA library HD-503	Zyagen
KOD buffer 10 x	Merck Millipore
KOD polymerase buffer 10 x	Merck Millipore
LB agar (Lennox)	Carl Roth
LB-medium (Lennox)	Carl Roth
Microfluiding sorting chip (100 µM)	Sony biotechnology
O'RangeRuler 20 bp DNA ladder, ready-to-use	Thermo Scientific
Phusion GC buffer 5 x	NEB
Phusion HF buffer 5 x	NEB
purple DNA loading dye 6 x	NEB
Q5 buffer 10 x	NEB
rCutSmart buffer 10 x	NEB
Spot On Flow Cell MK 1 R9	Oxford Nanopore Technologies
Streptavidin-Fluorescein Isothiocyanate (SAv-FITC)	Biolegend
Streptavidin-Fluorescein Phycoerythrin (SAv-PE)	Biolegend
T4 ligase buffer10 x	NEB
Thermo Polymerase buffer 10x	NEB

Commercial Kits and Mastermixes

Name	Manufacturer/Distributor
Ligation Sequencing Kit (SQK-LSK110)	Oxford Nanopore Technologies
Flow Cell Priming Kit	Oxford Nanopore Technologies
NEBuilder HiFi assembly master mix	NEB
1.33 x Self-made Gibson assembly mastermix	Summerer lab
NucleoSpin Gel and PCR Clean-up kit	Macherey & Nagel
NucleoSpin Plasmid DNA isolation kit	Macherey & Nagel
Quick Blunting Kit	NEB

Chemicals

Name	CAS Number	Supplier
2-(4-(2-Hydroxyethyl)-1-piperazine)-ethanesulfonic acid (HEPES)	7365-45-9	Carl Roth
1,4-Dithiothreitol (DTT)	3483-12-3	Carl Roth
Agarose LE (molecular biology grade)	9012-36-6	Biozym
Ammonium sulfate	7783-20-2	Carl Roth
ATP (25 mM)	-	Lucigen
Boric acid	10043-35-3	Carl Roth
Bovine serum albumin (ultra pure)	048-46-8	Cell Signaling Technology
Calcium chloride	10043-52-4	Fisher Scientific
Chloramphenicol	56-75-7	Carl Roth
D(+)-Biotin	58-85-5	Carl Roth
D(+)-glucose (molecular biology grade)	50-99-7	Carl Roth
dATP (100 mM)	-	NEB
dCTP (100 mM)	-	NEB
dGTP (100 mM)	-	NEB
Dimethyl sulfoxide (DMSO)	67-68-5	Carl Roth
dTTP (100 mM)	-	NEB
Ethanol, abs. (HPLC grade)	64-17-5	Fisher Scientific
Ethidium bromide solution (1%)	1239-45-8	Carl Roth
Ethylenediaminetetraacetic acid (EDTA) disodium salt dihydrate	6381-92-6	Carl Roth
Glycerol	56-81-5	Carl Roth
Isopropanol	67-63-0	Fisher Scientific

L(+)-arabinose	5328-37-0	Carl Roth
Magnesium sulfate heptahydrate	heptahydrate	Merck
Magnesium chloride hexahydrate	7791-18-6	Acros
Polyethylene glycol 8000 (PEG-8000)	25322-68-3	Promega
Potassium chloride	7447-40-7	Carl Roth
Sodium acetate	127-09-3	Sigma Aldrich
Sodium chloride	7647-14-5	Merck
Tris base	77-86-1	Carl Roth
Tris(2-carboxyethyl)phosphine hydrochloride (TCEP)	51805-45-9	Carl Roth
Triton X-100	9002-93-1	Fluka
Tryptone	91079-40-2	Carl Roth
Yeast extract	8013-01-2	Carl Roth
β-Nicotinamide adenine dinucleotide hydrate (NAD)	53-84-9	Sigma Aldrich

Enzymes

Name	Manufacturer/Distributor
AfeI	NEB
DpnI	NEB
HindIII-HF	NEB
KOD polymerase	Merck Millipore
NdeI	NEB
Pfu polymerase	Summerer group
Phusion-HF DNA polymerase	NEB
Q5 polymerase	NEB
Shrimp alkaline phosphatase (rSAP)	NEB
SpeI-HF	NEB
T4 DNA ligase	NEB
T4 polynucleotide kinase	NEB
T5 Exonuclease	NEB
Taq DNA Ligase	NEB
Taq DNA polymerase	NEB
TEV-protease	Summerer group
XhoI	NEB
SalI-HF	NEB

Antibodies

Antibody	Source	Manufacturer/Distributor
anti-Myc APC antibody 223896	Mouse	Abcam
anti-Myc APC antibody ab72580	Mouse	Abcam

Cell strains

E. coli strain	Genotype	Origin
GH371	F$^-$ *mcrA* Δ(*mrrhsdRMSmcrBC*)φ80*lac*ZΔM15 Δ(*lac*)X74 *recA1 endA1 araD139* Δ(*araAleu*)7697 *galU galK rpsL*(Str^r) *nupG fhuA::IS2 upp*$^-$	obtained from J. W. Chin
Tuner™(DE3)	F$^-$ *hsd*S$_B$ (r^-_B m^-_B) *gal dcm ompT lacY1*(DE3)	Novagen™ Merck
DH10B (TOP10™)	F$^-$ *mcrA* Δ(*mrrhsdRMSmcrBC*)φ80*lac*ZΔM15 Δ(*lac*)X74 *recA1 endA1 araD139* Δ(*araA-leu*)7697 *galU galK rpsL*(Str^r) *nupG*	Invitrogen™ Thermo Fisher Scientific

Buffers and Solutions

Name	Components
Cam Stock	34 mg mL^{-1} chloramphenicol in ethanol
CutSmart Buffer 1 x (NEB)	20 mM Trisacetate, 50 mM KOAc, 10 mM Mg(OAc)$_2$, 0.1 mg mL^{-1} BSA, pH = 7.9
EMSA buffer 10 x	200 mM HEPES, 300 mM KCl, 10 mM EDTA, 10 mM (NH$_4$)$_2$SO$_4$
Gibson master mix 1.33x	2.66x isothermal buffer, 6.66 mU µL^{-1} T5 exonuclease, 41.66 mU µL^{-1} Phusion DNA polymerase, 6.66 U µL^{-1} Taq DNA ligase
IDT annealing buffer 10 x	300 mM HEPES, 1 M KOAc
Isothermal reaction buffer 5 x	25 % PEG-8000, 500 mM Tris-HCl pH 7.5, 50 mM MgCl$_2$, 50 mM DTT, 1 mM dATP, 1 mM dTTP, 1 mM dCTP, 1mM dGTP, 5 mM NAD

LB Medium	10 g L^{-1} tryptone, 5 g L^{-1} yeast extract, 10 g L^{-1} NaCl, pH = 7
PBS 1 x	10 mM Na_2HPO_4, 1.8 mM K_2HPO_4, 137 mM NaCl, 2.7 mM KCl, pH = 7.2
Pfu buffer 10 x	400 mM Tris, 500 mM KCl, 4 mM $MgCl_2$, 4% Triton X-100, pH = 8.8
Purple loading dye	NEB
SOC medium	0.58 g L-1 NaCl, 2.03 g L-1 MgCl2 hexahydrate, 2.46 g L-1 MgSO4 heptahydrate, 5 g L-1 yeast extract, 20 g L-1 tryptone, 1 M glucose, pH = 7.5
T4 DNA ligase buffer 1 x (NEB)	50 mM TrisHCl, 10 mM MgCl2, 10 mM dithioerythritol, 1 mM adenosine triphosphate, pH = 7.5
TBE Buffer 1 x	89 mM Tris base, 89 mM boric acid, 2 mM EDTA, pH = 8.3

DNA oligonucleotides

Oligo	Purpose	Sequence
o877	PCR: M13 forward primer	CAGGAAACAGCTATGACC
o878	PCR: M13 reverse primer	TGTAAAACGACGGCCAGT
o4219	Gibson: Int' ORF in pBAD33.1	CTTTAAGAAGGAGATATACAAATGATTACTCATGGTTGTTATAC
o4220	Gibson: Int' ORF in pBAD33.1	CTCATCCGCCAAAACAGCCACAAGTCCTCTTCAGAAATG
o4386	PCR: anchored poly-dT primer	TGAGCTTTTGCTCCTCCGCATTTTTTTTTTTTTTTTTTTTTTTTTTTTTR
o4387	PCR: anchored poly-dT primer	TGAGCTTTTGCTCCTCCGCATTTTTTTTTTTTTTTTTTTTTTTTTTTTTCV
o4388	Gibson: Kozak primer acc. to Lee et al. for assembly with pDmS2569	CGTGCCGCTTCTGGCCCGGAATCCCCCGCCGCCACCATGG
o4389	Gibson: Kozak primer acc. to Lee et al. for assembly with pDmS2569	CGTGCCGCTTCTGGCCCGGAATCCCCCGCCGCCGCCATGG
o4390	Gibson: Kozak primer acc. to Lee et al. for assembly with pDmS2569	CGTGCCGCTTCTGGCCCGGAATCCDDDHDDHDAAAGATGH
o4391	Gibson: Kozak primer acc. to Lee et al. for assembly with pDmS2569	CGTGCCGCTTCTGGCCCGGAATCCDDDHDDHDKGKWATGH
o4430	Gibson: anchored poly-dT primer for assembly with pDmS2569	TCCGCACTAGTCATGGCCAATTTTTTTTTTTTTTTTTTTTTTTTTTTTTR
o4431	Gibson: anchored poly-dT primer for assembly with pDmS2569	TCCGCACTAGTCATGGCCAATTTTTTTTTTTTTTTTTTTTTTTTTTTTTCV

o4445	PCR: DpH-background oligo generation	TCACCCTTTCATTCATTCCC
o4446	PCR: DpH-background oligo generation	CTTCTCCTTTACTACTCAATTC
o4447	template: DpH-background oligo	TCACCCTTTCATTCATTCCCNNNNNDHNNNNGAATTGAGTAGTAAAGGAGAAG
o4449	template: 8NxC probe	TCACCCTTTCATTCATTCCCNNNNNCGNNNNGAATTGAGTAGTAAAGGAGAAG
o4451	PCR: 8NxC and 30NC probe generation	[TEG-biotin]TCACCCTTTCATTCATTCCC
o4452	PCR: 8NxC and 30NC probe generation	[TEG-biotin]CTTCTCCTTTACTACTCAATTC
o4453	QuikChange: XhoI in p2569	AGCGCAATTCCTCGAGCGCCCCTGGGTGC
o4454	QuikChange: XhoI in p2569	GCACCCAGGGGCGCTCGAGGAATTGCGCT
o4474	Custom sequencing primer for display vectors	GATGACTTCAGCACTTAATG
o4500	hybridization: Introducing linkers + TEV-site in pDmS2640	TCGAGTTGGGGGGGGAAGTGGTGGCGGATCAGAGAACCTGTACTTCCAGGGTGGGGGGGGAAGTGGTGGCGGATCAAA
o4501	hybridization: Introducing linkers + TEV-site in pDmS2640	AGCTTTTGATCCGCCACCACTTCCCCCCCCACCCTGGAAGTACAGGTTCTCTGATCCGCCACCACTTCCCCCCCCAAC
o4576	Gibson: HA-tag in pNaj2649	TTCCAGGGTGGGGGGGGAAGTGGTGGCGGATCAAAAGCTTATCCCTACGACGTACCAGATTATGCGGGGGGGGGA
o4577	Gibson: HA-tag in pNaj2649	CAAGTCCTCTTCAGAAATGAGCTTTTGCTCCTCCGCACTAGTTCCTCCGCCACCGGCGCGCCAT
o4578	template: HA-tag	AGCTTATCCCTACGACGTACCAGATTATGCGGGGGGGGGAAGTGGTGGCGGATGGCGCGCCGGTGGCGGAGGAA
o4579	template: HA-tag	CTAGTTCCTCCGCCACCGGCGCGCCATCCGCCACCACTTCCCCCCCCCGCATAATCTGGTACGTCGTAGGGATA
o4632	Gibson: hMBD2-domain in pNaJ2707	AGCTTATCCCTACGACGTACCAGATTATGCGGGGGGGGGAAGTGGTGGCGGATGGGATTGCCCCGCACTGCCTCCCGGATGGAAAAAG
o4633	Gibson: hMBD2-domain in pNaJ2707	CAAGTCCTCTTCAGAAATGAGCTTTTGCTCCTCCGCACTAGTGCATCATTTTACCGGTACGGAAATCGAACGAGGACAAGTCAACCGT
o4643	PCR: Kozak primer acc. to Lee et al.	ATCCCCCGCCGCCACCATGG
o4644	PCR: Kozak primer acc. to Lee et al.	ATCCCCCGCCGCCGCCATGG
o4645	PCR: Kozak primer acc. to Lee et al.	ATCCDDDHDDHDAAAGATGH

o4646	PCR: Kozak primer acc. to Lee et al.	ATCCDDDHDDHDKGKWATGH
o4681	PCR: VEGFA probe generation	TTTCCAAAGCCCATTCCCT
o4777	QuikChange: Removal of AfeI site from pDmS2899	CACCGCCGGACATCTTCGCTAGCGGAGTGT
o4778	QuikChange: Removal of AfeI site from pDmS2899	ACACTCCGCTAGCGAAGATGTCCGGCGGTG
o4814	Gibson: AfeI site in pNaJ2707	GGAAGTGGTGGCGGATCAAAAGCTTATCCCTACGACGTACCAGATTATGCG
o4815	Gibson: AfeI site in pNaJ2707	TGAGCTTTTGCTCCTCCGCATTCCTCCGCCACCGGCGA
o4852	PCR: VEGFA probe generation	[TEG-biotin]AGTGACCCCTGGCCT
o5206	PCR: incorporation of partial true-seq NGS adapter for pDmS2900 inserts	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGGTGGCGGAGC
o5207	PCR: incorporation of reverse NGS adapter for pDmS2900 inserts	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCACCGGCGAGC
o5208	PCR: amplification and incorporation of partial true-seq NGS adapter for pDmS2569 inserts	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTCTGGCCCGGA
o5209	PCR: incorporation of reverse NGS adapter for pDmS2569 inserts	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTTTTGCTCCTCCGCA
o5210	PCR: universal NGS amplification primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
o5211	PCR: Amplification and incorporation of i7-index 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
o5214	PCR: Amplification and incorporation of i7-index 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
o5216	PCR: Amplification and incorporation of i7-index 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
o5220	PCR: Amplification and incorporation of i7-index 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
o5298	template: 30NC probe	TCACCCTTTCATTCATTCCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAATTGAGTAGTAAAGGAGAAG
o5488	template: VEGFA promoter sequence probe	TTTCCAAAGCCCATTCCCTCTTTAGCCAGAGCCGGGGTGTGCAGACGGCAGTCACTAGGGGGCGCTCGGCCACCACAGGGAAGCTGGGTGAATGGAGCGAGCAGCGTCTTCGAGAGTGAG

	GACGTGTGTGTCTGTGTGGGTGAGTGAGTGTGTGCGTGTGGGGTTGAGGGCGTTGGAGCGGGGAGAAGGCCAGGGGTCACT

Plasmid vectors

Plasmid	Description	Source	Resistance
pAskInt110	Int', E-epitope	Wentzel et al.[242]	cm^r
pBAD33.1	araBAD promoter	Addgene	cm^r
pDmS2569	entry vector for Int' surface display under araBAD control, includes the E-epitope	this work	cm^r
pDmS2640	inserted XhoI site in pDmS2569	this work	cm^r
pDmS2701	pool of the full-length CDS thyroid library sorted 1x 8NmC(on), 8NC(off)	this work	cm^r
pDmS2705	pool of the full-length CDS thyroid library sorted 2x 8NmC(on), 8NC(off)	this work	cm^r
pDmS2711	pool of the full-length CDS thyroid library sorted 3x 8NmC(on), 8NC(off)	this work	cm^r
pDmS2716	pool of the full-length CDS thyroid library after induction	this work	cm^r
pDmS2899	inserted AfeI site to pNaJ2707	this work	cm^r
pDmS2900	entry vector for Int' surface display under araBAD control, includes AfeI site for blunt-end cloning	this work	cm^r
pDmS3491	pool of the fragmented prostate library sorted 1x for Myc-tag	this work	cm^r
pDmS3492	pool of the fragmented prostate library sorted 2x for Myc-tag	this work	cm^r
pDmS3493	pool of the fragmented prostate library sorted 3x for Myc-tag	this work	cm^r
pDmS3494	pool of the fragmented prostate library after induction	this work	cm^r
pDmS3495	pool of the fragmented prostate library after induction	this work	cm^r
pDmS3496	pool of the fragmented prostate library sorted 1x for 30NC probe	this work	cm^r

pDmS3497	pool of the fragmented prostate library sorted 2x for 30NC probe	this work	cm^r
pDmS3498	pool of the fragmented prostate library sorted 3x for 30NC probe	this work	cm^r
pNaJ2649	inserted (G)$_3$S(G)$_3$ linkers and TEV site in pDmS2640	this work	cm^r
pNaJ2707	inserted HA-tag in pNaJ2649	this work	cm^r
pNaJ2722	inserted hMBD2 MBD domain in pNaJ2707	this work	cm^r

12.2 Methods

12.2.1 General Methods

Bacterial Cell culture

If not mentioned otherwise liquid bacterial cultures were incubated in LB broth at 180 rpm at 37 °C overnight. Agar plates were incubated at 37 °C overnight. Plasmid maintenance of transformed bacteria was ensured by adding 34 µg mL^{-1} chloramphenicol.

Generation of chemically competent *E. coli* cells

Bacterial strains (strain GH371) were streaked to single clones on LB agar and cultivated overnight. 800 mL LB broth was inoculated from the fresh overnight culture at 200 rpm, 37 °C until the OD600 had reached 0.5 cm^{-1}. Then, the culture was cooled on ice for 15 min and the cells were harvested at 4 °C, 4,000 × g for 15 min. The pellet was washed once with 80 mL ice-cold, sterile 100 mM $MgCl_2$ and once with 50 mM $CaCl_2$. After incubating on ice for 30 minutes cells were centrifuged and resuspended in 4 mL sterile 50 mM $CaCl_2$ containing 15% (v/v) glycerol. Afterward, the cell suspension was aliquoted into fractions of 50 µL, snap-frozen in liquid nitrogen, and stored at –80 °C until further use.

Generation of electrocompetent *E. coli* cells

For the preparation of competent cells, *E. coli* strains Tuner™(DE3) and TOP10 were cultivated and harvested as mentioned above but resuspended once in 400 mL ice-cold sterile water and then washed again once in 200 mL and in 100 mL 10% (v/v) glycerol. Subsequently, the supernatant was discarded and the pellet was finally resuspended in 5 mL 10% (v/v) glycerol. Aliquots of 25 µL were snap-frozen and stored at –80 °C.

Transformation of bacterial cells via heat shock

25 µL of chemical-competent cells (GH371) were mixed with 1-5 µL (50-100 ng) of plasmid DNA or linear linearized plasmid DNA and incubated on ice for 15 minutes. The cells were placed in a heating block at 42 °C for 30 seconds and subsequently chilled on ice for 2 minutes. Afterward, the cells were rescued by the addition of 500 µL

of 37 °C pre-warmed S.O.C. medium for one hour at 37 °C and 750 rpm. Subsequently, 10 - 80 µL of the cell suspension were plated on an LB agar plate and incubated overnight at 37 °C.

Transformation of bacterial cells via electroporation

Typically, 18 ng library plasmid DNA and 25 µL electrocompetent cells were mixed and transferred into a pre-chilled 0.1 mm electroporation cuvette. After application of an electric charge of approx. 1800 V for 5 ms the cells were transferred to 500 µL 37 °C pre-warmed S.O.C.-medium and incubated for 1 hour at 37 °C and 750 rpm. Subsequently, a dilution series from the pooled S.O.C. cultures was performed and plated on LB agar plates. The majority of the cells were pelletized and resuspended in LB before they were plated on a LB agar plates as well.

Cryopreservation of bacterial cultures

50% sterile glycerol was added to a bacterial cell suspension to a total concentration of 15% and mixed carefully. Afterward, the cell suspension was aliquoted and kept at - 80 °C for long-term storage.

12.2.2 Biochemical Methods

Purification of PCR products and linearized plasmids

Linear dsDNA products of PCRs and restriction enzyme digests were purified with a PCR clean-up kit according to the manufacturers protocol and eluted in 2 steps in a total volume of 30 µL. 8NxC probes were purified enzymatically as described in the respective section.

Isolation of plasmid DNA

Plasmid DNA was isolated with the NucleoSpin Plasmid DNA isolation kit according to the manufacturer protocol and eluted in 50 µL elution buffer.

DNA concentration determination

The concentration of DNA solutions was determined via the absorption at 260 nm by NanoDrop after blank measurement with the respective buffer.

Agarose Gel Electrophoresis

2.5-5 µL of the DNA sample were mixed with 1x loading dye and loaded on an agarose gel (1% or 3% agarose in 0.5x TBE buffer) together with 5 µL of 1 kb Plus DNA Ladder or O'RangeRuler 20 bp Ladder The DNA was separated in 0.5x TBE buffer by applying a voltage of 100 V for 30-35 minutes. The gel was stained in 10 ng mL^{-1} ethidium bromide solution, then destained in ddH$_2$O and subsequently imaged under UV light.

12.2.3 Molecular Biology Methods

Restriction enzyme digest of plasmids

The digestion of plasmid DNA was executed according to the manufacturers protocol in a volume of 25 µL or 50 µL. The success of the reaction was verified by agarose gel electrophoresis and the linearized plasmid DNA was either purified by column purification or, concentrated and desalted by ethanol precipitation for library cloning,.

Dephosphorylation of the 5′-OH of dsDNA

The dephosphorylation of free 5′-OH end was executed in parallel to the restriction digest. Shrimp alkaline phosphatase was added to the plasmid restriction digest reaction in an amount of 2 U per µg of plasmid.

Phosphorylation of the 5′-OH of dsDNA

To perform T4 ligase-mediated ligation of dsDNA, the insert DNA was previously either phosphorylated by T4 polynucleotide kinase (NEB) or during dsDNA end-repair via the Quick Blunting™ Kit (NEB) according to the manufacturers guidelines.

Ethanol precipitation of DNA

For ethanol precipitation of DNA, 0.1 volumes of 3 M sodium acetate (pH = 5.1), 3 volumes ice cold (-20 °C) 100% ethanol (HPLC grade), and 1 volume of DNA solution were mixed and vortexed briefly. The DNA was precipitated at -80 °C for 1-2 h or overnight. After pelletizing the DNA at >16 × g at 4 °C for 30 min, the pellet was washed twice with 0.5 mL of -20 °C cold 75% EtOH and pelletized for 15 minutes. After removing residual ethanol, the pellet was air-dried for 30 s and resuspended in 10 µL MilliQ water.

Hybridization of complementary ssDNA oligonucleotides

3.8 µM of the complementary oligos were mixed in a 1:1 ratio in 1 x IDT annealing buffer. The mixture was incubated in a thermocycler at 90 °C for one minute and slowly cooled down to 20 °C with a rate of 0.1 °C min^{-1}.

Concentration of DNA samples

DNA samples were typically concentrated in a vacuum concentrator system using the "V-AQ" mode at 45 °C to prevent dsDNA melting.

Gibson assembly

Gibson assembly[362] of libraries was performed with the NEBuilder® HiFi DNA Assembly Master Mix according to the manufacturers protocol. The DNA amounts were adjusted to the specific needs of the library and are mentioned in the respective section. For assemblies that required only a low diversity, the linearized plasmid and the DNA insert of a total volume of 5 µL were mixed with 15 µL of a 1.33x self-made Gibson master mix and incubated at 50 °C for 1 h.

dsDNA ligation using T4 DNA ligase

The linearized plasmid and the DNA insert were mixed with T4 DNA ligase and ligated in a volume of 50 µL (see below) at 25 °C for 10 minutes or in the case of library ligation at 16 °C for 16 h. The enzyme was afterwards inactivated at 65 °C for 10 minutes.

Reagent	Volume / µL
Linearized plasmid	Varying
Insert DNA	Varying
10 x T4 ligase buffer	5
ATP (10 mM)	0.5 - 0.8
MilliQ water	Up to 45
T4 DNA ligase	5

Site-directed mutagenesis

For site-directed mutagenesis, the polymerases Pfu or KOD were used. The template plasmid was amplified as a whole by PCR (protocol see below) with primers baring one or more mismatched nucleotides thereby introducing the desired point mutations. After DpnI digest of the parent plasmid, the linear product was heat-shock transformed into GH371 cells, and the whole culture was plated on an agar plate after recovery.

	Pfu reaction	KOD reaction
Reagent	Volume	Volume
buffer	10x Pfu buffer (5 µL)	10x KOD buffer (2.5 µL)
dNTP mix	1.5 µL of 10 mM	2.5 µL of 2 mM
Fw primer (10 µM)	2.5 µL	0.75 µL
Rv primer (10 µM)	2.5 µL	0.75 µL
Plasmid template	1 µL	0.5 µL
$MgSO_4$ (25 mM)	-	1.5 µL
MilliQ	Up to 49 µL	Up to 24.5 µl
DNA polymerase	Pfu polymerase (1 µL)	KOD polymerase (0.5 µL)

Pfu protocol		
T / °C	t / s	Cycles
95	30	1
95	30	25
variable	60	
70	900	

KOD protocol		
T / °C	t / s	Cycles
95	120	1
95	20	25
variable	10	
70	240	

Creation of the entry vector for full-length CDS display (pDmS2569)

Plasmid pASKInt110[363], containing the truncated intimin sequence (Int′), was kindly provided by Prof. Dr. Harald Kolmar. The Int′ and E-epitope ORF was amplified with primers o4219 and o4220 using Phusion-HF polymerase and cloned into the NdeI- and HindIII-HF-digested plasmid pBAD33.1 (Addgene, Watertown, Massachusetts, USA) via Gibson assembly resulting in the entry vector for full-length CDS libraries pDmS2569. PCR conditions see below.

Reagent	**Volume / µL**
5x Phusion HF buffer	10
dNTP mix (10 mM)	1
o4219 (10 µM)	2.5
o4220 (10 µM)	2.5
pASKInt110 (p2510)	1 µL (19.9 ng µL^{-1})
MilliQ	32.5
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
98	30	1
98	10	25
56	30	
72	150	

Creation of the entry vector for fragmented cDNA surface display

For the display of protein fragments, an additional XhoI recognition site was introduced by site-directed mutagenesis of the entry vector pDmS2569 with primers o4453 and o4454 using self-made Pfu polymerase. Chemically competent *E. coli* GH371 cells were transformed with the PCR product and plated on an LB agar plate. After isolating the plasmid from liquid overnight culture, the success of the mutagenesis was verified by Sanger sequencing and resulted in plasmid pDmS2640.

The introduction of $(G)_3S(G)_3$ linkers downstream of the TEV recognition site was facilitated by ligation with T4 ligase. Oligos o4500 and o4501 were hybridized and the product was 5′-OH-phosphorylated. The hybridization product was ligated with the XhoI- and HindIII-HF-digested plasmid pDmS2640 in a ratio of 1:5 resulting in the plasmid pNaJ2649.

The implementation of the HA-tag and linker downstream of the TEV recognition site was facilitated by Gibson assembly. The purified PCR product of o4576 and o4577 and the template oligos o4578 and o4579 using the Phusion HF polymerase with the HindIII-HF- and SpeI-HF-digested plasmid pNaJ2649 resulted in the plasmid pNaJ2707. The PCR protocol is shown below.

Reagent	**Volume / µL**
5x Phusion HF buffer	10
dNTP mix (10 mM)	1
o4576 (10 µM)	2.5
o4577 (10 µM)	2.5
o4578 (10 µM)	0.1
o4579 (10 µM)	0.1
MilliQ	33.3
Phusion HF Polymerase	0.5

T / °C	**t / s**	**Cycles**
94	15	
72	30	30
72	30	

The implementation of the AfeI restriction site for blunt-end ligation was achieved by removing a part of the plasmid by digestion with HindIII-HF and SpeI-HF and replacing it with the same sequence including an AfeI restriction site. The insert sequence was amplified from the ssDNA template o4816 with Phusion polymerase (NEB) and primers o4814 and o4815 thereby introducing Gibson overhangs.

Reagent	Volume / µL
5x Phusion HF buffer	10
dNTP mix (10 mM)	1
o4814 (10 µM)	2.5
o4815 (10 µM)	2.5
o4816 (10 µM)	0.1 µL
MilliQ	33.4
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
98	30	1
98	10	
72	30	25
72	30	

The purified digested plasmid pNaJ2707 was assembled with the purified PCR product using the self-made Gibson mix which resulted in the plasmid pDmS2899. Removal of an additional AfeI site present in the backbone by site-directed mutagenesis with o4777 and o4778 and KOD-polymerase resulted in the entry vector for fragmented libraries pDmS2900.

Sanger sequencing of DNA samples

500 ng of plasmid DNA were mixed with 2 µM of custom sequencing primers in a volume of 12 µL and submitted to Sanger sequencing at Microsynth SeqLab. Alternatively, standard sequencing primers provided by Microsynth SeqLab were used. The sequence alignment and analysis of the results were facilitated with SnapGene[364].

12.2.4 Creation of cDNA encoded protein libraries for Int′-mediated surface display

Creation of the full-length human thyroid CDS display library

The commercially available human thyroid cDNA library HD-503 (Zyagen) was amplified with an anti-Kozak primer pool designed by Lee et al.[281] (o4388-o4391) and anchored poly-dT[365] primers (o4430 & o4431) thereby introducing overhangs for Gibson assembly. 5 µL of the PCR product were loaded on a 1% agarose gel and the remaining product was concentrated and desalted via ethanol precipitation. The compositions of the primer mixes as well as the PCR protocol are shown below.

Kozak primer mix	
Primer	**Ratio**
o4388 (10 µM)	12
o4389 (10 µM)	8
o4390 (10 µM)	15
o4391 (10 µM)	15

Anchored poly-dT primer mix	
Primer	**Ratio**
o4430 (10 µM)	2
o4431 (10 µM)	1

Reagent	Volume / µL
5x Phusion GC buffer	10
dNTP mix (10 mM)	1
Kozak primer mix (10 µM)	2.5
Anchored-poly-dT-primer mix (10 µM)	2.5
Thyroid cDNA	1
MilliQ	32.5
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
94	15	
50	30	30
72	360	

The display entry vector pDmS2569 was digested with HindIII-HF and SpeI-HF, dephosphorylated, and afterward concentrated via ethanol precipitation. 100 ng of digested pDmS2569 were assembled with 219 ng of thyroid cDNA amplicons using NEBuilder HiFi mix according to the manufacturers protocol in a 10 µL reaction volume. Additionally, a control reaction without cDNA amplicons was prepared. The library assembly, as well as the control, were each resuspended in a total volume of 10 µL MilliQ water after ethanol precipitation. 5 x 2 µL of the precipitated library assembly and 2 x 2 µL of the control were transformed in 25 µL electrocompetent *Escherichia coli* TOP10 cells (OD_{600} = 25) and recovered in SOC medium for 1 h, 37 °C, 750 rpm. All cells of the library assembly as well as a dilution series of assembly and the control (1:1000 and 1:10,000) were plated on LB agar plates and incubated at 37 °C overnight. The plasmid pool diversity was determined by counting the colonies of the dilution series plates and considering the dilution factor and the respective number of plasmids without inserts as determined by the control. Cells from library assemblies were washed off the plates with LB medium and the plasmid pool was isolated using plasmid purification kit, which gave the plasmid library pDmS2684. 8 x 1 µL of the plasmid pool was then retransformed in 35 µL electrocompetent *E. coli* Tuner^{TM} (DE3) cells (OD_{600} = 25) and treated as previously described. The diversity of this final screening library was determined by a dilution series (1:1000 and 1:10,000, 1:100,000) under consideration of the initial plasmid pool size. Cells were washed off the plates, prepared for cryopreservation, aliquoted, and stored at -80 °C. The success of the assembly was validated by Sanger sequencing.

Creation of the fragmented human prostate cDNA display library

For the fragmented prostate library, the commercially available plasmid prostate cDNA library 10108-A (BioCat) was purchased as pExpress plasmid pool in a cryopreserved bacterial strain. The plasmids were isolated from the overnight culture via a plasmid purification kit and the cDNA inserts were initially amplified by using M13 primers flanking the insertion site of the pExpress plasmid. Afterwards, the template plasmid was removed with DpnI and the amplicons were purified using a PCR purification kit. The PCR conditions are shown below.

Reagent	Volume / µL
5x Phusion GC buffer	10
dNTP mix (10 mM)	1
o877 (10 µM)	2.5
o878 (10 µM)	2.5
Prostate cDNA plasmid library	30 ng
MilliQ	Up to 49.5
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
98	30	1
98	30	35
55	30	
72	150	
72	300	1

The purified amplicons served as the template for the amplification with the anti-Kozak primer pool (o4643-o4646) and poly-dT primers (o4386 and o4387) without Gibson overhangs. Primer mix compositions and PCR conditions are shown below.

Kozak primer mix	
Primer	Ratio
o4643 (10 µM)	12
o4644 (10 µM)	8
o4645 (10 µM)	15
o4646 (10 µM)	15

Anchored poly-dT primer mix	
Primer	Ratio
o4386 (10 µM)	2
o4387 (10 µM)	1

Reagent	Volume / µL
5x Phusion GC buffer	10
dNTP mix (10 mM)	1
Kozak primer mix (10 µM)	2.5
Anchored-poly-dT-primer mix (10 µM)	2.5
Template amplicons	100 ng

MilliQ	Up to 49.5 µL
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
98	30	1
98	30	35
58	30	
72	150	
72	300	1

The resulting PCR product was purified and sheared by sonication to an average size of 500 bp in a volume of 100 µL and a concentration of 9 ng µL^{-1} for 24 cycles (5 s on, 90 s off). The fragments were end-repaired and phosphorylated using the Quick Blunting^{TM} kit according to the manufacturers protocol. 520 ng of AfeI-digested and dephosphorylated pDmS2900 were ligated with 520 ng of end-repaired and phosphorylated fragments using T4 DNA ligase. Transformation procedures were identical to the ones of full-length libraries and resulted in the plasmid library pNaK3454. The average fragment size of library inserts was determined via Sanger sequencing of 14 single clones.

Creation of an Int′-MBD2$_{151\text{-}214}$ display vector

The MBD of hMBD2 (residues 151-214) was amplified with Phusion HF polymerase and primers o4632 and o4633 from plasmid pBeB1567 created by Buchmuller et al.[109].

Reagent	Volume / µL
5x Phusion HF buffer	10
dNTP mix (10 mM)	1
o4632 (10 µM)	2.5
o4633 (10 µM)	2.5
pBeB1567	50.2 ng
MilliQ	Up to 49.5
Phusion HF Polymerase	0.5

T / °C	t / s	Cycles
94	30	1
98	15	30
65	30	
72	30	

After DpnI digestion of the template plasmid and purification, the PCR products were assembled with AscI- and SpeI-HF-digested pNaJ2707 in a mass ratio of 2:1 in the self-made Gibson mix. 25 µL of electrocompetent *E. coli* TOP10 cells were transformed with 2 µL of the crude assembly reaction and spread on an LB agar plate. Plasmid were isolated from overnight cultures of single clones and the successful incorporation of the hMBD2 sequence was verified by Sanger sequencing.

12.2.5 FACS related methods

The general concept as well as protocols for culturing bacterial cells, composition of the staining mixes, and FACS settings were based on the work of Buchmuller et al.[366] and optimized for the needs of the surface display of cDNA-encoded protein libraries.

Creation of 8NxC-probes, 8NDpH-background oligos, and VEGFA promoter probe

8NxC probes, as well as the 8NDpH-background oligos, were generated via PCR with Phusion polymerase using biotinylated primers o4451/52 for probes or unbiotinylated primers o4445/6 for 8NDpH-background. They bind the orthogonal flanking sequences of the template oligos o4449 (probe) or o4447 (DpH background). The dNTP mix (NEB) or 5mdCTP containing dNTP mix were used to incorporate 5mC or the unmodified cytosine. The orthogonal flanking sequences were modified from Subramanian et al.[367] to avoid the incorporation of cytosine during the PCR. Afterward, the remaining primers were removed using 5 U Exonuclease I and dephosphorylated with 1 U rSAP for 15 min at 37 °C. Heat-induced inactivation of the enzymes was omitted to avoid melting and subsequent misalignment of the randomized part of the dsDNA. Afterward, the products were desalted, concentrated via ethanol precipitation, and resuspended in MilliQ water. The concentration was determined via nanodrop and the stocks were adjusted with MilliQ water to 142 nM (probes) and 5 µM (DpH background). Afterward the success was controlled via agarose gel electrophoresis. PCR conditions and cycles are shown below.

	8NxC probe	8NDpH-background
5 x Phusion HF buffer	10 µL	10 µL
dNTP mix (10 mM, +/- 5mdCTP)	1 µL dNTPs	1 µL dNTPs
Fw primer (10 µM)	2.5 µL o4451	2.5 µL o4445
Rv primer (10 µM)	2.5 µL o4452	2.5 µL o4446
Fw template (10 µM)	0.05 µL o4449	0.05 µL o4447
Rv template (10 µM)	0.05 µL o4450	0.05 µL o4448
MilliQ	32.4 µL	32.4 µL
Phusion HF Polymerase	0.5 µL	0.5 µL

T / °C	t / s	Cycles
98 °C	30 s	1
98 °C	10 s	35 (DpH background), 26 (probes)
57 °C	5 s	
72 °C	15 s	
4 °C	hold	

The methylated VEGFA promoter probe was generated via PCR with Phusion polymerase using the biotinylated primers o4681 and o4852 and the template oligo o5488. The 5mdCTP containing dNTP mix was used to incorporate 5mC during the PCR. PCR conditions and cycles are shown below.

Reagent	Volume / µL
5 x Phusion HF buffer	50
dNTP mix (10 mM, +/- 5mdCTP)	5
o4681 (10 µM)	12.5
o4852 (10 µM)	12.5
o5488 (10 µM)	0.5 µL template
MilliQ	167
Phusion HF Polymerase	2.5

T / °C	t / s	Cycles
98 °C	30 s	1
98 °C	10 s	28
62 °C	15 s	
72 °C	15 s	
4 °C	hold	

Creation of 30mer random dsDNA probe (30NC)

The 30mer random probe (30NC) was created via PCR with Phusion polymerase using biotinylated primers o4451/52 binding to the orthogonal flanking sequences (of the template oligo o5298. The product was subsequently purified via a PCR purification kit and the success of the reaction was controlled via agarose gel electrophoresis. The concentration was determined via Nanodrop and adjusted to 246.9 nM with MilliQ water. PCR conditions are shown below.

Reagent	Volume / µL
5 x Phusion HF buffer	10
dNTP mix (10 mM)	1
o4451 (10 µM)	2.5
o4452 (10 µM)	2.5
o5298 (10 µM)	0.0125
MilliQ	33.5
Phusion Pol.	0.5

T / °C	t / s	Cycles
98 °C	30 s	1
98 °C	10 s	30
57 °C	5 s	
72 °C	15 s	

FACS laser and detector settings

Sorting and analysis of bacterial cells was executed with a SH800S Cell Sorter with a 100 µm microfluidic sorting chip. The instrument was equipped with 488 nm, 405 nm (not used), 638 nm and 561 nm lasers, and the "Filter Pattern 2" was used to disentangle the collinear beams. FITC fluorescence was detected in detector FL2 at 65% PMT, phycoerythrin (PE) in FL3 at 45% PMT, and allophycocyanin (APC) in FL4 at 45% PMT. Regular cells from a control population were gated at a forward scatter intensity of 16 a.u. and 40% PMT of the backward scatter.

Bacterial surface display of protein libraries and single proteins

For screening purposes, 3 mL LB broth was inoculated with the respective cryopreserved bacterial library to an $OD_{600} = 0.05$ and incubated at 37 °C, 180 rpm. Expression of the Int'-fusion proteins was induced after at OD = 0.4-0.6 by the addition of 0.05% arabinose. Cells were incubated at 30 °C, 150 rpm and a multiple of OD of 0.4 was harvested. Cells were washed with 1 mL of ice-cold PBS, resuspended in a respective multiple of 200 µL PBS, and chilled on ice for 1 h. The same protocol was used when displaying the Myc-tag or $hMBD2_{151\text{-}214}$ on bacterial cells harboring the plasmids pDmS2569, pNaJ2707, and pNaJ2722.

FACS screening of the fragmented human prostate cDNA library for Myc-tag display

One volume (V) of the suspension of *E. coli* Tuner™(DE3) cells displaying the fragmented human prostate library (first selection: 200 µL, second: 50 µL, third: 50 µL) was stained with an APC-labeled anti-Myc antibody (223896, Abcam) in a final dilution of 1:500 and incubated for 20 min at 20 °C, 700 rpm. Afterward, the cells were pelletized at 8000 x g, 2 min, 4 °C, and washed with 6V ice-cold 1x PBS. The cells were pelletized again and resuspended in 6.5 V 1x PBS. Before loading, the samples were carefully resuspended by pipetting up and down. Myc-tag positive cells were then sorted in purity mode at 50-300 events per second in a benchtop FACS system and collected in 15 mL tubes with 4 mL LB broth. The sorted cells were incubated at 37 °C, 180 rpm overnight. The overnight culture was used for inoculation of the culture for the next sorting round. Additionally, 800 µL were used for the preparation of a cryopreserved culture, and the plasmid pool was isolated from the remaining 3 mL of the culture via a plasmid purification kit. Furthermore, the plasmid pool of 3 mL culture of the initial culture was isolated after protein expression and harvesting. All sorting steps were executed in the same manner.

In the same way, 20 µL of the suspension of *E. coli* Tuner™(DE3) cells harboring the empty entry vector p2900 were stained, and subsequently analyzed by flow cytometry. These cells only displayed the HA-tag, linker, and TEV site and served as a negative control.

Single- and two-color staining of cells for FACS sorting or analysis

All dsDNA staining mixes consisted of components A and B.

Component A comprised DpH-background oligos, 3 x EMSA-T buffer, and the respective probe. Component B contained Streptavidin-phycoerythrin (SAv-PE) or SAv-FITC, TCEP, and ultrapure BSA. Both components were mixed in a 1:1 ratio and chilled on ice for 1 h to obtain the single staining mixes.

To obtain the double staining mix, the red (PE) and green (FITC) single staining mixes were mixed in a 1:1 ratio under the addition of 3.6 µM biotin (dissolved in DMSO) to saturate free binding sites of SAv.

Typically, cells and staining mixes were mixed in a ratio of 2:1 for flow cytometry analysis and FACS. The concentrations of all components of single-staining mixes as well as in the final staining solution are given below.

Staining	probe	Reagent	Red staining mix	Green staining mix	Final mix with cells
Single color	30NC	probe	120 nM	-	**40 nM**
		SAv-PE	776 nM		**259 nM**
		TCEP	4 mM		**1.33 mM**
		BSA	0.52 µg µL^{-1}		**0.173 µg µL^{-1}**
Two colors	8NxC	probe	39.76 nM	39.76 nM	**Each: 6.02 nM**
		SAv-PE	576 nM	-	**87.27 nM**
		SAv-FITC	-	576 nM	**87.27 nM**
		EMSA-T	3x	3x	**0.91x**
		TCEP	2.95 mM	2.95 mM	**0.89 mM**
		BSA	0.38 µg µL^{-1}	0.38 µg µL^{-11}	**0.115 µg µL^{-1}**
		8NDpH oligos	300 nM	300 nM	**90.9 nM**
		Biotin	-	-	**3.6 µM**

FACS screening of the fragmented human prostate cDNA library for general dsDNA binders

One volume (V) of the suspension of *E. coli* Tuner™(DE3) cells displaying the fragmented human prostate library (first round of screening: 150 µL, second: 40 µL, third: 20 µL) was stained with 0.5V of the 30NC-red single staining mix and incubated for 20 min at 20 °C, 700 rpm. All following procedures were identical to the Myc-tag screening and cells showing PE signal were sorted accordingly.

20 µL of the suspension of *E. coli* Tuner™(DE3) cells harboring the empty entry vector p2900 were stained in the same way and subsequently analyzed by flow cytometry a. These cells only displayed the HA-tag, linker, and TEV site and served as a negative control

FACS analysis of the intimin-mediated Myc-tag and hMBD2$_{151-214}$ display and surface cleavage by TEV

If not mentioned otherwise, all culturing, expression, and staining conditions were kept identical to the respective staining protocols for library sorting.

20 µL of the cell suspension of *E. coli* Tuner™(DE3) cells that harbored the plasmid pDmS2569 and displayed the Myc-tag were treated and stained with the anti-Myc (ab72580, Abcam) antibody to verify the display success via flow cytometry. Additionally, cells whose expression was not induced were stained accordingly and served as negative control.

1 µL TEV protease (1.2 mg mL^{-1}, self-made) was added to 49 µL of the cell suspension of *E. coli* Tuner™(DE3) cells that harbored the plasmid pNaJ2707 and displayed the HA-tag, linker, TEV recognition sequence, and the Myc-tag and subsequently incubated for 1 h at 22 °C, 750 rpm. After chilling on ice for 1 h, 20 µL of the cells were stained with the anti-Myc-antibody (ab72580, Abcam) and served as the negative control. 49 µL of the same cell suspension to which no TEV protease was added were treated and stained in the same way. The successful display of the Myc-tag and its removal by TEV cleavage were analyzed by flow cytometry.

49 µL of the cell suspension of *E. coli* Tuner™(DE3) cells that harbored the plasmid pNaJ2722 and displayed the hMBD2$_{151-214}$ were treated with TEV in the same way as

mentioned before. After chilling on ice for 1 h, 20 µL of the cells were stained with 5 nM of SAv-PE-labeled 8NmC probe and served as the negative control. 49 µL of the same cell suspension to which no TEV protease was added were treated and stained identically. The successful display of the MBD and its removal from the surface by TEV cleavage were analyzed by flow cytometry.

FACS screening of a full-length thyroid cDNA library for readers of 5mC

One volume (V) of the suspension of *E. coli* Tuner™(DE3) cells displaying the full-length human thyroid library (first round of screening: 80 µL, second: 30 µL, third: 10 µL) was stained with 0.5V of the 8NmC/8NC-double staining mix and incubated for 20 min at 20 °C, 700 rpm. All following procedures were identical to the Myc-tag screening and screening for general dsDNA binders.

Additionally, 47 regular-shaped single cells were sorted on FACS-compatible LB agar plates in single-cell mode during the last sorting round for single clone analysis via Sanger sequencing.

20 µL of the library cell suspension of the library after the second selection round were further incubated with 10 µL of 1 x PBS and analyzed via flow cytometry. These cells served as a control for the fluorescence background signal not based on probe binding.

Viability determination of E. coli Tuner™(DE3) cells displaying cDNA libraries

Cultures of *E. coli* Tuner™(DE3) cells harboring the fragmented prostate cDNA or the full-length thyroid CDS library were treated according to the single-color FACS screening protocol. Before induction, the cultures were split and only one subculture was induced while the other served as control. Cells harboring the prostate library were incubated with a mock green staining mix lacking the DNA probe while the cells harboring the thyroid library were incubated with a red-staining mix containing 9.3 nM VEGFA promoter sequence probe. Regular-shaped single cells were sorted on FACS-compatible LB agar plates in single-cell mode and the plates were incubated at 37 °C overnight. The next day the colony-forming units (cfu) were counted and compared to the number of sorted cells to determine the viable cell population.

12.2.6 Next generation sequencing (NGS) sample preparation and data analysis

Illumina sequencing of library pool inserts

The plasmid pools of the cultures after protein expression (parent libraries) and of the overnight cultures (sorted libraries) were isolated from 2 mL culture volume using a plasmid isolation kit. 40 ng of the plasmid pools from the ORF screening (p3491-p3494), the screening for general DNA binders (p3495-p3498), and for 5mC-selective readers (p2716, p2701, p2705, p2711), were used for an initial PCR with primers flanking the insertion site of the cDNA and introducing forward partial true-seq and reverse NGS adapters. For amplification of the pools p3491-p3494 and p3495-p3498, the reverse primer was shifted 66 bp downstream to ensure a minimal final insert size of 100 bp to avoid running into the flow cell during sequencing.

Reagent	Volume / µL
10 x Q5 buffer	10
dNTP mix (10 mM)	1
Template	40 or 150 ng
o5206/8 (10 uM)	0.2
o5207/9 (10 uM)	0.2
MilliQ water	Up to 49.5
Q5 Polymerase	0.5

T / °C	t / s	Cycles
98 °C	60 s	1
98 °C	30 s	4
66°C/ 64°C	20 s	
72 °C	45 s	

After enzymatic removal of the primers with Exonuclease I for an extended time of 1 h at 37 °C, the PCR product was used as a template for the second PCR thereby introducing Illumina i7 index sequences for multiplex sequencing.

Reagent	Volume / µL
10 x Thermo Polymerase buffer	20
dNTP mix (10 mM)	4
Template	50
o5210 (10 uM)	1
i7 BC primer (10 uM)	1
MilliQ water	Up to 200
Taq polymerase	1

T / °C	t / s	Cycles
95 °C	60 s	1
95 °C	30 s	30
72 °C	120 s	
4 °C	inf.	1.

5 µL of the PCR products were analyzed on a 1% agarose gel and the rest was purified via a PCR purification kit. No-template controls as well as controls for the respective empty backbones were executed and analysed identically. Afterwards, the mass concentration was measured via Nanodrop and the molar concentration was calculated under consideration of the average PCR product size determined by agarose gel electrophoresis. Subsequently, the amplicons of the enrichment sets were pooled under consideration of the estimated library diversity and sent for 150 bp paired-end multiplex Illumina sequencing at Novogene (Cambridge, UK) on a NovaSeq platform (San Diego, USA) with a data outcome of 1-3 GB.

Sub library name	Target	Display library	Selection step	Primer (1st PCR)	Primer (2nd PCR)	i7 index	i7 sequence	Molar fraction in sequencing pool (%)
p3494	Myc tag	Fragmented human prostate cDNA	Parent	o5206 & o5230	o5210 & o5211	1	`ATCACG`	89.49
p3491			1st		o5210 & o5214	4	`TGACCA`	3.72
p3492			2nd		o5210 & o5216	6	`GCCAAT`	3.34
p3493			3rd		o5210 & o5220	12	`CTTGTA`	3.45
p3495	30mer dsDNA	Fragmented human prostate cDNA	Parent	o5206 & o5230	o5210 & o5211	1	`ATCACG`	83.42
p3496			1st		o5210 & o5214	4	`TGACCA`	11.16
p3497			2nd		o5210 & o5216	6	`GCCAAT`	2.76
p3498			3rd		o5210 & o5220	12	`CTTGTA`	2.66
p2716	5mCpG	Full-length CDS ORF human thyroid cDNA	Parent	o5208 & o5209	o5210 & o5211	1	`ATCACG`	86.42
p2701			1st		o5210 & o5214	4	`TGACCA`	5.10
p2705			2nd		o5210 & o5216	6	`GCCAAT`	4.64
p2711			3rd		o5210 & o5220	12	`CTTGTA`	3.84

Trimming of Illumina data of library pool inserts

Demultiplexing of pooled samples was executed by Novogene (Cambridge, UK). Trimming of Illumina adapters, remaining bases of the flanking plasmid backbone sequences, and poly-G overhangs, as well as base overlap correction, and read-quality control, was done with fastp[368] without applying length filtering. The resulting data was analyzed via transcriptome alignment of the DNA inserts or proteome alignment of the resulting in-silico translated peptides. The number of quality control-passed reads can be seen above.

Transcriptome mapping of Illumina data of the fragmented prostate library

For transcriptome alignment, the QC-passed R1 reads were mapped against the human transcriptome GRCh38[369] (Genome Reference Consortium) with bowtie2 [370] and converted into bam file format with SAMtools[371]. The mapped reads were processed and plotted in *R* using dplyr, tidyverse, and ggplot packages as well as packages from the *R* Bioconductor suite. All transcripts, except for the top 10 most abundant, were summed as "rest" and plotted together with the most abundant ones.

Proteome mapping based on Illumina data of library pool inserts

Data processing was executed via *R* in RStudio generally using the packages tidyverse[372], dplyr[373], ggplot2[374], svglite[375], data.table[376] and ggbreak[377] as well as additional packages from the *R* Bioconductor suite. For proteome alignment, the QC-passed forward reads (R1) were imported in *R* for further processing via the microseq package[378]. The first reading frame of the DNA sequences was translated via the Biostrings package[379] to the corresponding amino acids sequence while solving ambiguous codons as undefined ("X"). The ORFs were obtained by C-terminal trimming of the amino acid sequences until a stop codon was reached. Additional trimming for the N-terminally fused linker sequence (SGGG) and a typical motif resulting from Kozak sequences (SPAA) were removed via stringr package[380] before the sequences were exported via the microseq package as fasta-file for subsequent proteome alignment via BLAST+ suite[303]. The blastp-short command of BLAST+ was used for the peptide sequence alignment to the Swissprot human proteome database release version 2023_2[294] while only considering the alignment with the best score.

Analysis of the enrichment of mapped peptide sequences

Analysis of the mapped data was executed in *R* by using dplyr, tidyverse, ggplot and ggbreak packages as well as packages from *R* Bioconductor suite. At first, the data was filtered for the alignment quality (e-value ≤ 0.01), minimal alignment length (min. 15 amino acids), and maximal two gap openings. Query and alignment lengths were calculated via their respective start and end points and were used to calculate the size of the displayed protein fragments of each sub-library. Peptide counts of the same proteins were summed and the mean values and standard deviations for start and end parameters, alignment lengths, e-values, and percent identities were calculated. In this way, the peptide composition of the sub-libraries was determined

Enrichment factor determination between two sub-libraries was achieved by normalizing the peptide counts of each protein on the trimmed and quality control-passed forward read number of the respective sub-library and calculating the ratios between these normalized counts for each protein between the sub-libraries. Proteins with 0 counts were set to 0.9 to include them in the enrichment factor calculation. Monitoring of the consecutive enrichment throughout the selection process was achieved by multiplying the enrichment value with the one of previous sorting steps. Only proteins with an absolute count ≥ 5 in the sorted library were considered for the final analysis.

Analysis of the composition of peptides and proteins in the reads of the sub libraries was achieved by determining the amount of unique (based on the alignment parameters), mapped, and filtered peptides and their respective proteins.

Nanopore sequencing and transcriptome mapping of cDNA

For analysis of the inserts of full-length screening library pDmS2684, the plasmid pool was linearized with SalI-HF according to the manufacturers manual. The prostate cDNA library 10108-A was amplified with Kozak and M13 primers. After PCR clean-up, the DNA was prepared for nanopore sequencing by the SQK-LSK110 kit according to the manual for amplicon ligation for Flongle flow cells or Spot On Flow Cells cells. 90 ng of the adapter-ligated library were loaded and sequenced on a Spot On Flow Cell (Kozak-M13 and M13-only amplicons) or Flongle Flow cell (pDmS2684 DNA). Base-calling was executed by the MinKnow software. The adapter and plasmid backbone sequences were removed from data with porechop[381.] The trimmed sequences were mapped against the human transcriptome GrCh37[382] mimicking the blastn settings in minimap2[285]. Only primary alignments were kept. Afterward, the mapped data was converted to bam-file using SAMtools and imported in *R*. Notably, the data was not filtered for mapq values, since accurate mapping positions of the transcripts were not required. The data was further analyzed with packages from the *R* Bioconductor suite.

Nanopore sequencing resulted in the following read output of each library:

Library	Raw reads
Thyroid display library	155,524
Kozak-M13 amplicons	1,110,000
M13-only amplicons	589,423

Data analysis of transcriptome-mapped Nanopore reads of Kozak-primer impact analysis

The mapped reads were either summed per RNA type or genomic locus, and their fractions within the libraries were determined in *R* using dplyr, tidyverse, ggplot and ggbreak packages as well as packages from the *R* Bioconductor suite.

The fraction of the transcript number of each genomic locus was compared and fold-change values were determined. Genomic loci that were found to be reduced or five-fold enriched in the Kozak-M13 amplicons were plotted against the fold-change values.

Data analysis of transcriptome-mapped Nanopore reads of the thyroid cDNA display library

The mapped reads were processed and plotted in *R* using dplyr, tidyverse, and ggplot packages as well as packages from the *R* Bioconductor suite. All transcripts, except for the top 10 most abundant, were summed as “rest” and plotted together with the most abundant ones.

12.2.7 Source code

Commands for Nanopore read trimming and transcriptome mapping

The sequence of the backbone for trimming of the thyroid cDNA reads was included in the porechop adapter sequence file.

```
porechop -t 12 --input ./reads --output ./trimmed_reads.fastq --
extra_end_trim 17 --verbosity 1

wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_37/gencode.
v37.transcripts.fa.gz
gunzip gencode.v37.transcripts.fa.gz

minimap2 -k 11 -A 2 -B 3 -O 5 -E2 -d ref.mmi gencode.v37.transcripts.fa.gz
(indexing parameters like blastn)
minimap2 -a --secondary=no ref.mmi ./trimmed_reads.fastq > ./
mapped_reads.sam

samtools view -Sb mapped_reads.sam > mapped_reads.sam.tmp
samtools sort mapped_reads.sam.tmp > mapped_reads.bam
samtools index mapped_reads.bam
```

Source code for Figure 24 b

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(ggbreak)

#read-in mapped data and sum by RNA type and calculate RNA type fraction
type_Kozak <- read_csv(paste0("./file_mapped_reads_Kozak-M13.csv")) %>%
 group_by(type) %>%
 summarise(mapped = sum(mapped)) %>%
 mutate(fraction=mapped/sum(mapped))
type_M13 <- read_csv(paste0("./file_mapped_reads_M13.csv")) %>%
 group_by(type) %>%
 summarise(mapped = sum(mapped)) %>%
 drop_na() %>%
 mutate(fraction=mapped/sum(mapped))

#join by RNA type
comp_join <- full_join(type_M13, type_Kozak, by = c("type"))
#calculate the difference of each fraction
comp_join <- comp_join %>% mutate(diff = fraction.y - fraction.x)
#calculate relative change in fraction in percent
comp_join <- comp_join %>% mutate(factor_fraction =
(fraction.y/fraction.x)*100)

#for clarity types of low abundance (fraction < 1e-2) are summed as "rest"
comp_join_rest <- comp_join %>%
 subset(fraction.x < 1e-2 & fraction.y < 1e-2) %>%
 summarise(fraction.x=sum(fraction.x), fraction.y=sum(fraction.y)) %>%
 mutate(type="Rest")

comp_join <- comp_join %>%
 subset(fraction.x > 1e-2 & fraction.y > 1e-2) %>%
 select(c('type','fraction.x', 'fraction.y'))

comp_incl_rest <- rbind(comp_join, comp_join_rest) %>% as.data.frame()

#plot
comp_incl_rest %>%
  ggplot(aes(x=reorder(type, -fraction.y))) +
  geom_bar(aes(y=(fraction.y*100)), stat="identity", position="identity",
  alpha=.8, fill='azure3', color='black')+
  geom_bar(aes(y=(fraction.x*100)), stat="identity", position ="identity",
alpha=.8,
  fill='darkslategray', color='black')+
  scale_y_break(c(15, 80))+
  ylim(0,100)+
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    legend.title=element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(colour = "black"),
    text = element_text(size = 8),
    axis.text = element_text(color="black"),
    axis.ticks = element_line(color = "black"))
```

Source code of Figure 24 c & d

```
library(dplyr)
library(tidyverse)
library(ggplot2)

#Read-in data and filter for protein coding transcripts
comp_Kozak <- read_csv(paste0("./file_mapped_reads_Kozak-M13.csv"))
comp_Kozak <- dplyr::filter(comp_Kozak, type %in% c("protein_coding"))
comp_M13 <- read.csv(paste0("./file_mapped_reads_M13.csv"), sep=",")
comp_M13 <- dplyr::filter(comp_M13, type %in% c("protein_coding"))

#remove not required columns
comp_Kozak <- comp_Kozak %>%
select(-c("gene", "isoform", "transcript", "id_1", "id_2", "unmapped",
"length"))
comp_M13 <- comp_M13 %>%
select(-c("gene", "isoform", "transcript", "id_1", "id_2", "unmapped",
"length"))

comp_join <- full_join(comp_M13, comp_Kozak, by = c("gene_symbol")) %>%
unique()
#summarize fraction per gene
comp_join <- comp_join %>%  group_by(gene_symbol) %>%
summarise(fraction.x = sum(fraction.x), fraction.y = sum(fraction.y))

#calculate the difference of each fraction
comp_join <- comp_join %>% mutate(diff = fraction.y - fraction.x)
#calculate relative change in fraction
comp_join <- comp_join %>% mutate(factor_fraction = fraction.y/fraction.x)

#plot of more than five-fold enriched genes; remove genes that not present
in both libraries
comp_join %>% na.omit() %>% filter(is.finite(factor_fraction)) %>%
 filter(factor_fraction > 5) %>%
  ggplot(aes(x = reorder(gene_symbol, -factor_fraction),
    y = factor_fraction)) +
  geom_bar(stat = "identity") +
  labs(y= "Fold increase of protein coding transcripts",
    x = "HGNC annotated genomic loci", legend)+
  theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
    legend.title=element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(colour = "black"),
    text = element_text(size = 7), axis.text = element_text(color="black"),
    axis.ticks = element_line(color = "black"))

#plot of reduced genes; remove genes that not present in both libraries
comp_join %>% na.omit() %>% filter(is.finite(factor_fraction)) %>%
 filter(factor_fraction < 1 ) %>%
  ggplot((aes(x=reorder(gene_symbol,-factor_fraction),
    y=(-1/factor_fraction)))) +
  geom_bar(stat = "identity", na.rm=TRUE)+
  labs(y= "Fold reduction of protein coding transcripts",
    x = "HGNC annotated genomic loci", legend)+
  ylim(-100, 0)+
  theme_bw() + theme(panel.border = element_blank(),
    panel.grid.major = element_blank(),
    legend.title=element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(colour = "black"),
    text = element_text(size = 7), axis.text = element_text(color="black"),
    axis.ticks = element_line(color = "black"))
```

Source code for Figure 25 & Table S3

```
library(tidyverse)
library(dplyr)
library(ggplot2)

library <- read.csv("./file_mapped_reads_transcriptome.csv", header = TRUE,
sep=";") %>%
 filter(mapped > 0)

fraction <- library %>% mutate(fraction = mapped/sum(mapped)) %>%
arrange(desc(fraction))

#Donut plot; all transcripts except of TOP10 are summed as "rest"
TOP10 <- head(fraction, n=10) %>% select(transcript_name, fraction)
rest <- tail(fraction,-10) %>% select(transcript_name, fraction)
rest_vector <- c(transcript_name = "rest", fraction = (sum(rest$fraction)))
TOP10_rest <- rbind(TOP10, rest_vector)

TOP10_rest$ymax <- cumsum(TOP10_rest$fraction)
TOP10_rest$ymin <- c(0, head(TOP10_rest$ymax, n=-1))

TOP10_rest$labelPosition <- (TOP10_rest$ymax + TOP10_rest$ymin) / 2
TOP10_rest$label <- paste0(TOP10_rest$transcript_name, "\n",
round(as.numeric((TOP10_rest$fraction))*100, digits=2), "%")

ggplot(TOP10_rest, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3,
fill=transcript_name)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=label), size = 2.46) +
  scale_fill_brewer(palette="BrBG") +
  coord_polar(theta="y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

Commands for Illumina read trimming

Illumina read QC and trimming of all 150 bp the R1 and R2 read files was executed with fastp. Trimmed sequences are depicted for the entry vector pDmS2900.

```
#disabling length filter -L
#enabling base correcting for overlapping PE reads -c
#adding adapter sequences including plasmid sequences for r1 and r2

fastp -w 8 -i data_R1.fq.gz -I data_R2.fq.gz -o out_R1.fq.gz -O
out_R2.fq.gz -L -c --adapter_sequence=GCTCGCCGGTGGCGGAGGAATGCGGAGGAGCAAAAGC
TCATTTCTGAAGAGGACTTGTGGCTGTTTTGGCGGATGAGAGAAGATTTAGATCGGAAGAGCACACGTCTGAACT
CCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG --adapter_sequence_r2=CAAGCAGAAGACGG
CATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAATCTTCTCTCATCCGCCAAAACA
GCCACAAGTCCTCTTCAGAAATGAGCTTTTGCTCCTCCGCATTCCTCCGCCACCGGCGAGTG
```

Commands and source code for Figure 27b

Mapping of R1 reads against the human transcriptome was executed with bowtie2. Samtools was used to convert the resulting .sam files into .bam files.

```
bowtie2-build --threads 4 -f ./GRCh38_latest_rna.fna GRCh38_latest_rna_ind

bowtie2 -p 8 -x ./GRCh38_latest_rna_ind  -U ./out_R1.fq.gz -S
./transcriptome_mapped_reads_R1.sam

samtools view -bS transcriptome_mapped_reads_R1.sam >
transcriptome_mapped_reads_R1.bam

samtools sort transcriptome_mapped_reads_R1.bam -o
transcriptome_mapped_reads_R1_sorted.bam

samtools index transcriptome_mapped_reads_R1_sorted.bam
```

Analysis of the library composition and peptide length was achieved with *R* (see next page).

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(biomaRt)

read_file <- "transcriptome_mapped_reads_R1_sorted.bam"
aln <- scanBam(read_file)
reads <- as.data.frame(aln)
mapped_reads <- reads[reads$flag != '4',]

mapped_reads_filtered <- mapped_reads %>% subset(mapq >=0)
mapped_reads_filtered$rname <- lapply(mapped_reads_filtered$rname, as.char-
acter)
mapped_reads_unique <- mapped_reads_filtered %>%  count(rname, seq)
mapped_reads_unique <- arrange(mapped_reads_unique,desc(n))
mapped_reads_unique$refseq<-gsub(" ", "_", mapped_reads_unique$rname)
mapped_reads_unique$refseq <- gsub("\\..*","", mapped_reads_unique$refseq)
mapped_reads_unique <- arrange(mapped_reads_unique,desc(n))

#Convert ref_seq annotation into hgnc_symbol
ensembl <- biomaRt::useEnsembl(biomart = "genes")
ensembl <- biomaRt::useDataset(dataset = "hsapiens_gene_ensembl", mart =
ensembl)

x <- biomaRt::getBM(attributes = c("refseq_mrna", "hgnc_symbol"),
      filters = "refseq_mrna",
      values = mapped_reads_unique$refseq,
      mart = ensembl) %>% rename(refseq=refseq_mrna)
b <- biomaRt::getBM(attributes = c("refseq_ncrna", "hgnc_symbol"),
                    filters = "refseq_ncrna",
                    values = mapped_reads_unique$refseq,
                    mart = ensembl) %>% rename(refseq=refseq_ncrna)
x <- rbind(x,b)

b <- biomaRt::getBM(attributes = c("refseq_mrna_predicted", "hgnc_symbol"),
                    filters = "refseq_mrna_predicted",
                    values = mapped_reads_unique$refseq,
                    mart = ensembl) %>% rename(refseq=refseq_mrna_pre-
dicted)
x <- rbind(x,b)

b <- biomaRt::getBM(attributes = c("refseq_ncrna_predicted", "hgnc_sym-
bol"),
                    filters = "refseq_ncrna_predicted",
                    values = mapped_reads_unique$refseq,
                    mart = ensembl) %>% rename(refseq=refseq_ncrna_pre-
dicted)
x <- rbind(x,b)
```

XR and XM entries that were not covered by the ensemble database were added manually.

```
#the following list contained manual HGNC_symbols annotations
#for predicted mRNA (XM) and ncRNA (XR)
manual_hgnc <-read.table("manually_HGNC-annotated_transcripts.csv", header
= TRUE, sep = ",",colClasses = c("character", "character", "NULL"))
x <- rbind(x, manual_hgnc)

list_df = list(mapped_reads_unique,x)
mapped_reads_final <- merge(x=mapped_reads_unique,y=x,
                            by="refseq", all.x=TRUE) %>% distinct()

mapped_reads_final <- arrange(mapped_reads_final,desc(n))
mapped_reads_final$hgnc_symbol[is.na(mapped_reads_final$hgnc_symbol)] <-
mapped_reads_final$rname[is.na(mapped_reads_final$hgnc_symbol)]

#sum transcripts over each protein & sum "rest" fraction
fraction <- mapped_reads_final %>% group_by(hgnc_symbol) %>%
  summarise(counts=sum(n), .groups = 'drop')

#Donut plot; all transcripts except of TOP10 are summed as "rest"
TOP10 <- head(fraction, n=10)
rest <- tail(fraction,-10)
rest_vector <- c(hgnc_symbol = "rest", counts = sum(rest$counts), fraction
= sum(rest$fraction) )
TOP10_rest <- rbind(TOP10, rest_vector)

TOP10_rest$ymax <- cumsum(TOP10_rest$fraction)
TOP10_rest$ymin <- c(0, head(TOP10_rest$ymax, n=-1))
TOP10_rest$labelPosition <- (TOP10_rest$ymax + TOP10_rest$ymin) / 2
TOP10_rest$label <- paste0(TOP10_rest$hgnc_symbol, "\n",
round(as.numeric(TOP10_rest$fraction), digits=2), "%")

ggplot(TOP10_rest, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3,
fill=hgnc_symbol)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=label), size=6) +
  scale_fill_brewer(palette="RdYlBu") +
  coord_polar(theta="y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

Source code and commands for in silico translation and proteome mapping of Illumina reads

The following procedure and codes were used for all Illumina analyses via peptide mapping and are identical for Figures 27b & c, 29c, 30b, 33b, S6, S7, and S8 as well as Tables S6, S13, S9, S10, S14, and S15.

In silico translation of trimmed R1 reads was executed in *R*.

```
library(tidyverse)
library(stringr)
library(Biostrings)
library(dplyr)
library(microseq)

read_file <- paste0("out_R1.fq")
trimmed_reads <- microseq::readFastq(read_file)

#fuzzy, non-ambiguous codons will be translated, ambiguous will appear as X
trimmed_reads_lc <- trimmed_reads %>% mutate(Sequence = tolower(Sequence))
trimmed_reads$peptide <- trimmed_reads_lc$Sequence %>% DNAStringSet() %>%
Biostrings::translate(genetic.code=GENETIC_CODE, if.fuzzy.codon="solve")
%>% as.character()

#trim until stop codon
trimmed_reads$peptide <- gsub("\\*.*", "",trimmed_reads$peptide)

#trim primer-based & linker sequence
trimmed_reads$Sequence <- trimmed_reads$peptide %>%
 stringr::str_remove("^SGGG") %>%
 stringr::str_remove("^SPAA")

select(trimmed_reads, Header, Sequence) %>%
 writeFasta(out.file = paste0("./trimmed_peptides.fa"))
```

Proteome mapping of in silico translated peptides was done with the blastp-short command of the blast+ suite against the Swissprot database.

```
blastp -task blastp-short -query ./trimmed_peptides.fa -db
./BLAST+/blastX/swissprot -out ./proteome_mapped_peptides.txt -
num_alignments 1 -outfmt 6 -num_threads 8 -taxids 9606
```

Source code for Figures 27c, S7, S6, S8 and Tables S6 and S13

For clarity, the mapped peptides of the most abundant proteins were omitted from plots of Figure 27c and S6.

```
blast_p_in <- read.table(paste0("./proteome_mapped_peptides.txt"))  %>%
rename(
  "Subject_id" = "V2",
  "alignment_length" = "V4",
  "query_start" = "V7",
  "query_end" = "V8",
  "e_value" = "V11",
  "gap_openings" = "V6"
  ) %>% select("Subject_id", "alignment_length", "query_start",
"query_end", "e_value", "gap_openings") %>%
  subset(gap_openings <= 2) %>%
  subset(e_value <= 0.01) %>%
  unique()

#when highly abundant protein should be omitted include the following
blast_p_in_unbiased <-
blast_p_in[!grepl(c("Q9BZJ4|Q8TBP6|Q9NZ72|Q13813|Q9H4I3|Q04725|P06396"),
 blast_p_in$Subject_id),]
#in this cased the reduced table was used for plotting

blast_p_query_sum <- blast_p_in %>% mutate(m_query_length = query_end-
query_start+1) %>%
  group_by(m_query_length) %>% summarise(n = n()) %>%
arrange(desc(m_query_length))

ggplot(blast_p_query_sum, (aes(x=m_query_length))) +
  geom_col(aes(y=n), fill="grey10")+
  labs(y= "counts", x = "peptide length")+
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0))+
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major =
element_blank(),
  legend.title=element_blank(), panel.grid.minor = element_blank(),
  axis.line = element_line(colour = "black"), text = element_text(size = 7,
color = "black"),
  axis.ticks = element_line(color = "black"), axis.text =
element_text(color="black"))
```

Source code for Figure 29c

```r
libraries = c("p3494", "p3491", "p3492", "p3493")

for (i in libraries){
      blast_p_in <- read.table(paste0("./proteome_mapped_peptides.txt"))
      %>% rename(
        "Subject_id" = "V2",
        "alignment_length" = "V4",t
        "query_start" = "V7",
        "query_end" = "V8",
        "e_value" = "V11",
        "gap_openings" = "V6"
        ) %>% select("Subject_id", "alignment_length", "query_start",
      "query_end", "e_value",
        "gap_openings") %>%
        subset(gap_openings <= 2) %>%
        subset(e_value <= 0.01) %>%
        unique()

      blast_p_in <- blast_p_in[sample(1:nrow(blast_p_in)), ]
      blast_p_in <- blast_p_in[1:150097,]

      blast_p_in <- blast_p_in[,-2] %>% unique()
      blast_p_in_unbiased <-
      blast_p_in[!grepl(c("Q9BZJ4|Q8TBP6|Q9NZ72|Q13813|Q9H4I3|Q04725|P06396
      "),
       blast_p_in$Subject_id),]

      number_peptides <- nrow(blast_p_in_unbiased)
      number_proteins <- n_distinct(blast_p_in_unbiased$Subject_id)
      assign(paste0("peptide_number_",  i), number_peptides)
      assign(paste0("protein_number_",  i), number_proteins)
}

libs_overview <- data.frame(library = c("p3494", "p3491", "p3492",
"p3493"),
peptide_number = c(peptide_number_p3494, peptide_number_p3491,
peptide_number_p3492, peptide_number_p3493),
protein_number = c(protein_number_p3494, protein_number_p3491,
protein_number_p3492, protein_number_p3493))

#set “y = protein_number” to plot proteins
ggplot(libs_overview, aes(x = library, y = peptide_number)) +
  geom_col() +
  ylab("unique peptides") +
  xlab("Library")+
  theme_bw() + theme(panel.border = element_blank(), panel.grid.major =
element_blank(),
    legend.title=element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour =
"black"),
    text = element_text(size = 7, color = "black"), axis.ticks =
element_line(color = "black"),
    axis.text = element_text(color="black"))
```

Source code for Figures 30c, 33b and Tables S9, S10, S14, and S15

First, the mapped peptides of each library were summed by protein. The parameter means as well as standard deviations were calculated.

```
blast_p_in <- read.table(paste0("proteome_mapped_peptides", "c", ".txt"))
%>% rename(
  "V1" = "Query_id",
  "V2" = "Subject_id",
  "V3" = "percent_identity",
  "V4" = "alignment_length",
  "V5" = "mismatches",
  "V6" = "gap_openings",
  "V7" = "query_start",
  "V8" = "query_end",
  "V9" = "subject_start",
  "V10" = "subject_end",
  "V11" = "e_value",
  "V12" = "bit_score"
)

assign(paste0("blastp_in_", i), blast_p_in)
  blast_p_subset <- subset(get(paste0("blastp_in_", i)),
  alignment_length >= 15) %>%
  subset(gap_openings <= 2) %>%
  subset(e_value <= 0.01)

blast_p_subset <- blast_p_subset[order(blast_p_subset$Subject_id),]
blast_p_subset$Subject_id <- gsub("\\..*","", blast_p_subset$Subject_id)

ensembl <- biomaRt::useEnsembl(biomart = "genes")
ensembl <- biomaRt::useDataset(dataset = "hsapiens_gene_ensembl", mart =
ensembl)

hgnc <- biomaRt::getBM(attributes = c("uniprotswissprot", "hgnc_symbol"),
                      filters = "uniprotswissprot",
                      values = blast_p_subset$Subject_id,
                      mart = ensembl) %>% rename("uniprotswissprot"="Sub-
ject_id") %>%
                    unique()
blast_p_subset <- merge(x = blast_p_subset, y = hgnc, by = "Subject_id",
all.x = TRUE)

#calculate mean values
blast_p_subset_mean<- blast_p_subset %>%
  group_by(Subject_id, hgnc_symbol) %>%
  mutate(mean_subject_start = mean(subject_start)) %>%
  mutate(mean_subject_end = mean(subject_end)) %>%
  mutate(mean_e_value = mean(e_value)) %>%
  mutate(mean_alength = mean(alignment_length)) %>%
  mutate(mean_identity = mean(percent_identity)) %>%
  mutate(sd_subject_start = sd(subject_start)) %>%
  mutate(sd_subject_end = sd(subject_end)) %>%
  mutate(sd_e_value = sd(e_value)) %>%
  mutate(sd_alength = sd(alignment_length)) %>%
  mutate(sd_identity = sd(percent_identity)) %>%
  mutate_all(~replace(., is.na(.), 0))

#grouping by name and keep means & sds
blast_p_subset_mean_sum <- blast_p_subset_mean %>%
  group_by(Subject_id, hgnc_symbol, mean_subject_start, sd_subject_start,
  mean_subject_end, sd_subject_end, mean_e_value, sd_e_value, mean_alength,
  sd_alength, mean_identity, sd_identity) %>%
  summarize(total_counts = n())

write_csv(blast_p_subset_mean_sum, paste0("./summed_mapped_proteins.csv"))
```

Afterward, the enriched proteins between the libraries and their EFs were calculated.

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(viridis)

original = c("input_lib") #library which was the input
sorted = c("sorted_lib") #library after sorting

original_reads = as.numeric("377899") #trimmed R1-reads of the original
library
sorted_reads = as.numeric("413193")  #trimmed R1-reads of the sorted
library

original_lib <- read_csv(paste0("./summed_mapped_proteins_input.csv"))
sorted_lib <- read_csv(paste0("./ssummed_mapped_proteins_sorted.csv"))

x <- merge(x=original_lib,y=sorted_lib, by="Subject_id", all.x=TRUE)

x <- x %>%
  mutate(total_counts.y = coalesce(total_counts.y, 0.9)) %>%
  mutate(total_counts.x = coalesce(total_counts.x, 0.9))

#Normalize on R1 reads, calculate enrichment factor
#select for proteins with at least 5 mapped reads in the sorted library
x <- x %>% mutate(total_counts.x_norm = (total_counts.x/original_reads))%>%
  mutate(total_counts.y_norm = (total_counts.y/sorted_reads)) %>%
  mutate(enrichment_factor = (total_counts.y_norm/total_counts.x_norm)) %>%
  arrange(desc(enrichment_factor), desc(total_counts.y)) %>%
  subset(total_counts.y >= 5)

assign(paste0("merged_", original, "_", sorted), x)

write_csv(get(paste0("merged_", original, "_", sorted)),
"./enrichment/enrichment_list.csv")
```

The consecutive EFs for each protein were calculated and used for the heatmap plot (see next page).

```
#Calculate consecutive enrichment factor
enrichment <- c("parent_first_sorting", "first_sorting_second_sorting",
"second_sorting_third_sorting")

for (i in enrichment) {
  df <- read.csv(paste0("enrichment_list_" i, ".csv"))
  if (i == enrichment[1]){
    df_overview_full <- as.data.frame(cbind(hgnc_symbol.x =
df$hgnc_symbol.x,
    enrichment_factor = df$enrichment_factor)) #if parent library: store as
df
  } else {
    df <- dplyr::select(df, "hgnc_symbol.x", "enrichment_factor")#if sorted
library: full_join by hgnc_symbol
    df_overview_full <- df_overview_full %>% full_join(df, by =
"hgnc_symbol.x", all)
  }
}

df_overview_full$enrichment_factor.x <-
as.numeric(as.character(df_overview_full$enrichment_factor.x))
df_overview_full <- df_overview_full %>%
  mutate(enrichment_1 = enrichment_factor.x) %>% #keep first EF
  mutate(enrichment_2 = ifelse(is.na(enrichment_1), enrichment_factor.y,
enrichment_factor.y * enrichment_1)) #multiply first and second EF
df_overview_full <- df_overview_full %>%
  mutate(enrichment_3 = ifelse(is.na(enrichment_2), enrichment_factor,
enrichment_factor * enrichment_2))
df_overview_full <- df_overview_full %>%
rename(enrichment_factor1=enrichment_factor.x,
 enrichment_factor2=enrichment_factor.y,
enrichment_factor3=enrichment_factor)
write_csv(df_overview_full, paste0("./Enrichment_table.csv"))

#Heatmap tile plot
short_for_plot <- df_overview_full %>%
 gather(key = enrichment_round, value = value, enrichment_1:enrichment_3)
short_for_plot$value <- short_for_plot$value %>% log10()
short_for_plot <- left_join(short_for_plot, df_overview_full,
by="hgnc_symbol.x", copy=TRUE) %>%
 group_by(hgnc_symbol.x) %>% fct_inseq(enrichment3)

lmt <- sort(short_for_plot$value, TRUE)[1] #max value for color scale

ggplot(short_for_plot, aes(x=enrichment_round,y=fct_reorder(hgnc_symbol.x,
enrichment_3,
 .na_rm=FALSE, .desc=TRUE), fill=value)) +
  geom_tile()+
  scale_fill_viridis(option= "H", n.breaks=6, limits=c((-lmt+1.55),lmt))+
  scale_x_discrete(expand=c(0,0))+
  scale_y_discrete(expand=c(0,0))+
  labs(y= "Protein", x = "Round of enrichment", legend)+
  coord_fixed()+ theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
  legend.title=element_blank(), panel.grid.minor = element_blank(),
  axis.line = element_line(colour = "black"), text = element_text(size = 7,
color = "black"),
  axis.ticks = element_line(color = "black"), axis.text =
element_text(color="black"))
```

13 References

1. Watson, J. D. & Crick, F. H. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**, 123–131 (1953).
2. Watson, J. D. & Crick, F. H. C. A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
3. Pauling, L. & Corey, R. B. Specific Hydrogen-Bond Formation between Pyrimidines and Purines in Deoxyribonucleic Acids. *Arch Biochem Biophys* **66**, 164–181 (1956).
4. Chargaff, E., Lipshitz, R., Green, C. & Hodes, M. E. The Composition of the Desoxyribonucleic Acid of Salmon Sperm. *Journal of Biological Chemistry* **192**, 223–230 (1951).
5. Clark, D. P., Pazdernik, N. J. & McGehee, M. R. Nucleic Acids and Proteins. in *Molecular Biology* 63–94 (Elsevier, 2019). doi:10.1016/b978-0-12-813288-3.00003-3.
6. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry*. (W.H. Freeman and Company, 2006).
7. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **79**, 233–269 (2010).
8. Bębenek, A. & Ziuzia-Graczyk, I. Fidelity of DNA replication—a matter of proofreading. *Curr Genet* **64**, 985–996 (2018).
9. Méchali, M. Eukaryotic DNA replication origins: Many choices for appropriate answers. *Nat Rev Mol Cell Biol* **11**, 728–738 (2010).
10. Piovesan, A. *et al.* On the length, weight and GC content of the human genome. *BMC Res Notes* **12**, (2019).
11. Huberman, J. A. & Riggs, A. D. Autoradiography of chromosomal DNA fibers from chinese hamster cells. *PNAS* **16**, 599–606 (1966).
12. Meselson, M. & Stahl, F. W. The replication of DNA in Escherichia coli. *PNAS* **44**, 671–682 (1958).
13. Crick, F. H. C. On Protein Synthesis. *Symp. Soc. Exp. Biol.* 138–163 (1958).
14. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 1970 (1970).
15. Shapiro, J. A. Revisiting the central dogma in the 21st century. in *Annals of the New York Academy of Sciences* vol. 1178 6–28 (Blackwell Publishing Inc., 2009).
16. Dai, X., Zhang, S. & Zaleta-Rivera, K. RNA: interactions drive functionalities. *Mol Biol Rep* **47**, 1413–1434 (2020).

17. Lee, J. T. Epigenetic Regulation by Long Noncoding RNAs. *Science (1979)* **338**, 1435–1439 (2012).

18. Clemson, C. M. *et al.* An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Mol Cell* **33**, 717–726 (2009).

19. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621–637 (2018).

20. Ghosh, A. & Lima, C. D. Enzymology of RNA cap synthesis. *Wiley Interdiscip Rev RNA* **1**, 152–172 (2010).

21. Wilkinson, M. E., Charenton, C. & Nagai, K. Annual Review of Biochemistry RNA Splicing by the Spliceosome. *Annual Review of Biochemistry* **89**, 359–388 (2020).

22. Nicholson, A. L. & Pasquinelli, A. E. Tales of Detailed Poly(A) Tails. *Trends Cell Biol* **29**, 191–200 (2019).

23. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol* **3**, (2002).

24. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2002).

25. Prieto, E. I. & Maeshima, K. Dynamic chromatin organization in the cell. *Essays in Biochemistry* vol. 63 133–145 Preprint at https://doi.org/10.1042/EBC20180054 (2019).

26. Jansen, A. & Verstrepen, K. J. Nucleosome Positioning in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews* **75**, 301–320 (2011).

27. Finch, J. T. *et al.* Structure of nucleosome core particles of chromatin. *Nature* **269**, 29–36 (1977).

28. Ruiz-Carrillo, A., Jorcano, J. L., Eder, G. & Lurz, R. In vitro core particle and nucleosome assembly at physiological ionic strength. *Biochemistry* **76**, 3284–3288 (1979).

29. Belmont, A. S. Mitotic chromosome structure and condensation. *Curr Opin Cell Biol* **18**, 632–638 (2006).

30. Hennig, W. Heterochromatin. *Chromosoma* **108**, 1–9 (1999).

31. Lander, S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

32. Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, (2015).

33. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* **6**, 699–708 (2005).

34. Shortt, J. A., Ruggiero, R. P., Cox, C., Wacholder, A. C. & Pollock, D. D. Finding and extending ancient simple sequence repeat-derived regions in the human genome. *Mob DNA* **11**, (2020).

35. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691–703 (2009).

36. O'Donnell, K. A. & Burns, K. H. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mob DNA* **1**, (2010).

37. Hayward, A. & Gilbert, C. Transposable elements. *Current Biology* **32**, R904–R909 (2022).

38. Sharp, A. J. *et al.* Segmental Duplications and Copy-Number Variation in the Human Genome. *Am. J. Hum. Genet* **77**, 78–88 (2005).

39. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science (1979)* **376**, (2022).

40. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**, 559–571 (2010).

41. Tutar, Y. Pseudogenes. *Comp Funct Genomics* **2012**, (2012).

42. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29–59 (2006).

43. Kugel, J. F. & Goodrich, J. A. Finding the start site: redefining the human initiator element. *Genes Dev* **31**, 1–2 (2017).

44. Burley, S. K. The TATA box binding protein. *Curr Opin Struct Biol* **6**, 69–75 (1996).

45. Lim, C. Y. *et al.* The MTE, a new core promoter element for transcription by RNA poymerase II. *Genes Dev* **18**, 1606–1617 (2004).

46. Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene expression. *Curr Opin Genet Dev* **37**, 17–26 (2016).

47. Angeloni, A. & Bogdanovic, O. Sequence determinants, function, and evolution of CpG islands. *Biochem Soc Trans* **49**, 1109–1119 (2021).

48. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6–21 (2002).

49. Felsenfeld, G. A brief history of epigenetics. *Cold Spring Harb Perspect Biol* **6**, 1–10 (2014).

50. Riggs, A. D., Martienssen, R. A. & Russo, V. E. A. Introduction. in *Epigenetic Mechanisms of Gene Regulation* (eds. Russo, V. E. A., Martiensse, R. A. & Riggs, A. D.) 1–4 (Cold Spring Harbor Laboratory Press, 1996).

51. Riggs, A. D. X inactivation, differentiation, and DNA méthylation. *Cytogenet. Cell Genet* **14**, 9–25 (1975).

52. Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene Activity during Development. *Science (1979)* **187**, 226–232 (1975).

53. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

54. Zeng, Y. & Chen, T. DNA methylation reprogramming during mammalian development. *Genes (Basel)* **10**, (2019).

55. Wu, K. J. The epigenetic roles of DNA N6-Methyladenine (6mA) modification in eukaryotes. *Cancer Lett* **494**, 40–46 (2020).

56. Du, Q., Wang, Z. & Schramm, V. L. Human DNMT1 transition state structure. *PNAS* **113**, 2916–2921 (2016).

57. Gujar, H., Weisenberger, D. J. & Liang, G. The roles of human DNA methyltransferases and their isoforms in shaping the epigenome. *Genes (Basel)* **10**, (2019).

58. Buchmuller, B. C., Kosel, B. & Summerer, D. Complete Profiling of Methyl-CpG-Binding Domains for Combinations of Cytosine Modifications at CpG Dinucleotides Reveals Differential Read-out in Normal and Rett-Associated States. *Sci Rep* **10**, (2020).

59. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **6**, 1049–1055 (2014).

60. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol* **11**, 555–557 (2015).

61. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science (1979)* **324**, 929–930 (2009).

62. Wheldon, L. M. *et al.* Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. *Cell Rep* **7**, 1353–1361 (2014).

63. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Genetics* **89**, 1827–1831 (1992).

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

65. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853–862 (2005).

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

67. Carlberg, C. & Molnár, F. *Human epigenomics*. *Human Epigenomics* (Springer Singapore, 2018). doi:10.1007/978-981-10-7614-5.

68. Bai, L. *et al.* Proteome-Wide Profiling of Readers for DNA Modification. *Advanced Science* **8**, (2021).

69. Song, G. *et al.* An all-to-all approach to the identification of sequence-specific readers for epigenetic DNA modifications on cytosine. *Nat Commun* **12**, 1–16 (2021).

70. Raiber, E. A. *et al.* 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat Chem* **10**, 1258–1266 (2018).

71. Ji, S., Shao, H., Han, Q., Seiler, C. L. & Tretyakova, N. Y. Reversible DNA–Protein Cross-Linking at Epigenetic DNA Marks. *Angewandte Chemie* **129**, 14318–14322 (2017).

72. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **14**, 1–11 (2013).

73. Spruijt, C. G. *et al.* Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).

74. Ichimura, T. *et al.* Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. *Journal of Biological Chemistry* **280**, 13928–13935 (2005).

75. Schmolka, N. *et al.* Dissecting the roles of MBD2 isoforms and domains in regulating NuRD complex function during cellular differentiation. *Nat Commun* **14**, 1–13 (2023).

76. Sjolund, A. B., Senejani, A. G. & Sweasy, J. B. MBD4 and TDG: Multifaceted DNA glycosylases with ever expanding biological roles. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **743–744**, 12–25 (2012).

77. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).

78. Buck-Koehntop, B. A. *et al.* Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proc Natl Acad Sci U S A* **109**, 15229–15234 (2012).

79. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet* **17**, 551–565 (2016).

80. Sayeed, S. K., Zhao, J., Sathyanarayana, B. K., Golla, J. P. & Vinson, C. C/EBPβ (CEBPB) protein binding to the C/EBP|CRE DNA 8-mer TTGC|GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. *Biochim Biophys Acta Gene Regul Mech* **1849**, 583–589 (2015).

81. Tan, D. S. *et al.* The homeodomain of Oct4 is a dimeric binder of methylated CpG elements. *Nucleic Acids Res* **51**, 1120–1138 (2023).

82. Zhou, T. *et al.* Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2. *Mol Cell* **54**, 879–886 (2014).

83. Abhishek, S., Nakarakanti, N. K., Deeksha, W. & Rajakumara, E. Mechanistic insights into recognition of symmetric methylated cytosines in CpG and non-CpG DNA by UHRF1 SRA. *Int J Biol Macromol* **170**, 514–522 (2021).

84. Mancini, M., Magnani, E., Macchi, F. & Bonapace, I. M. The multi-functionality of UHRF1: Epigenome maintenance and preservation of genome integrity. *Nucleic Acids Res* **49**, 6053–6068 (2021).

85. Unoki, M. & Sasaki, H. The UHRF protein family in epigenetics, development, and carcinogenesis. *Proc Jpn Acad Ser B Phys Biol Sci* **98**, 401–415 (2022).

86. Khund-Sayeed, S. *et al.* 5-Hydroxymethylcytosine in E-box motifs ACAT|GTG and ACAC|GTG increases DNA-binding of the B-HLH transcription factor TCF4. *Integrative Biology (United Kingdom)* **8**, 936–945 (2016).

87. Torchy, M. P., Hamiche, A. & Klaholz, B. P. Structure and function insights into the NuRD chromatin remodeling complex. *Cellular and Molecular Life Sciences* **72**, 2491–2507 (2015).

88. Covington, K. R. & Fuqua, S. A. W. Role of MTA2 in human cancer. *Cancer and Metastasis Reviews* **33**, 921–928 (2014).

89. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* **8**, 328–330 (2012).

90. Wang, D. *et al.* MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res* **45**, 2396–2407 (2017).

91. Ho, K. L. *et al.* MeCP2 Binding to DNA Depends upon Hydration at Methyl-CpG. *Mol Cell* **29**, 525–531 (2008).

92. Nikolova, E. N., Stanfield, R. L., Dyson, H. J. & Wright, P. E. CH···O Hydrogen Bonds Mediate Highly Specific Recognition of Methylated CpG Sites by the Zinc Finger Protein Kaiso. *Biochemistry* **57**, 2109–2120 (2018).

93. Machado, A. C. D. *et al.* Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct Genomics* **14**, 61–73 (2015).

94. Boulasiki, P., Tan, X. W., Spinelli, M. & Riccio, A. The NuRD Complex in Neurodevelopment and Disease: A Case of Sliding Doors. *Cells* **12**, 1–14 (2023).

95. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).

96. Sharifi Tabar, M. *et al.* Unique protein interaction networks define the chromatin remodelling module of the NuRD complex. *FEBS Journal* **289**, 199–214 (2022).

97. Alendar, A. & Berns, A. Sentinels of chromatin: chromodomain helicase DNA-binding proteins in development and disease. *Genes Dev* **35**, 1403–1430 (2021).

98. Lai, A. Y. & Wade, P. A. Cancer biology and NuRD: A multifaceted chromatin remodelling complex. *Nat Rev Cancer* **11**, 588–596 (2011).

99. Zhang, K., Mosch, K., Fischle, W. & Grewal, S. I. S. Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin. *Nat Struct Mol Biol* **15**, 381–388 (2008).

100. Rothbart, S. B. *et al.* Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat Struct Mol Biol* **19**, 1155–1160 (2012).

101. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185–188 (1999).

102. Zaret, K. S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu Rev Genet* **54**, 1–19 (2020).

103. Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. Identification of a Mammalian Protein That Binds Specifically to DNA Containing Methylated CpGs. *Cell* **59**, 499–507 (1989).

104. Bird, A. & Macleod, D. Reading the DNA methylation signal. in *Cold Spring Harbor Symposia on Quantitative Biology* vol. 69 113–118 (2004).

105. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2013**, 1–16 (2013).

106. Mann, I. K. *et al.* CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res* **23**, 988–997 (2013).

107. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (1979)* **356**, (2017).

108. Kribelbauer, J. F. *et al.* Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep* **19**, 2383–2395 (2017).

109. Buchmuller, B. C. *et al.* Evolved DNA Duplex Readers for Strand-Asymmetrically Modified 5-Hydroxymethylcytosine/5-Methylcytosine CpG Dyads. *J Am Chem Soc* **144**, 2987–2993 (2022).

110. Fagerberg, L., Jonasson, K., Von Heijne, G., Uhlén, M. & Berglund, L. Prediction of the human membrane proteome. *Proteomics* **10**, 1141–1149 (2010).

111. Wang, X., Peterson, J. H. & Bernstein, H. D. Bacterial outer membrane proteins are targeted to the bam complex by two parallel mechanisms. *mBio* **12**, (2021).

112. Kinoshita, T. Glycosylphosphatidylinositol (GPI) anchors: Biochemistry and cell biology: Introduction to a thematic review series. *J Lipid Res* **57**, 4–5 (2016).

113. Braconi, D. *et al.* Surfome analysis of a wild-type wine Saccharomyces cerevisiae strain. *Food Microbiol* **28**, 1220–1230 (2011).

114. Guo, B., Styles, C. A., Feng, Q. & Fink, G. R. A Saccharomyces gene family involved in invasive growth, cell-cell adhesion, and mating. *PNAS* **97**, 12158–12163 (2000).

115. Jose, J. & Meyer, T. F. The Autodisplay Story, from Discovery to Biotechnical and Biomedical Applications. *Microbiology and Molecular Biology Reviews* **71**, 600–619 (2007).

116. Leo, J. C., Oberhettinger, P., Schütz, M. & Linke, D. The inverse autotransporter family: Intimin, invasin and related proteins. *International Journal of Medical Microbiology* **305**, 276–282 (2015).

117. Benz, I. & Schmidt, M. A. Cloning and Expression of an Adhesin (AIDA-I) Involved in Diffuse Adherence of Enteropathogenic Escherichia coli. *Infect Immun* **57**, 1506–1511 (1989).

118. Smith, G. P. Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. *Science (1979)* **228**, 1315–1317 (1985).

119. Jaroszewicz, W., Morcinek-Orłowska, J., Pierzynowska, K., Gaffke, L. & Węgrzyn, G. Phage display and other peptide display technologies. *FEMS Microbiol Rev* **46**, 1–25 (2022).

120. Smith, G. P. *Nobel Lecture: Phage Display: Simple Evolution in a Petri Dish.* (2018).

121. Kieke, M. C., Cho, B. K., Boder, E. T., Kranz, D. M. & Wittrup, K. D. Isolation of anti-T cell receptor scFv mutants by yeast surface display. *Protein Eng* **10**, 1303–1310 (1997).

122. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 553–557 (1997).

123. Agterberg, M., Adriaanse, H., Van Bruggen, A., Karperien, M. & Tommassen, J. Outer-membrane PhoE protein of Escherichia coli K-12 as an exposure vector: possibilities and limitations. *Gene* **88**, 37–45 (1990).

124. Lee, S. Y., Choi, J. H. & Xu, Z. Microbial cell-surface display. *Trends Biotechnol* **21**, 45 (2003).

125. Xu, C. *et al.* Microcystin-LR nanobody screening from an alpaca phage display nanobody library and its expression and application. *Ecotoxicol Environ Saf* **151**, 220–227 (2018).

126. Xia, L., Teng, Q., Chen, Q. & Zhang, F. Preparation and characterization of anti-GPC3 nanobody against hepatocellular carcinoma. *Int J Nanomedicine* **15**, 2197–2205 (2020).

127. Banerjee, S., Singh, A., Rawat, J., Bansal, N. & Maan, S. Dataset of next-generation sequence reads of nanobody clones in phage display library derived from Indian desert camel (Camelus dromedarius L.). *Data Brief* **34**, (2021).

128. Xia, J., Hao, B. I., Qin, Y., Shen, Q. U. & Zong, Y. Construction of Human ScFv Phage Display Library against Ovarian Tumor. *Journal of Huazhong University of Science and Technology* **26**, 497–499 (2006).

129. Liu, J. *et al.* Phage display library selection of a hypoxia-binding scFv antibody for liver cancer metabolic marker discovery. *Oncotarget* **7**, 38105–38121 (2016).

130. Kim, H. Y., Wang, X., Wahlberg, B. & Edwards, W. B. Discovery of hapten-specific scFv from a phage display library and applications for HER2-positive tumor imaging. *Bioconjug Chem* **25**, 1311–1322 (2014).

131. Vandormael, P., Verschueren, P., De Winter, L. & Somers, V. cDNA phage display for the discovery of theranostic autoantibodies in rheumatoid arthritis. *Immunol Res* **65**, 307–325 (2017).

132. Veggiani, G. *et al.* High-affinity chromodomains engineered for improved detection of histone methylation and enhanced CRISPR-based gene repression. *Nat Commun* **13**, (2022).

133. Hiipakka, M., Poikonen, K. & Saksela, K. SH3 Domains with High Affinity and Engineered Ligand Specificities Targeted to HIV-1 Nef. *J Mol Biol* **293**, 1097–1106 (1999).

134. White, K. A. & Zegelbone, P. M. Directed evolution of a probe ligase with activity in the secretory pathway and application to imaging intercellular protein-protein interactions. *Biochemistry* **52**, 3728–3739 (2013).

135. Kumarasamy, J. *et al.* Production, characterization and in-vitro applications of single-domain antibody against thyroglobulin selected from novel T7 phage display library. *J Immunol Methods* **492**, (2021).

136. Xu, H. *et al.* Construction of a T7 phage display nanobody library for bio-panning and identification of chicken dendritic cell-specific binding nanobodies. *Sci Rep* **12**, (2022).

137. Krebs, P., Kurrer, M., Sahin, U., Tureci, O. & Ludewig, B. Autoimmunity seen through the SEREX-scope. *Autoimmun Rev* **2**, 339–345 (2003).

138. Caberoy, N. B., Zhou, Y., Jiang, X., Alvarado, G. & Li, W. Efficient identification of tubby-binding proteins by an improved system of T7 phage display. *Journal of Molecular Recognition* **23**, 74–83 (2010).

139. Caberoy, N. B., Maiguel, D., Kim, Y. & Li, W. Identification of tubby and tubby-like protein 1 as eat-me signals by phage display. *Exp Cell Res* **316**, 245–257 (2010).

140. Danner, S. & Belasco, J. G. T7 phage display: A novel genetic selection system for cloning RNA-binding proteins from cDNA libraries. *PNAS* **98**, 12954–12959 (2001).

141. Ishikawa, H. *et al.* A system for the directed evolution of the insecticidal protein from Bacillus thuringiensis. *Mol Biotechnol* **36**, 90–101 (2007).

142. Han, Z., Xiong, C., Mori, T. & Boyd, M. R. Discovery of a stable dimeric mutant of cyanovirin-N (CV-N) from a T7 phage-displayed CV-N mutant library. *Biochem Biophys Res Commun* **292**, 1036–1043 (2002).

143. Tillotson, B. J., De Larrinoa, I. F., Skinner, C. A., Klavas, D. M. & Shusta, E. V. Antibody affinity maturation using yeast display with detergent-solubilized membrane proteins as antigen sources. *Protein Engineering, Design and Selection* **26**, 101–112 (2013).

144. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 553–557 (1997).

145. Wadle, A. *et al.* Serological identification of breast cancer-related antigens from a Saccharomyces cerevisiae surface display library. *Int J Cancer* **117**, 104–113 (2005).

146. Bidlingmaier, S. & Liu, B. Construction and application of a yeast surface-displayed human cDNA library to identify post-translational modification-dependent protein-protein interactions. *Molecular and Cellular Proteomics* **5**, 533–540 (2006).

147. Bidlingmaier, S. & Liu, B. Interrogating yeast surface-displayed human proteome to identify small molecule-binding proteins. *Molecular and Cellular Proteomics* **6**, 2012–2020 (2007).

148. Bidlingmaier, S., Ha, K., Lee, N. K., Su, Y. & Liu, B. Proteome-wide identification of novel ceramide-binding proteins by yeast surface cdna display and deep sequencing. *Molecular and Cellular Proteomics* **15**, 1232–1245 (2016).

149. Fushimi, T. *et al.* Mutant firefly luciferases with improved specific activity and dATP discrimination constructed by yeast cell surface engineering. *Appl Microbiol Biotechnol* **97**, 4003–4011 (2013).

150. Lipovšek, D. *et al.* Selection of Horseradish Peroxidase Variants with Enhanced Enantioselectivity by Yeast Surface Display. *Chem Biol* **14**, 1176–1185 (2007).

151. Antipov, E., Cho, A. E., Dane Wittrup, K. & Klibanov, A. M. Highly L and D enantioselective variants of horseradish peroxidase discovered by an ultrahigh-throughput selection method. *PNAS* **105**, 17694–17699 (2008).

152. Han, S. yan, Zhang, J. hui, Han, Z. lin, Zheng, S. ping & Lin, Y. Combination of site-directed mutagenesis and yeast surface display enhances Rhizomucor miehei lipase esterification activity in organic solvent. *Biotechnol Lett* **33**, 2431–2438 (2011).

153. Salema, V. *et al.* Selection of Single Domain Antibodies from Immune Libraries Displayed on the Surface of E. coli Cells with Two β-Domains of Opposite Topologies. *PLoS One* **8**, 1–18 (2013).

154. Salema, V. & Fernández, L. Á. Escherichia coli surface display for the selection of nanobodies. *Microb Biotechnol* **10**, 1468–1484 (2017).

155. Salema, V. *et al.* High affinity nanobodies against human epidermal growth factor receptor selected on cells by E. coli display. *MAbs* **8**, 1286–1301 (2016).

156. Palei, S., Becher, K. S., Nienberg, C., Jose, J. & Mootz, H. D. Bacterial Cell-Surface Display of Semisynthetic Cyclic Peptides. *ChemBioChem* **20**, 72–77 (2019).

157. Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L. & Georgiou, G. Antibody affinity maturation using bacterial surface display. *Protein Eng* **11**, 825–832 (1998).

158. Kim, Y.-S., Jung, H.-C. & Pan, J.-G. Bacterial Cell Surface Display of an Enzyme Library for Selective Screening of Improved Cellulase Variants. *Appl Environ Microbiol* **66**, 788–793 (2000).

159. Jung, H. C. *et al.* Bacterial cell surface display of lipase and its randomly mutated library facilitates high-throughput screening of mutants showing higher specific activities. *J Mol Catal B Enzym* **26**, 177–184 (2003).

160. Cho, C. M. H., Mulchandani, A. & Chen, W. Bacterial cell surface display of organophosphorus hydrolase for selective screening of improved hydrolysis of organophosphate nerve agents. *Appl Environ Microbiol* **68**, 2026–2030 (2002).

161. Park, M. Surface display technology for biosensor applications: A review. *Sensors (* **20**, 1–17 (2020).

162. Han, L., Zhao, Y., Cui, S. & Liang, B. Redesigning of Microbial Cell Surface and Its Application to Whole-Cell Biocatalysis and Biosensors. *Appl Biochem Biotechnol* **185**, 396–418 (2018).

163. Machera, S. J., Niedziółka-Jönsson, J. & Szot-Karpińska, K. Phage-based sensors in medicine: A review. *Chemosensors* **8**, 1–25 (2020).

164. Teymennet-Ramírez, K. V., Martínez-Morales, F. & Trejo-Hernández, M. R. Yeast Surface Display System: Strategies for Improvement and Biotechnological Applications. *Front Bioeng Biotechnol* **9**, (2022).

165. Volpicella, M., Gallerani, R., Ceci, L. R., Jongsma, M. A. & Beekwilder, J. Functional expression on bacteriophage of the mustard trypsin inhibitor MTI-2. *Biochem Biophys Res Commun* **280**, 813–817 (2001).

166. McCafferty, J., Jackson, R. H. & Chiswell, D. J. Phage-enzymes: expression and affinity chromatography of functional alkaline phosphatase on the surface of bacteriophage. *Protein Eng* **4**, 955–961 (1991).

167. Rosenberg, A., Griffin, K., Studier, F. W. & Mccormick, M. T7Select® Phage Display System: A Powerful New Protein Display System Based on Bacteriophage T7. *inNovations - Newsletter of Novagen* (1996).

168. Saw, P. E. & Song, E. W. Phage display screening of therapeutic peptide for cancer targeting and therapy. *Protein Cell* **10**, 787–807 (2019).

169. Bratkovič, T. Progress in phage display: Evolution of the technique and its applications. *Cellular and Molecular Life Sciences* **67**, 749–767 (2010).

170. Qi, H., Lu, H., Qiu, H. J., Petrenko, V. & Liu, A. Phagemid vectors for phage display: Properties, characteristics and construction. *J Mol Biol* **417**, 129–143 (2012).

171. Devlin, J. J., Panganiban, L. C. & Devlin, P. E. Random Peptide Libraries: A Source of Specific Protein Binding Molecules. *Science (1979)* **249**, 404–406 (1990).

172. Deutscher, S. Phage Display to Detect and Identify Autoantibodies in Disease. *New England Journal of Medicine* **381**, 89–91 (2019).

173. Yang, F. *et al.* Phage Display-Derived Peptide for the Specific Binding of SARS-CoV-2. *ACS Omega* **7**, 3203–3211 (2022).

174. Liu, K. C. *et al.* Affinity-Selected Bicyclic Peptide G-Quadruplex Ligands Mimic a Protein-like Binding Mechanism. *J Am Chem Soc* **142**, 8367–8373 (2020).

175. Cao, J. *et al.* Phage-display based discovery and characterization of peptide ligands against WDR5. *Molecules* **26**, (2021).

176. Lyu, X. Y. *et al.* WDR5 promotes the tumorigenesis of oral squamous cell carcinoma via CARM1/β-catenin axis. *Odontology* **110**, 138–147 (2022).

177. Fernandez-Gacio, A., Uguen, M. & Fastrez, J. Phage display as a tool for the directed evolution of enzymes. *Trends Biotechnol* **21**, 408–414 (2003).

178. Blum, T. R. *et al.* Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity. *Science (1979)* **371**, 803–810 (2021).

179. Schofield, D. J. *et al.* Application of phage display to high throughput antibody generation and characterization. *Genome Biol* **8**, (2007).

180. Ledsgaard, L. *et al.* Advances in antibody phage display technology. *Drug Discov Today* **27**, 2151–2169 (2022).

181. Alfaleh, M. A. *et al.* Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Front Immunol* **11**, 1–37 (2020).

182. Li, W. & Caberoy, N. B. New perspective for phage display as an efficient and versatile technology of functional proteomics. *Appl Microbiol Biotechnol* **85**, 909–919 (2010).

183. Crameri, R. & Suter, M. Display of biologically active proteins on the surface of filamentous phages: a cDNA cloning system for selection of functional gene products linked to the genetic information responsible for their production. *Gene* **137**, 69–75 (1993).

184. Cicchini, C. *et al.* Searching for DNA-protein interactions by lambda phage display. *J Mol Biol* **322**, 697–706 (2002).

185. Könning, D. & Kolmar, H. Beyond antibody engineering: Directed evolution of alternative binding scaffolds and enzymes using yeast surface display. *Microb Cell Fact* **17**, 1–17 (2018).

186. Lu, C.-F. *et al.* Glycosyl Phosphatidylinositol-dependent Cross-linking of α-Agglutinin and β1,6-Glucan in the Saccharomyces cerevisiae Cell Wall. *J Cell Biol* **128**, 333–340 (1995).

187. Yang, X. *et al.* Development of novel surface display platforms for anchoring heterologous proteins in Saccharomyces cerevisiae. *Microb Cell Fact* **18**, (2019).

188. Marcel Van Der Vaart, J. *et al.* Comparison of Cell Wall Proteins of Saccharomyces cerevisiae as Anchors for Cell Surface Expression of Heterologous Proteins. *Appl Environ Microbiol* **63**, 615–620 (1997).

189. Linciano, S., Pluda, S., Bacchin, A. & Angelini, A. Molecular evolution of peptides by yeast surface display technology. *Medchemcomm* **10**, 1569–1580 (2019).

190. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C. M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Engineering, Design and Selection* **23**, 155–159 (2010).

191. Feldhaus, M. J. *et al.* Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat Biotechnol* **21**, 163–170 (2003).

192. Murai, T., Ueda, M., Kawaguchi, T., Arai, M. & Tanaka, A. Assimilation of Cellooligosaccharides by a Cell Surface-Engineered Yeast Expressing-Glucosidase and Carboxymethylcellulase from Aspergillus aculeatus. *Appl Environ Microbiol* **64**, 4857–4861 (1998).

193. Doerner, A., Rhiel, L., Zielonka, S. & Kolmar, H. Therapeutic antibody engineering by high efficiency cell screening. *FEBS Lett* **588**, 278–287 (2014).

194. Kim, J., Kim, H. K., Jang, H. J., Kim, E. & Kim, M. K. Optimization of yeast surface-displayed cDNA library screening for low abundance targets. *J Microbiol Biotechnol* **25**, 547–553 (2015).

195. Huisman, B. D., Balivada, P. A. & Birnbaum, M. E. Yeast display platform for expression of linear peptide epitopes to assess peptide-MHC-II binding in high-throughput. *Journal of Biological Chemistry* **229**, 1–11 (2023).

196. Wen, F., Esteban, O. & Zhao, H. Rapid identification of CD4+ T-cell epitopes using yeast displaying pathogen-derived peptide library. *J Immunol Methods* **336**, 37–44 (2008).

197. Bowen, J. *et al.* Screening of yeast display libraries of enzymatically treated peptides to discover macrocyclic peptide ligands. *Int J Mol Sci* **22**, 1–20 (2021).

198. Bacon, K. *et al.* Isolation of Chemically Cyclized Peptide Binders Using Yeast Surface Display. *ACS Comb Sci* **22**, 519–532 (2020).

199. Kimura, R. H. *et al.* Functional mutation of multiple solvent-exposed loops in the Ecballium elaterium trypsin inhibitor-II cystine knot miniprotein. *PLoS One* **6**, (2011).

200. Cherf, G. M. & Cochran, J. R. Applications of yeast surface display for protein engineering. *Methods in Molecular Biology* **1319**, 155–175 (2015).

201. Duina, A. A., Miller, M. E. & Keeney, J. B. Budding yeast for budding geneticists: A primer on the Saccharomyces cerevisiae model system. *Genetics* **197**, 33–48 (2014).

202. Jin, D. J., Cagliero, C. & Zhou, Y. N. Growth rate regulation in Escherichia coli. *FEMS Microbiol Rev* **36**, 269–287 (2011).

203. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr Opin Struct Biol* **17**, 467–473 (2007).

204. Agterberg, M., Adriaanse, H., Van Bruggen, A., Karperien, M. & Tommassen, J. Outer-membrane PhoE protein of Escherichia coli K-12 as an exposure vector: possibilities and limitations. *Gene* **88**, 37–45 (1990).

205. Tozakidis, I. E. P., Sichwart, S. & Jose, J. Going beyond E. coli: autotransporter based surface display on alternative host organisms. *N Biotechnol* **32**, 644–650 (2015).

206. van Bloois, E., Winter, R. T., Kolmar, H. & Fraaije, M. W. Decorating microbes: Surface display of proteins on Escherichia coli. *Trends Biotechnol* **29**, 79–86 (2011).

207. Rice, J. J. & Daugherty, P. S. Directed evolution of a biterminal bacterial display scaffold enhances the display of diverse peptides. *Protein Engineering, Design and Selection* **21**, 435–442 (2008).

208. Schneewind, O. & Missiakas, D. M. Protein secretion and surface display in Gram-positive bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 1123–1139 (2012).

209. Green, E. R. & Mecsas, J. Bacterial Secretion Systems: An Overview. *Microbiol Spectr* **4**, (2016).

210. Dhillon, J. K., Drew, P. D. & Porter, A. J. R. Bacterial surface display of an anti-pollutant antibody fragment. *Lett Appl Microbiol* **28**, 350–354 (1999).

211. Michel, L. V. *et al.* Dual orientation of the outer membrane lipoprotein pal in Escherichia coli. *Microbiology (United Kingdom)* **161**, 1251–1259 (2015).

212. Earhart, C. F. Use of an Lpp-OmpA fusion vehicle for bacterial surface display. *Methods Enzymol* **326**, 506–516 (2000).

213. Georgiou, G. *et al.* Display of heterologous proteins on the surface of microorganisms: From the screening of combinatorial libraries to live recombinant vaccines. *Nat Biotechnol* **15**, 29–34 (1997).

214. Malmborg, A.-C., Söderlind, E., Frost, L. & Borrebaeck, C. A. K. Selective Phage Infection Mediated by Epitope Expression on F Pilus. *J Mol Biol* **273**, 544–551 (1997).

215. Stentebjerg-Olesen, B., Pallesen, L., Jensen, L. B., Christiansen, G. & Klemm, P. Authentic display of a cholera toxin epitope by chimeric type I

fimbriae: effects of insert position and host background. *Microbiology (N Y)* **143**, 2027–2038 (1997).

216. Martineau, P., Charbit, A., Leclerc1, C., Werts, C. & Hofnung, M. A Genetic system to elicit and monitor anti-peptide antibodies without peptide synthesis. *Nat Biotechnol* **9**, 170–172 (1991).

217. Janssen, R., Wauben, M., Van Der Zee, R. & Tommassen, J. Immunogenicity of a mycobacterial T-cell epitope expressed in outer membrane protein PhoE of Escherichia coli. *Vaccine* **12**, 406–409 (1994).

218. Taylor, I. M., Harrison, J. L., Timmis, K. N. & O'Connor, C. D. The TraT lipoprotein as a vehicle for the transport of foreign antigenic determinants to the cell surface of Escherichia coli K12: structure–function relationships in the TraT protein. *Mol Microbiol* **4**, 1259–1268 (1990).

219. Sousa, C. *et al.* Metalloadsorption by Escherichia coli Cells Displaying Yeast and Mammalian Metallothioneins Anchored to the Outer Membrane Protein LamB. *J Bacteriol* **180**, 2280–2284 (1998).

220. Bae, W., Chen, W., Mulchandani, A. & Mehra, R. K. Enhanced bioaccumulation of heavy metals by bacterial cells displaying synthetic phytochelatins. *Biotechnol Bioeng* **70**, 518–524 (2000).

221. Xu, Z. & Lee, S. Y. Display of Polyhistidine Peptides on the Escherichia coli Cell Surface by Using Outer Membrane Protein C as an Anchoring Motif. *Appl Environ Microbiol* **65**, 5142–5147 (1999).

222. Bae, W., Mulchandani, A. & Chen, W. Cell surface display of synthetic phytochelatins using ice nucleation protein for enhanced heavy metal bioaccumulation. *J Inorg Biochem* **88**, 223–227 (2002).

223. Richins, R. D., Kaneva, I., Mulchandani, A. & Chen, W. Biodegradation of organophosphorus pesticides by surface-expressed organophosphorus hydrolase. *Nat Biotechnol* **15**, 984–987 (1997).

224. Francisco, J. A., Earhartt, C. F. & Georgiou, G. Transport and anchoring of β-lactamase to the external surface of Escherichia coli. *Biochemistry* **89**, 2713–2717 (1992).

225. Han, M. J. Novel Bacterial Surface Display System Based on the Escherichia coli Protein MipA. *J Microbiol Biotechnol* **30**, 1097–1103 (2020).

226. Jung, H.-C., Lebeault, J.-M. & Pan, J.-G. Surface display of Zymomonas mobilis levansucrase by using the ice-nucleation protein of Pseudomonas syringae. *Nat Biotechnol* **16**, 576–580 (1998).

227. Jung, H.-C., Park, J.-H., Park, S.-H., Lebeault, J.-M. & Pan, J.-G. Expression of carboxymethylcellulase on the surface of Escherichia coli using

Pseudomonas syringae ice nucleation protein. *Enzyme Microb Technol* **22**, 348–354 (1998).

228. Shimazu, M., Mulchandani, A. & Chen, W. Simultaneous degradation of organophosphorus pesticides and p-nitrophenol by a genetically engineered Moraxella sp. with surface-expressed organophosphorus hydrolase. *Biotechnol Bioeng* **76**, 318–324 (2001).

229. Spatola, B. N., Murray, J. A., Kagnoff, M., Kaukinen, K. & Daugherty, P. S. Antibody repertoire profiling using bacterial display identifies reactivity signatures of celiac disease. *Anal Chem* **85**, 1215–1222 (2013).

230. Pantazes, R. J. *et al.* Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. *Sci Rep* **6**, (2016).

231. Li, A., Voleti, R., Lee, M., Gagoski, D. & Shah, N. H. High-throughput profiling of sequence recognition by tyrosine kinases and SH2 domains using bacterial peptide display. *Elife* **12**, (2023).

232. Tucker, A. T. *et al.* Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries. *Cell* **172**, 618-628.e13 (2018).

233. Kjaergaard, K., Sørensen, J. K. & Schembri, M. A. Sequestration of Zinc Oxide by Fimbrial Designer Chelators. *Appl Environ Microbiol* **66**, 10–14 (2000).

234. Nicolay, T., Vanderleyden, J. & Spaepen, S. Autotransporter-based cell surface display in Gram-negative bacteria. *Crit Rev Microbiol* **41**, 109–123 (2013).

235. Jose, J. Autodisplay: Efficient bacterial surface display of recombinant proteins. *Appl Microbiol Biotechnol* **69**, 607–614 (2006).

236. Van Ulsen, P., Zinner, K. M., Jong, W. S. P. & Luirink, J. On display: Autotransporter secretion and application. *FEMS Microbiol Lett* **365**, (2018).

237. Albenne, C. & Ieva, R. Job contenders: roles of the β-barrel assembly machinery and the translocation and assembly module in autotransporter secretion. *Mol Microbiol* **106**, 505–517 (2017).

238. Maurer, J., Jose, J., And, ‡, Meyer, T. F. & Biologie, A. M. Characterization of the Essential Transport Function of the AIDA-I Autotransporter and Evidence Supporting Structural Predictions. *J Bacteriol* **181**, 7014–7020 (1999).

239. Touzé, T., Hayward, R. D., Eswaran, J., Leong, J. M. & Koronakis, V. Self-association of EPEC intimin mediated by the β-barrel-containing anchor

domain: A role in clustering of the Tir receptor. *Mol Microbiol* **51**, 73–87 (2004).

240. Fairman, J. W. *et al.* Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic E. coli and enteropathogenic Y. pseudotuberculosis. *Structure* **20**, 1233–1243 (2012).

241. Luo, Y. *et al.* Crystal structure of enteropathogenic Escherichia coli intimin-receptor complex. *Nature* **405**, 1073–1077 (2000).

242. Wentzel, A., Christmann, A., Adams, T. & Kolmar, H. Display of passenger proteins on the surface of Escherichia coli K-12 by the enterohemorrhagic E. coli intimin EaeA. *J Bacteriol* **183**, 7273–7284 (2001).

243. Denks, K. *et al.* The Sec translocon mediated protein transport in prokaryotes and eukaryotes. *Mol Membr Biol* **31**, 58–84 (2014).

244. Jacob-Dubuisson, F., Fernandez, R. & Coutte, L. Protein secretion through autotransporter and two-partner pathways. *Biochim Biophys Acta Mol Cell Res* **1694**, 235–257 (2004).

245. Saurí, A., Ten Hagen-Jongman, C. M., Van Ulsen, P. & Luirink, J. Estimating the size of the active translocation pore of an autotransporter. *J Mol Biol* **416**, 335–345 (2012).

246. Leyton, D. L. *et al.* Size and conformation limits to secretion of disulfide-bonded loops in autotransporter proteins. *Journal of Biological Chemistry* **286**, 42283–42291 (2011).

247. Adams, T. M., Wentzel, A. & Kolmar, H. Intimin-mediated export of passenger proteins requires maintenance of a translocation-competent conformation. *J Bacteriol* **187**, 522–533 (2005).

248. Jong, W. S. P. *et al.* Limited tolerance towards folded elements during secretion of the autotransporter Hbp. *Mol Microbiol* **63**, 1524–1536 (2007).

249. Kang'ethe, W. & Bernstein, H. D. Charge-dependent secretion of an intrinsically disordered protein via the autotransporter pathway. *Proc Natl Acad Sci U S A* **110**, (2013).

250. Kjærgaard, K., Hasman, H., Schembri, M. A. & Klemm, P. Antigen 43-mediated autotransporter display, a versatile bacterial cell surface presentation system. *J Bacteriol* **184**, 4197–4204 (2002).

251. Kramer, U., Rizos, K., Apfel, H., Autenrieth, I. B. & Lattemann, C. T. Autodisplay: Development of an efficacious system for surface display of antigenic determinants in Salmonella vaccine strains. *Infect Immun* **71**, 1944–1952 (2003).

252. Ming-Ju Chen, Kreuter, J. Y.-T. K. Presentation of Functional Organophosphorus Hydrolase Fusions on the Surface of Escherichia coli

by the AIDA-I Autotransporter Pathway. *Biotechnol Bioeng* **00**, 485–90 (2008).

253. Jose, J. & Von Schwichow, S. Autodisplay of active sorbitol dehydrogenase (SDH) yields a whole cell biocatalyst for the synthesis of rare sugars. *ChemBioChem* **5**, 491–499 (2004).

254. Yang, T. H., Kwon, M. A., Song, J. K., Pan, J. G. & Rhee, J. S. Functional display of Pseudomonas and Burkholderia lipases using a translocator domain of EstA autotransporter on the cell surface of Escherichia coli. *J Biotechnol* **146**, 126–129 (2010).

255. Mukherjee, S. & De Buck, J. Autotransporter-based surface expression and complementation of split TreA fragments utilized for the detection of antibodies against bovine leukemia virus. *J Immunol Methods* **495**, (2021).

256. Kaeßler, A., Olgen, S. & Jose, J. Autodisplay of catalytically active human hyaluronidase hPH-20 and testing of enzyme inhibitors. *European Journal of Pharmaceutical Sciences* **42**, 138–147 (2011).

257. Valls, M., Atrian, S., De Lorenzo, V. & Fernández, L. A. Engineering a mouse metallothionein on the cell surface of Ralstonia eutropha CH34 for immobilization of heavy metals in soil. *Nat Biotechnol* **18**, 661–665 (2000).

258. Schumacher, S. D. & Jose, J. Expression of active human P450 3A4 on the cell surface of Escherichia coli by Autodisplay. *J Biotechnol* **161**, 113–120 (2012).

259. Jose, J., Betscheider, D. & Zangen, D. Bacterial surface display library screening by target enzyme labeling: Identification of new human cathepsin G inhibitors. *Anal Biochem* **346**, 258–267 (2005).

260. Binder, U., Matschiner, G., Theobald, I. & Skerra, A. High-throughput Sorting of an Anticalin Library via EspP-mediated Functional Display on the Escherichia coli Cell Surface. *J Mol Biol* **400**, 783–802 (2010).

261. Friedrich, L. *et al.* Selection of an Anticalin® against the membrane form of Hsp70 via bacterial surface display and its theranostic application in tumour models. *Biol Chem* **399**, 235–252 (2018).

262. Becker, S., Michalczyk, A., Wilhelm, S., Jaeger, K. E. & Kolmar, H. Ultrahigh-throughput screening to identify E. coli cells expressing functionally active enzymes on their surface. *ChemBioChem* **8**, 943–949 (2007).

263. Fleetwood, F., Andersson, K. G., Ståhl, S. & Löfblom, J. An engineered autotransporter-based surface expression vector enables efficient display of Affibody molecules on OmpT-negative E. coli as well as protease-mediated secretion in OmpT-positive strains. *Microb Cell Fact* **13**, (2014).

264. Jong, W. S. P. *et al.* A structurally informed autotransporter platform for efficient heterologous protein secretion and display. *Microb Cell Fact* **11**, (2012).

265. Jose, J., Krämer, J., Klauser, T., Pohlner, J. & Meyer, T. F. Absence of periplasmic DsbA oxidoreductase facilitates export of cysteine-containing passenger proteins to the Escherichia coli cell surface via the Iga autotransporter pathway. *Gene* **178**, 107–110 (1996).

266. Jose, J. & Zangen, D. Autodisplay of the protease inhibitor aprotinin in Escherichia coli. *Biochem Biophys Res Commun* **333**, 1218–1226 (2005).

267. Maurer, J., Jose, J., Meyer, T. F. & Molekulare Biologie, A. Autodisplay: One-Component System for Efficient Surface Display and Release of Soluble Recombinant Proteins from Escherichia coli. *J Bacteriol* **179**, 794–804 (1997).

268. Pardavé-Alejandre, H. D. *et al.* Autodisplay of an avidin with biotin-binding activity on the surface of Escherichia coli. *Biotechnol Lett* **40**, 591–600 (2018).

269. Wilhelm, S. *et al.* Functional cell-surface display of a lipase-specific chaperone. *ChemBioChem* **8**, 55–60 (2007).

270. Kranen, E., Detzel, C., Weber, T. & Jose, J. Autodisplay for the co-expression of lipase and foldase on the surface of E. coli: Washing with designer bugs. *Microb Cell Fact* **13**, (2014).

271. Hörnström, D., Larsson, G., van Maris, A. J. A. & Gustavsson, M. Molecular optimization of autotransporter-based tyrosinase surface display. *Biochim Biophys Acta Biomembr* **1861**, 486–494 (2019).

272. Nicchi, S. *et al.* Decorating the surface of Escherichia coli with bacterial lipoproteins: a comparative analysis of different display systems. *Microb Cell Fact* **20**, (2021).

273. Salema, V., López-Guajardo, A., Gutierrez, C., Mencía, M. & Fernández, L. Á. Characterization of nanobodies binding human fibrinogen selected by E. coli display. *J Biotechnol* **234**, 58–65 (2016).

274. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **14**, 1–11 (2013).

275. Spruijt, C. G. *et al.* Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).

276. de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. *Trends in Genetics* **31**, 128–139 (2015).

277. Jacobo Goebbels, N. Bacterial Surface Display of Human Brain cDNA library: Evaluation of Display Coverage. *Bachelor thesis* (2021).

278. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. & Vitale, L. GeneBase 1.1: A tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database* **2016**, (2016).

279. Renz, P. F., Valdivia Francia, F. & Sendoel, A. Some like it translated: small ORFs in the 5′UTR. *Exp Cell Res* **396**, (2020).

280. Kozak, M. An analysis of S′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**, (1987).

281. Lee, S., Lee, I., Jung, Y., McConkey, D. & Czerniak, B. In-Frame cDNA Library Combined with Protein Complementation Assay Identifies ARL11-Binding Partners. *PLoS One* **7**, 5–12 (2012).

282. Nam, D. K. *et al.* Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *PNAS* **99**, 6152–6156 (2002).

283. Wang, S. M., Fears, S. C., Zhang, L., Chen, J.-J. & Rowley, J. D. Screening poly(dAdT)- cDNAs for gene identification. *PNAS* **97**, 4162–4167 (2000).

284. New England Biolabs Inc. *Instruction Manual NEBuilder® HiFi DNA Assembly Master Mix/ NEBuilder HiFi DNA Assembly Cloning Kit - Version 5.0*. (2023).

285. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

286. Wang, S. M., Fears, S. C., Zhang, L., Chen, J. & Rowley, J. D. Screening Poly [dA/dT(−)] cDNA for Gene Identification. in *Methods in Molecular Biology* vol. 221 197–205 (2000).

287. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 1–16 (2020).

288. Kataev, N. Bacterial Surface Display of human protein fragment libraries: Library construction and screening for 5mC-specific readers. (Technical University of Dortmund, 2023).

289. Wheelan, S. J., Marchler-Bauer, A. & Bryant, S. H. Domain size distributions can predict domain boundaries. *Bioinformatics* **16**, 613–618 (2000).

290. Xiong, J. *et al.* Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. *Mol Cell* **64**, 913–925 (2016).

291. Jin, S. G. *et al.* Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. *Cell Rep* **14**, 493–505 (2016).

292. Hashimoto, H. *et al.* Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev* **28**, 2304–2313 (2014).

293. Natan, E. *et al.* Interaction of the p53 DNA-binding domain with its N-terminal extension modulates the stability of the p53 tetramer. *J Mol Biol* **409**, 358–368 (2011).

294. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).

295. Webb, H. *et al.* The FOXP2 forkhead domain binds to a variety of DNA sequences with different rates and affinities. *J Biochem* **162**, 45–54 (2017).

296. Lambert, M. *et al.* Direct and indirect targeting of HOXA9 transcription factor in acute myeloid leukemia. *Cancers (Basel)* **11**, 1–38 (2019).

297. Kalniņa, Z. *et al.* Evaluation of T7 and lambda phage display systems for survey of autoantibody profiles in cancer patients. *J Immunol Methods* **334**, 37–50 (2008).

298. Lin, H. S. *et al.* Autoantibody approach for serum-based detection of head and neck cancer. *Cancer Epidemiology Biomarkers and Prevention* **16**, 2396–2405 (2007).

299. Caberoy, N. B., Zhou, Y., Alvarado, G., Fan, X. & Li, W. Efficient identification of phosphatidylserine-binding proteins by ORF phage display. *Biochem Biophys Res Commun* **386**, 197–201 (2009).

300. Bidlingmaier, S., Ha, K., Lee, N. K., Su, Y. & Liu, B. Proteome-wide identification of novel ceramide-binding proteins by yeast surface cdna display and deep sequencing. *Molecular and Cellular Proteomics* **15**, 1232–1245 (2016).

301. Kronqvist, N., Löfblom, J., Jonsson, A., Wernérus, H. & Ståhl, S. A novel affinity protein selection system based on staphylococcal cell surface display and flow cytometry. *Protein Engineering, Design and Selection* **21**, 247–255 (2008).

302. Siegele, D. A. & Hu, J. C. Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *PNAS* **94**, 8168–8172 (1997).

303. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, (2009).

304. Li, W. ORF phage display to identify cellular proteins with different functions. *Methods* **58**, 2–9 (2012).

305. Deangelo, D. J., DeFalco, J., Rybacki, L. & Childs, G. The Embryonic Enhancer-Binding Protein SSAP Contains a Novel DNA-Binding Domain

Which Has Homology to Several RNA-Binding Proteins. *Mol Cell Biol* **15**, 1254–1264 (1995).

306. Basu, A., Dong, B., Krainer, A. R. & Howe, C. C. The Intracisternal A-Particle Proximal Enhancer-Binding Protein Activates Transcription and Is Identical to the RNA-and DNA-Binding Protein p54 nrb /NonO. *Mol Cell Biol* **17**, 677–686 (1997).

307. Newberry, E. P., Latifi, T. & Towler, D. A. The RRM domain of MINT, a novel Msx2 binding protein, recognizes and regulates the rat osteocalcin promoter. *Biochemistry* **38**, 10678–10690 (1999).

308. Hamimes, S., Bourgeon, D., Stasiak, A. Z., Stasiak, A. & Van Dyck, E. Nucleic acid-binding properties of the RRM-containing protein RDM1. *Biochem Biophys Res Commun* **344**, 87–94 (2006).

309. Hwang, S., Guo, Z. & Kuznetsov, I. B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **23**, 634–636 (2007).

310. Yan, J. & Kurgan, L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic Acids Res* **45**, (2017).

311. Iliev, D. B., Skjæveland, I. & Jørgensen, J. B. CpG oligonucleotides bind TLR9 and RRM-Containing proteins in Atlantic Salmon (Salmo salar). *BMC Immunol* **14**, (2013).

312. Wang, Y. *et al.* Mapping the transcription repressive domain in the highly conserved human gene hnulp1. *Frontiers of Biology in China* **3**, 137–142 (2008).

313. Nachmias, D. *et al.* Asgard ESCRT-III and VPS4 reveal conserved chromatin binding properties of the ESCRT machinery. *ISME Journal* **17**, 117–129 (2023).

314. Cho, S. J., Lee, H. S., Dutta, S., Seog, D. H. & Moon, I. S. Translation elongation factor-1A1 (eEF1A1) localizes to the spine by domain III. *BMB Rep* **45**, 227–232 (2012).

315. Negrutskii, B., Vlasenko, D., Mirande, M., Futernyk, P. & El'skaya, A. mRNA-Independent way to regulate translation elongation rate in eukaryotic cells. *IUBMB Life* **70**, 192–196 (2018).

316. Enders, M., Neumann, P., Dickmanns, A. & Ficner, R. Structure and function of spliceosomal DEAH-box ATPases. *Biol Chem* **404**, 851–866 (2023).

317. Cai, Z. *et al.* hnulp1, a basic helix-loop-helix protein with a novel transcriptional repressive domain, inhibits transcriptional activity of serum response factor. *Biochem Biophys Res Commun* **343**, 973–981 (2006).

318. Henne, W. M., Buchkovich, N. J. & Emr, S. D. The ESCRT Pathway. *Dev Cell* **21**, 77–91 (2011).

319. Yorikawa, C. *et al.* Human CHMP6, a myristoylated ESCRT-III protein, interacts directly with an ESCRT-II component EAP20 and regulates endosomal cargo sorting. *Biochem. J* **387**, 17–26 (2005).

320. Goliand, I., Nachmias, D., Gershony, O. & Elia, N. Inhibition of ESCRT-II-CHMP6 interactions impedes cytokinetic abscission and leads to cell death. *Mol Biol Cell* **25**, 3740–3748 (2014).

321. Talledge, N. *et al.* Preprint: The ESCRT-III proteins IST1 and CHMP1B assemble around nucleic acids. *bioRxiv* (2018) doi:10.1101/386532.

322. Stauffer, D. R., Howard, T. L., Nyun, T. & Hollenberg, S. M. CHMP1 is a novel nuclear matrix protein affecting chromatin structure and cell-cycle progression. *J Cell Sci* **114**, 2383–2393 (2001).

323. Patrick, W. M., Firth, A. E. & Blackburn, J. M. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng* **16**, 451–457 (2003).

324. Roloff, T. C., Ropers, H. H. & Nuber, U. A. Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics* **4**, (2003).

325. Kato, H. *et al.* Architecture of the high mobility group nucleosomal protein 2-nucleosome complex as revealed by methyl-based NMR. *PNAS* **108**, 12283–12288 (2011).

326. Huang, Z. L. *et al.* Identification of G-Quadruplex-Binding Protein from the Exploration of RGG Motif/G-Quadruplex Interactions. *J Am Chem Soc* **140**, 17945–17955 (2018).

327. Zhu, X., Bührer, C. & Wellmann, S. Cold-inducible proteins CIRP and RBM3, a unique couple with activities far beyond the cold. *Cellular and Molecular Life Sciences* **73**, 3839–3859 (2016).

328. Sun, H. *et al.* CRIP1 cooperates with BRCA2 to drive the nuclear enrichment of RAD51 and to facilitate homologous repair upon DNA damage induced by chemotherapy. *Oncogene* **40**, 5342–5355 (2021).

329. Gorleku, O. A., Barns, A. M., Prescott, G. R., Greaves, J. & Chamberlain, L. H. Endoplasmic reticulum localization of DHHC palmitoyltransferases mediated by lysine-based sorting signals. *Journal of Biological Chemistry* **286**, 39573–39584 (2011).

330. Solis, G. P. *et al.* Local and substrate-specific S-palmitoylation determines subcellular localization of Gαo. *Nat Commun* **13**, (2022).

331. Nartey, N., Banerjee, D. & George Cherian, M. Immunohistochemical localization of the metallothionein in cell nucleus and cytoplasm of fetal human liver and kidey and its changes during development. *Pathology* **19**, 233–238 (1987).

332. Boeynaems, S. *et al.* Preprint: Aberrant phase separation is a common killing strategy of positively charged peptides in biology and human disease. *bioRxiv* (2023) doi:10.1101/2023.03.09.531820.

333. Das, L., Quintana, V. G. & Sweasy, J. B. NTHL1 in genomic integrity, aging and cancer. *DNA Repair (Amst)* **93**, (2020).

334. Jian, H. *et al.* Alteration of mRNA 5-Methylcytosine Modification in Neurons After OGD/R and Potential Roles in Cell Stress Response and Apoptosis. *Front Genet* **12**, (2021).

335. Pirok, E. W. *et al.* APBP-1, a DNA/RNA-binding protein, interacts with the chick aggrecan regulatory region. *Journal of Biological Chemistry* **280**, 35606–35616 (2005).

336. Kim, Y.-M. & Hong, S. Controversial roles of cold-inducible RNA-binding protein in human cancer (Review). *Int J Oncol* **59**, (2021).

337. Hu, Y. *et al.* RBM3 is an outstanding cold shock protein with multiple physiological functions beyond hypothermia. *J Cell Physiol* **237**, 3788–3802 (2022).

338. Jayawardena, D. P., Heinemann, I. U. & Stillman, M. J. Zinc binds non-cooperatively to human liver metallothionein 2a at physiological pH. *Biochem Biophys Res Commun* **493**, 650–653 (2017).

339. Carroll, B. L. *et al.* Caught in motion: human NTHL1 undergoes interdomain rearrangement necessary for catalysis. *Nucleic Acids Res* **49**, 13165–13178 (2021).

340. Shinmura, K. *et al.* Defective repair capacity of variant proteins of the DNA glycosylase NTHL1 for 5-hydroxyuracil, an oxidation product of cytosine. *Free Radic Biol Med* **131**, 264–273 (2019).

341. Fromme, J. C. & Verdine, G. L. Structure of a trapped endonuclease III-DNA covalent intermediate. *EMBO J* **22**, 3461–3471 (2003).

342. Temiz, N. A., Donohue, D. E., Bacolla, A., Luke, B. T. & Collins, J. R. The role of methylation in the intrinsic dynamics of B- and Z-DNA. *PLoS One* **7**, (2012).

343. Westwood, M. N. *et al.* Single-Base Lesions and Mismatches Alter the Backbone Conformational Dynamics in DNA. *Biochemistry* **60**, 873–885 (2021).

344. He, D. Y., Neasta, J. & Ron, D. Epigenetic regulation of BDNF expression via the scaffolding protein RACK1. *Journal of Biological Chemistry* **285**, 19043–19050 (2010).

345. Yang, I. V. *et al.* Relationship of DNA methylation and gene expression in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **190**, 1263–1272 (2014).

346. Thompson, N. A. *et al.* Combinatorial CRISPR screen identifies fitness effects of gene paralogues. *Nat Commun* **12**, (2021).

347. Anver, S. *et al.* Yeast X-chromosome-associated protein 5 (Xap5) functions with H2A.Z to suppress aberrant transcripts. *EMBO Rep* **15**, 894–902 (2014).

348. Sedlacek, Z. *et al.* Human and Mouse XAP-5 and XAP-5-like (X5L) Genes: Identification of an Ancient Functional Retroposon Differentially Expressed in Testis. *Genomics* **61**, 125–132 (1999).

349. Man Chan, H. & George Cherian, M. Ontogenic changes in hepatic metallothionein isoforms in prenatal and newborn rats. *Biochem. Cell Biol.* **71**, 133–140 (1993).

350. Sabolić, I. *et al.* Expression and immunolocalization of metallothioneins MT1, MT2 and MT3 in rat nephron. *Journal of Trace Elements in Medicine and Biology* **46**, 62–75 (2018).

351. Ebadi, M. *et al.* Metallothionein-mediated neuroprotection in genetically engineered mouse models of Parkinson's disease. *Molecular Brain Research* **134**, 67–75 (2005).

352. Elfaki, I., Bayer, P. & Mueller, J. W. A potential transcriptional regulator is out-of-frame translated from the metallothionein 2A messenger RNA. *Anal Biochem* **409**, 159–161 (2011).

353. Ludyga, N. *et al.* The impact of Cysteine-Rich Intestinal Protein 1 (CRIP1) in human breast cancer. *Mol Cancer* **12**, (2013).

354. Xie, H. *et al.* Combining peptide and DNA for protein assay: CRIP1 detection for breast cancer staging. *ACS Appl Mater Interfaces* **6**, 459–463 (2014).

355. Itoh, H. *et al.* Identification of hepatocyte growth factor activator inhibitor type 2 (HAI-2)-related small peptide (H2RSP): Its nuclear localization and generation of chimeric mRNA transcribed from both HAI-2 and H2RSP genes. *Biochem Biophys Res Commun* **288**, 390–399 (2001).

356. Naganuma, S. *et al.* Nuclear translocation of H2RSP is impaired in regenerating intestinal epithelial cells of murine colitis model. *Virchows Archiv* **448**, 354–360 (2006).

357. Luo, Q. *et al.* Immortalization-upregulated protein promotes pancreatic cancer progression by regulating NPM1/FHL1-mediated cell-cycle-checkpoint protein activity. *Cell Biol Toxicol* **39**, 2069–2087 (2022).

358. Kim, J.-K., Ryll, R., Ishizuka, Y. & Kato, S. Identification of cDNAs encoding two novel nuclear proteins, IMUP-1 and IMUP-2, upregulated in SV40-immortalized human fibroblasts. *Gene* **257**, 327–334 (2000).

359. Tan, H. W. *et al.* Single-gene knockout-coupled omics analysis identifies C9orf85 and CXorf38 as two uncharacterized human proteins associated with ZIP8 malfunction. *Front Mol Biosci* **9**, (2022).

360. Li, H. & Chang, Q. Regulation and function of stimulus-induced phosphorylation of MeCP2. *Front Biol (Beijing)* **9**, 367–375 (2014).

361. Tan, C. P. & Nakielny, S. Control of the DNA Methylation System Component MBD2 by Protein Arginine Methylation. *Mol Cell Biol* **26**, 7224–7235 (2006).

362. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).

363. Wentzel, A., Christmann, A., Adams, T. & Kolmar, H. Display of Passenger Proteins on the Surface of Escherichia coli K-12 by the Enterohemorrhagic E. coli Intimin EaeA. *J Bacteriol* **183**, 7273–7284 (2001).

364. Dotmatics. SnapGene® software.

365. Nam, D. K. *et al.* Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *PNAS* **99**, 6152–6156 (2002).

366. Buchmuller, B. C. *et al.* Evolved DNA Duplex Readers for Strand-Asymmetrically Modified 5 - Hydroxymethylcytosine/5-Methylcytosine CpG Dyads. *J Am Chem Soc* (2022) doi:10.1021/jacs.1c10678.

367. Subramanian, S. K., Russ, W. P. & Ranganathan, R. A set of experimentally validated, mutually orthogonal primers for combinatorially specifying genetic components. *Synth Biol* **3**, 1–5 (2018).

368. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* (2023) doi:10.1002/imt2.107.

369. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849–864 (2017).

370. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

371. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

372. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, (2019).

373. Wickham H, François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. Preprint at (2023).

374. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

375. Wickham, H. *et al.* svglite: An 'SVG' Graphics Device. Preprint at (2023).

376. Dowle, M. S. A. _data.table: Extension of `data.frame`_. Preprint at (2023).

377. Xu, S. *et al.* Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front Genet* **12**, (2021).

378. Snipen, L. L. K. _microseq: Basic Biological Sequence Handling_. Preprint at (2021).

379. Pagès, H. A. P. G. R. D. S. Biostrings: Efficient manipulation of biological strings. Preprint at (2022).

380. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. Preprint at (2022).

381. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, 1–7 (2017).

382. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol* **9**, (2011).

383. National Center for Biotechnology Information (US). *The NCBI Handbook [Internet]*. (2019).

384. Wernersson, R. Virtual Ribosome - A comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res* **34**, (2006).

14 Supplementary Information

14.1 Supplementary Figures

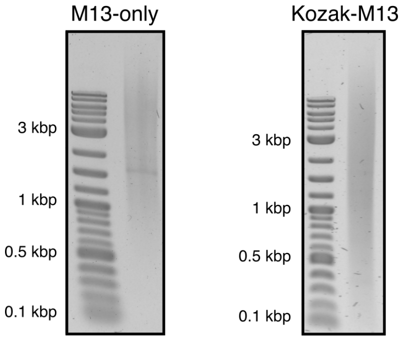

Figure S1: Amplicons of the human prostate cDNA library 10108-A (BioCat) with M13 fw and rv primer (M13-only) or anti-Kozak primer mix and M13 rv primer (Kozak-M13). These amplicons were later subjected to Nanopore sequencing.

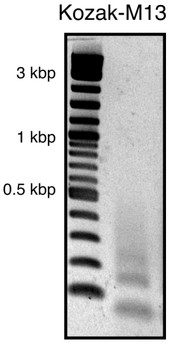

Figure S2: PCR product of human prostate cDNA (BioCat) with anti-Kozak primer mix and M13-rev primer.

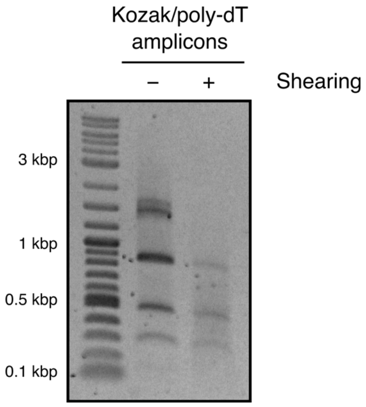

Figure S3: Amplicons of human prostate cDNA library before and after random shearing via sonication.

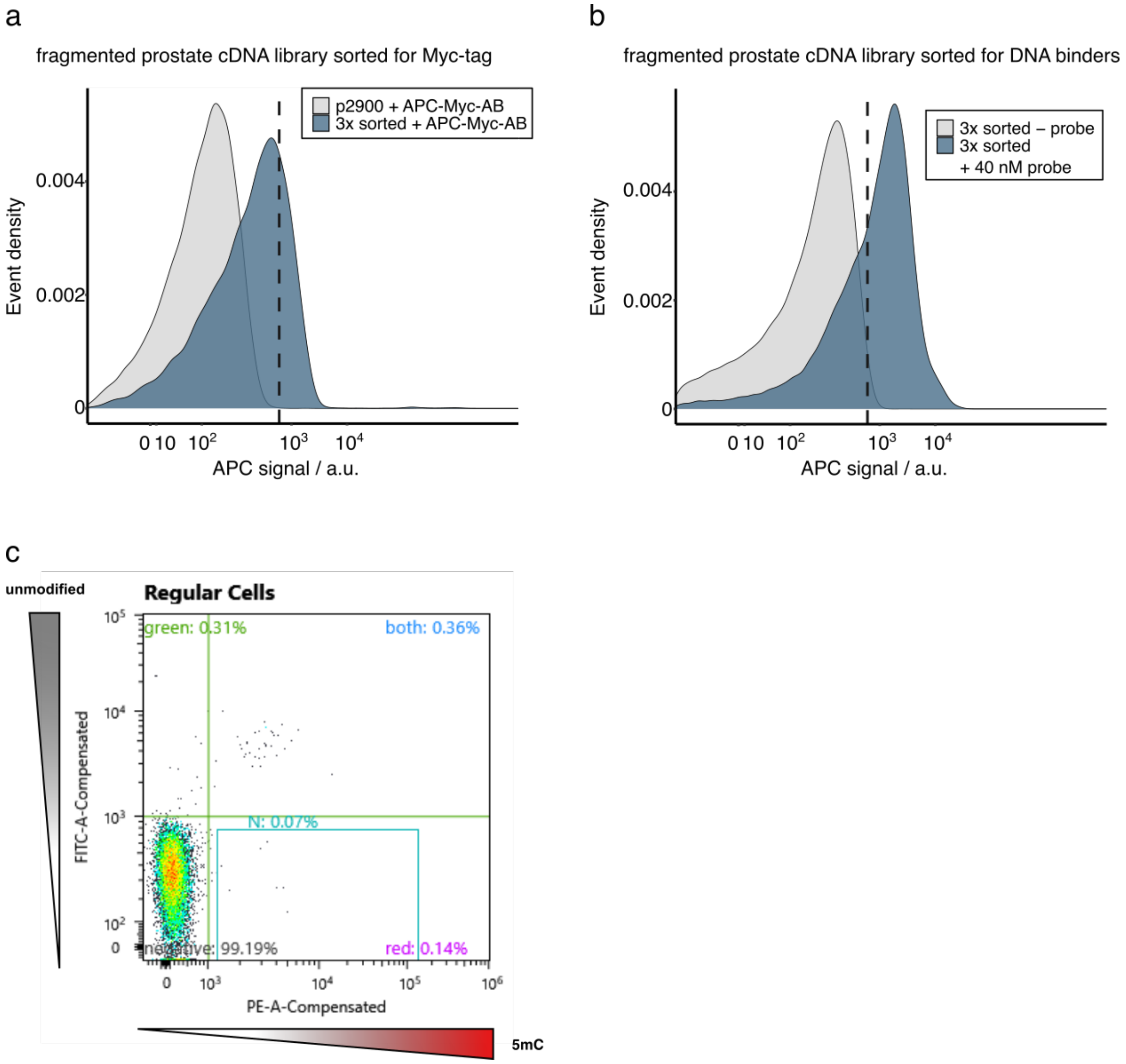

Figure S4: Negative control FACS signals of induced BL21(DE3) Tuner cells harboring (a) the empty entry vector for fragmented cDNA libraries p2900 stained with the Myc-AB (Myc-tag is out of frame), (b) the library enriched for DNA binder without the probe, and the respective library after three rounds of selection. (c) Cells harboring the empty entry vector for full-length CDS library p2569 stained with the 8NmC-PE/8NC-FITC double staining mix.

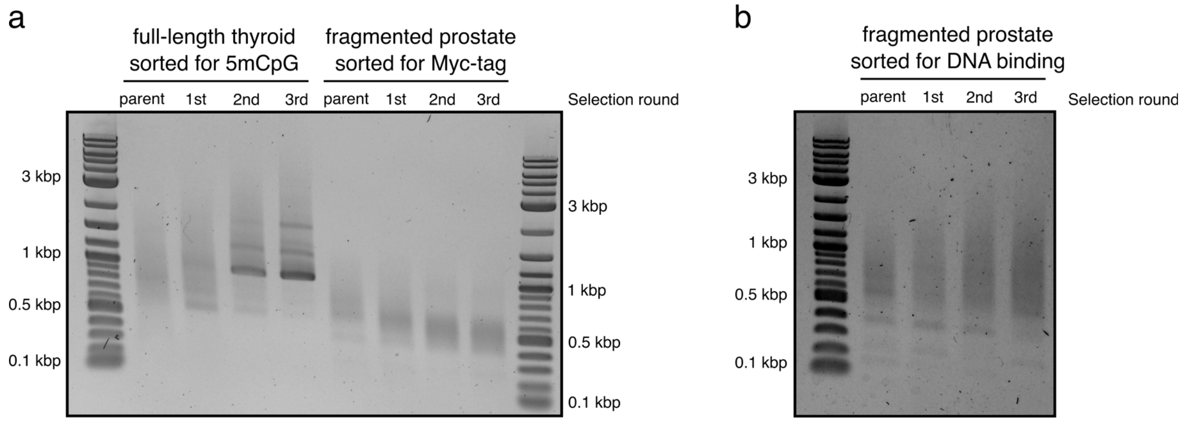

Figure S5: Amplicons of the two-step PCR introducing NGS adapters and barcodes. The isolated plasmid library of the parent libraries as well as the plasmid libraries of the sorted full-length CDS (thyroid) and fragmented CDS (prostate) selected for 8NmC probe and Myc-tag probes (a) and 30NC probe (b) were used as templates.

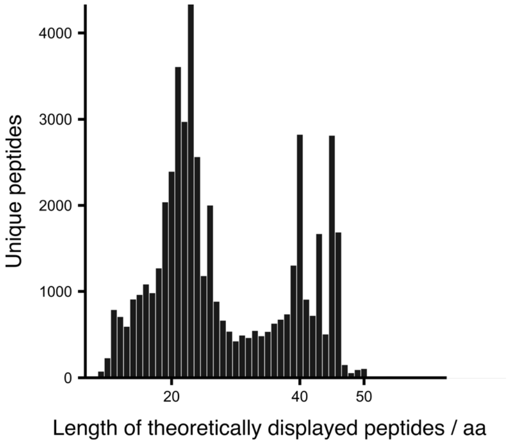

Figure S6: Distribution of the length of proteome-mapped ORFs of the fragmented prostate display library, that could theoretically be displayed, as identified from 150 bp Illumina reads. The length determination was limited to 50 residues by the Illumina read length.

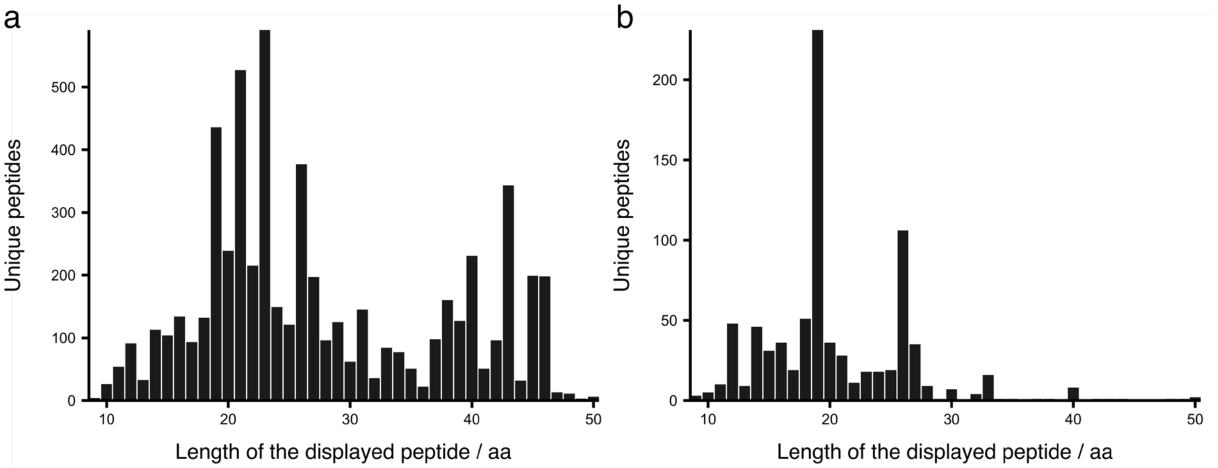

Figure S7: Length distribution of unique, proteome-mapped ORFs enriched for the C-terminal Myc-tag. (a) Distribution of all enriched, mapped ORFs, (b) distribution of enriched ORFs, omitting highly abundant proteins of the parent library.

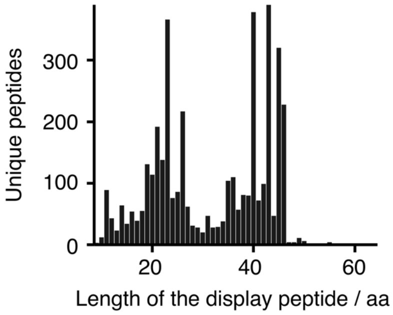

Figure S8: Length distribution of all proteome mapped, unique dsDNA probe-enriched ORFs.

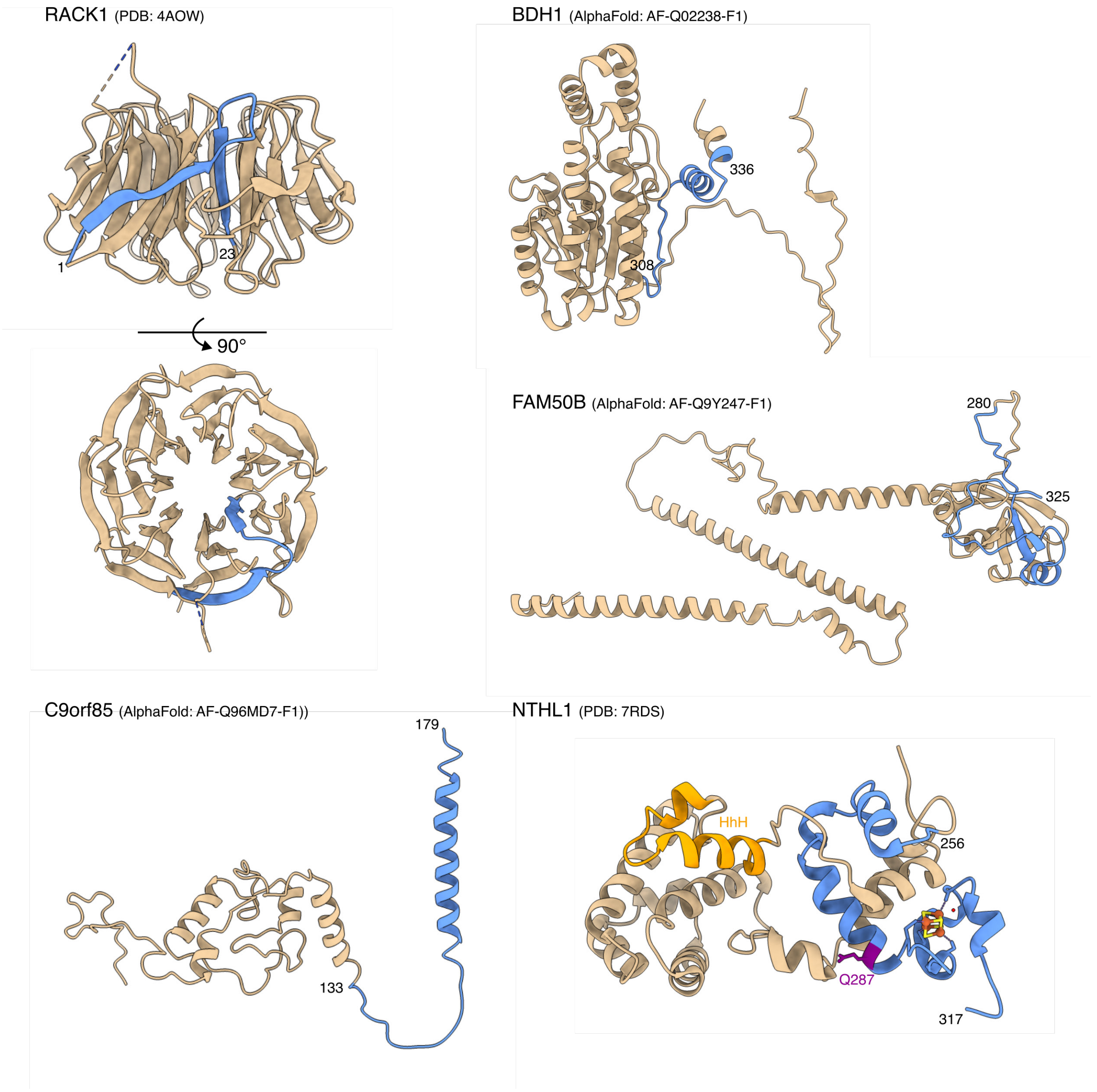
RACK1 (PDB: 4AOW)
1
23
90°
BDH1 (AlphaFold: AF-Q02238-F1)
336
308
FAM50B (AlphaFold: AF-Q9Y247-F1)
280
325
C9orf85 (AlphaFold: AF-Q96MD7-F1))
179
133
NTHL1 (PDB: 7RDS)
HhH
256
Q287
317

Figure S9: Crystal structure and AlphaFold models of the protein fragments enriched for 5mCpG binding. The enriched parts of the proteins in highlighted in blue. Numbers represent the start and potential end residues of the enriched fragments. The DNA backbone-interacting residue Q287 and the HhH domain of NTHL1 are depicted in purple and orange.

14.2 Supplementary Tables

Table S1: Diversity determination of the full-length CDS human thyroid library via counting of colony forming units (cfu) after transformation of the cloned plasmid pool. Dilutions of 1:1,000 and 1:10,000 of two transformations were plated on agar plates.

	1:1,000	**1:10,000**	**Mean**
Transformation 1	225	27	247,500 ± 22,500
Transformation 2	NA	NA	775,500 ± 54,500
Total (after pooling)			1,023,000 ± 77,000

Table S2: Analysis of the plasmid inserts of seven clones of the full-length thyroid CDS display library. DNA sequences were mapped against the human transcriptome with blastN sequence analysis tool.[383] The DNA sequence obtained from Sanger sequencing was translated in-frame with the Virtual Ribosome suite[384] to the intimin sequence and analysed regarding, whether it contained the native start codon and covered the full-length sequence.

	Covered start codon	Covered poly-A tail	Expressed ORF	Full-length sequence	transcript name
1	No	No	Primer derived residues		
2	No	No	Primer derived residues		
3	No	No	Primer derived residues	No	Homo sapiens acidic nuclear phosphoprotein 32 family member B (ANP32B), mRNA
4	Yes	Yes	MGREFGNLTRMRHVIS YSLSPFEQRAYPHVFTK GIPNVLRRIRESFFRVVP QFVVFYLIYTWGTEEFER SKRKNPAAYENDK	Yes	Homo sapiens ubiquinol-cytochrome c reductase complex III subunit VII (UQCRQ), mRNA
5	Yes	Yes	MAEEGIAAGGVMDVN TALQEVLKTALIHDGLA RGIREAAKALDKRQAH LCVLASNCDEPMYVKL VEALCAEHQINLIKVDD NKKLGEWVGLCKIDRE GKPRKVVGCSCVVVKD YGKESQAKDVIEEYFKC KK	Yes	Homo sapiens ribosomal protein S12 (RPS12), mRNA
6	Yes	Yes	MAEEGIAAGGVMDVN TALQEVLKTALIHDGLA RGIREAAKALDKRQAH LCVLASNCDEPMYVKV EALCAEHQINLIKVDDN KKLGEWVGLCKIDREG KPRKVVGCSCVVVKDY	Yes	Homo sapiens ribosomal protein S12 (RPS12), mRNA

			GRESQAKDVIEEYFKCK K		
7	Likely	Yes	XDTSRVXPIKLARVTKV LGRTGSQGQCXXVRVEF XDDTSRSIIRNVKGPVRE GDVLTLLESEREARRLR	Yes	Homo sapiens ribosomal protein S28 (RPS28), mRNA

Table S3: Transcript composition of the full-length human thyroid CDS display library. Transcript and gene IDs and symbols according to HGNC and ensemble databases for the top 200 most abundant transcripts as determined by mapping of Nanopore sequencing reads against the human transcriptome (GCh37) with minimap2[285].

Ensemble transcript ID	Ensemble gene ID	HGNC Transcript name	HGNC gene symbol	Length	RNA type	Fraction
ENST00000230050.4_2	ENSG00000112306.8_5	RPS12-201	RPS12	503	protein_coding	0.103578
ENST00000361427.6_2	ENSG00000198830.11_5	HMGN2-201	HMGN2	1940	protein_coding	0.094552
ENST00000234875.9_5	ENSG00000116251.11_8	RPL22-201	RPL22	2061	protein_coding	0.057345
ENST00000600659.3_2	ENSG00000233927.5_5	RPS28-203	RPS28	1330	protein_coding	0.050731
ENST00000344063.7_3	ENSG00000108107.15_6	RPL28-201	RPL28	4209	protein_coding	0.030475
ENST00000361436.10_4	ENSG00000197958.13_7	RPL12-201	RPL12	634	protein_coding	0.028896
ENST00000245185.6_2	ENSG00000125148.7_5	MT2A-201	MT2A	401	protein_coding	0.025172
ENST00000572008.5_3	ENSG00000134419.15_7	RPS15A-208	RPS15A	797	nonsense_mediated_decay	0.02398
ENST00000619352.4_1	ENSG00000198830.11_5	HMGN2-210	HMGN2	1148	protein_coding	0.022312
ENST00000563390.5_3	ENSG00000134419.15_7	RPS15A-203	RPS15A	525	protein_coding	0.01975
ENST00000559463.5_1	ENSG00000108107.15_6	RPL28-208	RPL28	852	protein_coding	0.019721
ENST00000519543.5_1	ENSG00000042832.12_5	TG-208	TG	3132	protein_coding	0.016503
ENST00000220616.9_2	ENSG00000042832.12_5	TG-201	TG	8455	protein_coding	0.015401
ENST00000460355.1_1	ENSG00000144426.18_7	NBEAL1-204	NBEAL1	3595	retained_intron	0.011916
ENST00000218388.9_2	ENSG00000102265.12_5	TIMP1-201	TIMP1	769	protein_coding	0.010694
ENST00000268668.11_2	ENSG00000140990.15_7	NDUFB10-201	NDUFB10	675	protein_coding	0.010456
ENST00000469257.2_2	ENSG00000173812.11_6	EIF1-203	EIF1	2338	protein_coding	0.010039
ENST00000377017.5_1	ENSG00000102265.12_5	TIMP1-202	TIMP1	626	protein_coding	0.009682
ENST00000273550.12_4	ENSG00000167996.16_9	FTH1-201	FTH1	1203	protein_coding	0.009056
ENST00000297258.11_2	ENSG00000164687.11_5	FABP5-201	FABP5	676	protein_coding	0.007775
ENST00000569148.1_1	ENSG00000140990.15_7	NDUFB10-204	NDUFB10	628	protein_coding	0.00703
ENST00000523756.5_1	ENSG00000042832.12_5	TG-218	TG	5065	nonsense_mediated_decay	0.006822
ENST00000484616.2_1	ENSG00000112306.8_5	RPS12-202	RPS12	639	retained_intron	0.006226
ENST00000322989.8_4	ENSG00000134419.15_7	RPS15A-201	RPS15A	2203	protein_coding	0.005243
ENST00000322028.5_4	ENSG00000177700.6_6	POLR2L-201	POLR2L	876	protein_coding	0.005124
ENST00000396394.7_3	ENSG00000129084.18_8	PSMA1-201	PSMA1	1195	protein_coding	0.005005
ENST00000358435.9_4	ENSG00000196683.11_6	TOMM7-201	TOMM7	434	protein_coding	0.004141

ENST00000378667.1_1	ENSG00000164405.11_5	UQCRQ-202	UQCRQ	429	protein_coding	0.004141
ENST00000378670.8_2	ENSG00000164405.11_5	UQCRQ-203	UQCRQ	1573	protein_coding	0.003664
ENST00000497965.5_4	ENSG00000116251.11_8	RPL22-207	RPL22	786	protein_coding	0.003664
ENST00000299166.9_4	ENSG00000166136.16_7	NDUFB8-201	NDUFB8	678	protein_coding	0.003515
ENST00000260379.11_2	ENSG00000137818.12_5	RPLP1-201	RPLP1	1174	protein_coding	0.003426
ENST00000480661.1_3	ENSG00000116251.11_8	RPL22-206	RPL22	2279	retained_intron	0.003396
ENST00000543683.6_1	ENSG00000140990.15_7	NDUFB10-202	NDUFB10	830	protein_coding	0.003396
ENST00000648033.1_3	ENSG00000285589.1_5	AC010422.8-201	AC010422.8	5025	nonsense_mediated_decay	0.003247
ENST00000037243.7_2	ENSG00000034713.8_5	GABARAPL2-201	GABARAPL2	974	protein_coding	0.003217
ENST00000570172.1_3	ENSG00000140990.15_7	NDUFB10-205	NDUFB10	495	protein_coding	0.003068
ENST00000362079.2_1	ENSG00000198938.2_1	MT-CO3-201	MT-CO3	784	protein_coding	0.002949
ENST00000313115.11_5	ENSG00000177556.12_8	ATOX1-201	ATOX1	471	protein_coding	0.00286
ENST00000620041.4_3	ENSG00000167996.16_9	FTH1-212	FTH1	825	protein_coding	0.002651
ENST00000498095.4_1	ENSG00000148362.11_8	PAXX-208	PAXX	1077	retained_intron	0.002562
ENST00000524142.5_4	ENSG00000177556.12_8	ATOX1-207	ATOX1	749	protein_coding	0.002502
ENST00000285605.8_2	ENSG00000155254.13_5	MARVELD1-201	MARVELD1	3213	protein_coding	0.002413
ENST00000299529.7_2	ENSG00000166426.8_5	CRABP1-201	CRABP1	738	protein_coding	0.002324
ENST00000474471.7_4	ENSG00000153446.16_7	C16orf89-203	C16orf89	1829	protein_coding	0.002175
ENST00000586629.5_3	ENSG00000153446.16_7	C16orf89-204	C16orf89	576	retained_intron	0.002115
ENST00000315997.5_3	ENSG00000153446.16_7	C16orf89-201	C16orf89	1872	protein_coding	0.002085
ENST00000497825.5_3	ENSG00000197958.13_7	RPL12-204	RPL12	864	processed_transcript	0.002085
ENST00000472572.8_4	ENSG00000153446.16_7	C16orf89-202	C16orf89	1360	protein_coding	0.001787
ENST00000299299.4_4	ENSG00000166228.9_6	PCBD1-201	PCBD1	787	protein_coding	0.001728
ENST00000565420.5_3	ENSG00000134419.15_7	RPS15A-205	RPS15A	802	protein_coding	0.001698
ENST00000343267.8_5	ENSG00000189058.9_8	APOD-201	APOD	862	protein_coding	0.001638
ENST00000565031.1_1	ENSG00000140990.15_7	NDUFB10-203	NDUFB10	958	retained_intron	0.001638
ENST00000492564.2_1	ENSG00000148362.11_8	PAXX-206	PAXX	701	retained_intron	0.001579
ENST00000346786.2_1	ENSG00000101335.10_5	MYL9-202	MYL9	1024	protein_coding	0.001549
ENST00000421807.7_4	ENSG00000132376.20_7	INPP5K-204	INPP5K	2706	protein_coding	0.001549
ENST00000417088.2_1	ENSG00000233927.5_5	RPS28-201	RPS28	334	retained_intron	0.001519
ENST00000361739.1_1	ENSG00000198712.1_1	MT-CO2-201	MT-CO2	684	protein_coding	0.00143
ENST00000519178.5_1	ENSG00000042832.12_5	TG-206	TG	3706	protein_coding	0.0014
ENST00000596731.7_6	ENSG00000105583.11_9	WDR83OS-202	WDR83OS	619	protein_coding	0.0014

ENST00000463765.1_1	ENSG00000148362.11_8	PAXX-202	PAXX	496	retained_intron	0.00137
ENST00000272438.9_4	ENSG00000144043.12_7	TEX261-201	TEX261	3365	protein_coding	0.001341
ENST00000334976.11_5	ENSG00000161671.17_9	EMC10-201	EMC10	9439	protein_coding	0.001281
ENST00000398733.8_3	ENSG00000131508.16_8	UBE2D2-202	UBE2D2	2530	protein_coding	0.001281
ENST00000370322.5_3	ENSG00000166136.16_7	NDUFB8-203	NDUFB8	713	protein_coding	0.001251
ENST00000622754.4_5	ENSG00000143409.15_10	MINDY1-207	MINDY1	2637	protein_coding	0.001251
ENST00000379607.10_4	ENSG00000173674.11_6	EIF1AX-202	EIF1AX	4414	protein_coding	0.001221
ENST00000380358.9_5	ENSG00000170004.17_11	CHD3-203	CHD3	7361	protein_coding	0.001221
ENST00000252891.8_5	ENSG00000105245.9_7	NUMBL-201	NUMBL	3561	protein_coding	0.001192
ENST00000279022.7_2	ENSG00000101335.10_5	MYL9-201	MYL9	2786	protein_coding	0.001132
ENST00000511884.6_2	ENSG00000169851.15_8	PCDH7-205	PCDH7	6769	protein_coding	0.001132
ENST00000312210.9_4	ENSG00000143409.15_10	MINDY1-201	MINDY1	2400	protein_coding	0.001102
ENST00000361227.2_1	ENSG00000198840.2_1	MT-ND3-201	MT-ND3	346	protein_coding	0.001072
ENST00000361936.9_5	ENSG00000143409.15_10	MINDY1-203	MINDY1	2817	protein_coding	0.001072
ENST00000512805.6_4	ENSG00000204628.12_9	RACK1-227	RACK1	1140	protein_coding	0.001072
ENST00000567300.1_1	ENSG00000125148.7_5	MT2A-205	MT2A	416	processed_transcript	0.001072
ENST00000256015.5_4	ENSG00000133639.6_7	BTG1-201	BTG1	4629	protein_coding	0.001043
ENST00000378024.9_8	ENSG00000124942.14_11	AHNAK-202	AHNAK	18761	protein_coding	0.001013
ENST00000398734.8_1	ENSG00000131508.16_8	UBE2D2-203	UBE2D2	901	nonsense_mediated_decay	0.001013
ENST00000378292.9_5	ENSG00000198467.16_11	TPM2-202	TPM2	1182	protein_coding	9.83E-04
ENST00000400137.9_5	ENSG00000088832.18_9	FKBP1A-204	FKBP1A	1514	protein_coding	9.83E-04
ENST00000202556.14_4	ENSG00000088808.18_10	PPP1R13B-201	PPP1R13B	5500	protein_coding	9.23E-04
ENST00000289577.9_4	ENSG00000158604.15_8	TMED4-201	TMED4	2203	protein_coding	9.23E-04
ENST00000395068.9_2	ENSG00000178719.17_6	GRINA-202	GRINA	1858	protein_coding	8.94E-04
ENST00000558131.1_1	ENSG00000108107.15_6	RPL28-205	RPL28	444	protein_coding	8.94E-04
ENST00000601780.5_1	ENSG00000161671.17_9	EMC10-208	EMC10	2178	nonsense_mediated_decay	8.94E-04
ENST00000647748.1_4	ENSG00000088808.18_10	PPP1R13B-218	PPP1R13B	5187	protein_coding	8.94E-04
ENST00000301904.4_2	ENSG00000168077.14_5	SCARA3-201	SCARA3	3787	protein_coding	8.64E-04
ENST00000371620.4_2	ENSG00000148362.11_8	PAXX-201	PAXX	808	protein_coding	8.64E-04
ENST00000569417.6_3	ENSG00000167965.18_8	MLST8-232	MLST8	1671	protein_coding	8.64E-04
ENST00000222190.9_3	ENSG00000105583.11_9	WDR83OS-201	WDR83OS	599	protein_coding	8.04E-04
ENST00000290101.8_4	ENSG00000076864.19_10	RAP1GAP-201	RAP1GAP	3584	protein_coding	8.04E-04
ENST00000373840.9_4	ENSG00000119396.11_7	RAB14-201	RAB14	4149	protein_coding	8.04E-04

ENST00000313269.5_1	ENSG00000178719.17_6	GRINA-201	GRINA	1968	protein_coding	7.75E-04
ENST00000566835.5_1	ENSG00000167965.18_8	MLST8-226	MLST8	1565	nonsense_mediated_decay	7.75E-04
ENST00000354546.10_4	ENSG00000197043.14_7	ANXA6-201	ANXA6	2889	protein_coding	7.45E-04
ENST00000361479.10_2	ENSG00000196199.14_5	MPHOSPH8-201	MPHOSPH8	4239	protein_coding	7.45E-04
ENST00000622181.4_4	ENSG00000211445.12_12	GPX3-210	GPX3	1760	protein_coding	7.45E-04
ENST00000262238.10_3	ENSG00000100811.14_9	YY1-201	YY1	6534	protein_coding	7.15E-04
ENST00000376918.7_5	ENSG00000161671.17_9	EMC10-202	EMC10	2014	protein_coding	7.15E-04
ENST00000536368.1_3	ENSG00000197958.13_7	RPL12-205	RPL12	499	protein_coding	7.15E-04
ENST00000568194.5_1	ENSG00000167965.18_8	MLST8-230	MLST8	2116	retained_intron	7.15E-04
ENST00000643350.1_1	ENSG00000197912.16_12	SPG7-236	SPG7	2475	processed_transcript	7.15E-04
ENST00000353703.9_3	ENSG00000166913.13_6	YWHAB-201	YWHAB	3020	protein_coding	6.85E-04
ENST00000457408.7_7	ENSG00000158604.15_8	TMED4-203	TMED4	2294	protein_coding	6.85E-04
ENST00000597793.5_1	ENSG00000104960.15_5	PTOV1-209	PTOV1	1716	retained_intron	6.85E-04
ENST00000357847.9_2	ENSG00000146416.19_8	AIG1-202	AIG1	2136	protein_coding	6.55E-04
ENST00000248975.6_2	ENSG00000128245.15_9	YWHAH-201	YWHAH	1751	protein_coding	6.26E-04
ENST00000360031.6_2	ENSG00000197586.13_12	ENTPD6-202	ENTPD6	2766	protein_coding	6.26E-04
ENST00000460976.5_3	ENSG00000175029.17_9	CTBP2-206	CTBP2	603	processed_transcript	6.26E-04
ENST00000482587.5_5	ENSG00000169905.13_7	TOR1AIP2-203	TOR1AIP2	7067	protein_coding	6.26E-04
ENST00000592634.5_3	ENSG00000174886.14_8	NDUFA11-205	NDUFA11	2154	protein_coding	6.26E-04
ENST00000215793.13_4	ENSG00000099995.19_11	SF3A1-201	SF3A1	5090	protein_coding	5.96E-04
ENST00000308961.5_4	ENSG00000174886.14_8	NDUFA11-201	NDUFA11	575	protein_coding	5.96E-04
ENST00000422725.4_4	ENSG00000228594.4_9	FNDC10-201	FNDC10	2124	protein_coding	5.96E-04
ENST00000520358.7_2	ENSG00000157570.12_7	TSPAN18-207	TSPAN18	4557	protein_coding	5.96E-04
ENST00000622663.1_1	ENSG00000111961.18_6	SASH1-204	SASH1	5075	protein_coding	5.96E-04
ENST00000624174.1_1	ENSG00000279322.1_7	AL035252.4-201	AL035252.4	288	TEC	5.96E-04
ENST00000645818.2_5	ENSG00000197912.16_12	SPG7-268	SPG7	3076	protein_coding	5.96E-04
ENST00000330720.7_4	ENSG00000105438.9_8	KDELR1-201	KDELR1	1550	protein_coding	5.66E-04
ENST00000388825.9_7	ENSG00000211445.12_12	GPX3-201	GPX3	1603	protein_coding	5.66E-04
ENST00000474879.7_8	ENSG00000100258.18_9	LMF2-202	LMF2	2593	protein_coding	5.66E-04
ENST00000485936.5_1	ENSG00000197586.13_12	ENTPD6-217	ENTPD6	2588	processed_transcript	5.66E-04
ENST00000565096.6_1	ENSG00000064666.15_10	CNN2-207	CNN2	2109	protein_coding	5.66E-04
ENST00000650687.2_4	ENSG00000109846.9_9	CRYAB-217	CRYAB	774	protein_coding	5.66E-04
ENST00000233813.5_4	ENSG00000115461.5_8	IGFBP5-201	IGFBP5	6239	protein_coding	5.36E-04

ENST00000284719.8_4	ENSG00000138430.16_6	OLA1-201	OLA1	4251	protein_coding	5.36E-04
ENST00000330233.11_1	ENSG00000213145.10_6	CRIP1-201	CRIP1	1311	protein_coding	5.36E-04
ENST00000354989.9_3	ENSG00000197586.13_12	ENTPD6-201	ENTPD6	2708	protein_coding	5.36E-04
ENST00000361144.9_1	ENSG00000189337.17_10	KAZN-201	KAZN	3838	protein_coding	5.36E-04
ENST00000376652.9_4	ENSG00000197586.13_12	ENTPD6-203	ENTPD6	4104	protein_coding	5.36E-04
ENST00000378665.1_1	ENSG00000164405.11_5	UQCRQ-201	UQCRQ	586	protein_coding	5.36E-04
ENST00000382120.4_3	ENSG00000109610.6_6	SOD3-201	SOD3	1416	protein_coding	5.36E-04
ENST00000561702.6_2	ENSG00000197912.16_12	SPG7-203	SPG7	3729	retained_intron	5.36E-04
ENST00000564572.1_1	ENSG00000064666.15_10	CNN2-206	CNN2	2266	retained_intron	5.36E-04
ENST00000264202.8_8	ENSG00000077549.19_10	CAPZB-201	CAPZB	1675	protein_coding	5.06E-04
ENST00000266079.5_2	ENSG00000101161.8_5	PRPF6-201	PRPF6	3047	protein_coding	5.06E-04
ENST00000301724.14_1	ENSG00000167965.18_8	MLST8-201	MLST8	1735	protein_coding	5.06E-04
ENST00000356805.9_5	ENSG00000115306.16_8	SPTBN1-202	SPTBN1	10211	protein_coding	5.06E-04
ENST00000361899.2_1	ENSG00000198899.2_1	MT-ATP6-201	MT-ATP6	681	protein_coding	5.06E-04
ENST00000481247.6_4	ENSG00000183826.18_8	BTBD9-206	BTBD9	8530	protein_coding	5.06E-04
ENST00000556597.1_5	ENSG00000088808.18_10	PPP1R13B-214	PPP1R13B	3473	nonsense_mediated_decay	5.06E-04
ENST00000392531.4_2	ENSG00000213145.10_6	CRIP1-202	CRIP1	672	protein_coding	4.77E-04
ENST00000429938.1_2	ENSG00000106211.10_7	HSPB1-202	HSPB1	527	protein_coding	4.77E-04
ENST00000463620.1_3	ENSG00000112983.18_8	BRD8-216	BRD8	764	retained_intron	4.77E-04
ENST00000490187.1_1	ENSG00000197586.13_12	ENTPD6-218	ENTPD6	1918	processed_transcript	4.77E-04
ENST00000579016.6_4	ENSG00000167280.17_7	ENGASE-205	ENGASE	4603	protein_coding	4.77E-04
ENST00000588627.1_1	ENSG00000213228.5_6	RPL12P38-202	RPL12P38	2319	processed_transcript	4.77E-04
ENST00000216080.5_7	ENSG00000100258.18_9	LMF2-201	LMF2	2658	protein_coding	4.47E-04
ENST00000263097.9_3	ENSG00000064666.15_10	CNN2-201	CNN2	2149	protein_coding	4.47E-04
ENST00000272227.8_6	ENSG00000143870.13_8	PDIA6-201	PDIA6	2279	protein_coding	4.47E-04
ENST00000376666.3_1	ENSG00000197586.13_12	ENTPD6-204	ENTPD6	2100	protein_coding	4.47E-04
ENST00000431098.1_1	ENSG00000204196.5_4	RPL12P16-201	RPL12P16	498	processed_pseudogene	4.47E-04
ENST00000433259.6_5	ENSG00000197586.13_12	ENTPD6-209	ENTPD6	2654	protein_coding	4.47E-04
ENST00000449223.3_1	ENSG00000233927.5_5	RPS28-202	RPS28	1523	retained_intron	4.47E-04
ENST00000472174.7_2	ENSG00000117410.14_6	ATP6V0B-205	ATP6V0B	987	protein_coding	4.47E-04
ENST00000527950.5_5	ENSG00000109846.9_9	CRYAB-207	CRYAB	887	protein_coding	4.47E-04
ENST00000540494.5_3	ENSG00000143870.13_8	PDIA6-207	PDIA6	2509	protein_coding	4.47E-04
ENST00000562958.6_1	ENSG00000064666.15_10	CNN2-205	CNN2	2194	protein_coding	4.47E-04

ENST00000617249.4_3	ENSG00000143870.13_8	PDIA6-208	PDIA6	2344	protein_coding	4.47E-04
ENST00000674361.1_2	ENSG00000241743.4_9	XACT-203	XACT	347375	lncRNA	4.47E-04
ENST00000266041.9_5	ENSG00000055955.17_9	ITIH4-201	ITIH4	3282	protein_coding	4.17E-04
ENST00000288398.10_3	ENSG00000140416.23_12	TPM1-202	TPM1	1295	protein_coding	4.17E-04
ENST00000296456.10_2	ENSG00000164062.13_8	APEH-201	APEH	2879	protein_coding	4.17E-04
ENST00000382533.8_3	ENSG00000150459.12_9	SAP18-201	SAP18	2298	protein_coding	4.17E-04
ENST00000396359.1_1	ENSG00000164687.11_5	FABP5-202	FABP5	986	protein_coding	4.17E-04
ENST00000397124.5_2	ENSG00000167965.18_8	MLST8-203	MLST8	1691	protein_coding	4.17E-04
ENST00000404371.6_3	ENSG00000143870.13_8	PDIA6-203	PDIA6	2682	protein_coding	4.17E-04
ENST00000409393.6_1	ENSG00000213145.10_6	CRIP1-203	CRIP1	677	protein_coding	4.17E-04
ENST00000460563.5_1	ENSG00000198830.11_5	HMGN2-202	HMGN2	1045	processed_transcript	4.17E-04
ENST00000465150.6_3	ENSG00000126768.12_5	TIMM17B-204	TIMM17B	964	protein_coding	4.17E-04
ENST00000471374.1_1	ENSG00000128245.15_9	YWHAH-205	YWHAH	1735	processed_transcript	4.17E-04
ENST00000532829.5_3	ENSG00000167996.16_9	FTH1-208	FTH1	912	nonsense_mediated_decay	4.17E-04
ENST00000561504.1_1	ENSG00000148671.14_6	ADIRF-203	ADIRF	700	nonsense_mediated_decay	4.17E-04
ENST00000609111.1_3	ENSG00000272734.1_7	ADIRF-AS1-203	ADIRF-AS1	3822	lncRNA	4.17E-04
ENST00000313146.10_3	ENSG00000143319.17_5	ISG20L2-201	ISG20L2	3303	protein_coding	3.87E-04
ENST00000320451.7_4	ENSG00000197961.12_6	ZNF121-201	ZNF121	6987	protein_coding	3.87E-04
ENST00000372013.8_2	ENSG00000148671.14_6	ADIRF-201	ADIRF	627	protein_coding	3.87E-04
ENST00000377767.9_4	ENSG00000083520.15_6	DIS3-201	DIS3	10571	protein_coding	3.87E-04
ENST00000397492.1_1	ENSG00000128245.15_9	YWHAH-202	YWHAH	1879	protein_coding	3.87E-04
ENST00000501122.2_1	ENSG00000245532.9_7	NEAT1-202	NEAT1	22743	lncRNA	3.87E-04
ENST00000562017.1_1	ENSG00000125148.7_5	MT2A-203	MT2A	903	retained_intron	3.87E-04
ENST00000586602.5_3	ENSG00000197961.12_6	ZNF121-202	ZNF121	7177	protein_coding	3.87E-04
ENST00000219204.8_2	ENSG00000102931.8_6	ARL2BP-201	ARL2BP	1969	protein_coding	3.57E-04
ENST00000263645.10_2	ENSG00000110651.12_7	CD81-201	CD81	1482	protein_coding	3.57E-04
ENST00000317508.11_4	ENSG00000052344.16_6	PRSS8-201	PRSS8	1837	protein_coding	3.57E-04
ENST00000336854.8_1	ENSG00000123395.14_4	ATG101-201	ATG101	1439	protein_coding	3.57E-04
ENST00000358278.7_3	ENSG00000140416.23_12	TPM1-206	TPM1	1717	protein_coding	3.57E-04
ENST00000359591.9_3	ENSG00000197324.9_8	LRP10-201	LRP10	6892	protein_coding	3.57E-04
ENST00000415496.5_1	ENSG00000102858.13_12	MGRN1-203	MGRN1	6405	protein_coding	3.57E-04
ENST00000445623.1_1	ENSG00000102265.12_5	TIMP1-204	TIMP1	595	protein_coding	3.57E-04
ENST00000466194.1_1	ENSG00000198830.11_5	HMGN2-205	HMGN2	952	processed_transcript	3.57E-04

ENST00000486685.3_2	ENSG00000158457.6_5	TSPAN33-201	TSPAN33	2774	protein_coding	3.57E-04
ENST00000526180.6_5	ENSG00000109846.9_9	CRYAB-205	CRYAB	977	protein_coding	3.57E-04
ENST00000529140.1_1	ENSG00000052841.15_6	TTC17-209	TTC17	1017	retained_intron	3.57E-04
ENST00000548547.5_1	ENSG00000167779.9_8	IGFBP6-203	IGFBP6	1177	protein_coding	3.57E-04
ENST00000576008.5_3	ENSG00000134419.15_7	RPS15A-212	RPS15A	589	nonsense_mediated_decay	3.57E-04
ENST00000592588.7_4	ENSG00000115268.10_6	RPS15-209	RPS15	501	protein_coding	3.57E-04
ENST00000607003.5_6	ENSG00000150459.12_9	SAP18-207	SAP18	2234	protein_coding	3.57E-04
ENST00000615901.4_2	ENSG00000115306.16_8	SPTBN1-207	SPTBN1	10226	protein_coding	3.57E-04

Table S4: Diversity determination of the fragmented CDS human prostate library via counting of colony forming units (cfu) after transformation of the cloned plasmid pool. Dilutions of 1:1,000 and 1:10,000 were plated on agar plates.

	1:1,000	1:10,000	Mean
cfu	370	33	350,000 ± 20,000

Table S5: Sanger sequencing results of the cDNA inserts of 14 single clones harboring the fragmented CDS prostate cDNA display library. Nucleic acid sequences were subjected to nucleotide blast search against the human transcriptome and proteome and the theor. displayed amino acid sequence was determined by in silico translation by the virtual ribosome 2.0 suite.[384]

Clone	Insert length / bp	Transcript (blastN)	Theor. displayed amino acid sequence	Number of displayed residues
1	198	Homo sapiens SCG10 like-protein, helicase-like protein NHL, M68, and ADP-ribosylation factor related protein 1	IPRRRHGFTTVTTTSLTGPPGDPAAPGLVSAPRDRVSRPAAPAGEAVVSDSPRGQGRAAECGREA	65
2	301	Homo sapiens spectrin alpha, non-erythrocytic 1 (SPTAN1), transcript variant 9, mRNA	IPRRHHGQVIAQGAQARFELLMVQPARLVLVKVFKHHAEFFEGLLSHTCCVPGLDLLLQVVLHAHAQLVQLVPLLGEAHGAVLRVLVVQDEGLLHGGGGG	100
3	257	Homo sapiens TLE family member 2, transcriptional corepressor (TLE2), transcript variant 4, mRNA	XTMATXSLLLHETLRGRGRCTPARGSLHSGPLGLHXRTVLLDSRXQLXLGSQGSRQSKXPRSXAXHQPGWAGGEGHRGCVVGQRL	85
4	287	No alignment	IPRRHHGHLNRLRPRRLGP	19

5	287	No alignment	IPRRHHGHLNRLRPRRLGP	19
6	437	Homo sapiens spectrin alpha, non-erythrocytic 1 (SPTAN1), transcript variant 9, mRNA	IXRRXHGCVEVRGQGSTTQEDLQLVIGSEVDGGSWHSHXHGGGGGYPPPPPWRRPSSWTTSTRSTAPWASPSSGTSWTSWACACSTTWSSRSRPGTQQV	99
7	339	Homo sapiens isolate CHM13 chromosome 22	SPSPPK	6
8	301	Homo sapiens spectrin alpha, non-erythrocytic 1 (SPTAN1), transcript variant 9, mRNA	IPRRHHGQVIAQGAQARFELLMVQPARLVLVKVFKHHAEFFEGLLSHTCCVPGLDLLLQVVLHAHAQLVQLVPXLGEAHGAVLRVLVVQDEGLLHGGGGG	100
9	301	Homo sapiens spectrin alpha, non-erythrocytic 1 (SPTAN1), transcript variant 9, mRNA	IPRRRHGGGXHXGQQVHGAQHRGPRPAVGPXGPXGHAHAAQPGAADPGQEHNRXD	55
10	776	Homo sapiens cDNA FLJ61693 complete cds, highly similar to Transducin-like enhancer protein 2	IPRRHHGETPTPRSLAPTQEVLKLGWCVGLGSSGPGNPRGTCCA	44
11	406	Homo sapiens pyruvate carboxylase (PC), transcript variant 2, mRNA; nuclear gene for mitochondrial product	IPRRHHGEQGPCGVPRPRCEDQHRLPAECAQQPAVPGRHCGHPVHRREPRAVPAAACTEPGPKAVALPRPCHGKRSNHPDSRQGQPQPHGPRCPCSAHRPAPGWFQRHPAARGA	114
12	NA	-	-	-

13	1182	Homo sapiens cDNA FLJ52887 complete cds, highly similar to Preimplantation protein 3		0
14	136	Homo sapiens solute carrier family 25 member 39 (SLC25A39), transcript variant 1, mRNA; nuclear gene for mitochondrial product	IPRRHHGDDCASYRHLLHCL	20
average	400±269			56±39

Table S6: Composition of the fragmented human prostate cDNA library. ORFs of human proteins in-frame with the intimin sequence. HGNC gene symbols, total read counts and fraction of the library as detected via proteome mapping of in silico translated Illumina reads. Only the 300 most abundant proteins are shown.

HGNC symbol	Total counts	Library fraction in %	HGNC symbol	Total counts	Library fraction in %
STMN3	235553	31.473749	SNRNP40	10	0.00133616
SPTAN1	177328	23.6939329	ZFYVE28	10	0.00133616
TRABD	165244	22.079312	AGRN	9	0.00120255
TLE2	67610	9.03380629	DENND4B	9	0.00120255
SLC25A39	34189	4.56821185	GRK6	9	0.00120255
GSN	9452	1.26294242	LRP2	9	0.00120255
SLC25A40	8534	1.14028255	UBXN2B	9	0.00120255
CIRBP	5684	0.75947574	NA	8	0.00106893
NOP53	5638	0.75332939	TUBB4A	8	0.00106893
MYC	4100	0.54782733	ZNF513	8	0.00106893
TTC38	3410	0.455632	FZD1	8	0.00106893
TLE1	2730	0.36477283	MYO1F	7	9.35E-04
FOXK1	2167	0.28954679	CRABP2	7	9.35E-04
RBM3	2160	0.28861147	PDSS1	7	9.35E-04
STMN2	1677	0.22407474	CORO6	7	9.35E-04
DEGS2	1557	0.20804077	BRSK1	7	9.35E-04
PGD	1510	0.2017608	PPP1R16B	7	9.35E-04
DTD1	1291	0.1724988	SLC39A8	7	9.35E-04
GPAA1	1057	0.14123256	MED14	6	8.02E-04
ABHD8	955	0.12760368	CALML3	6	8.02E-04
ALDH3A1	937	0.12519859	DNM1	6	8.02E-04
PPP2CB	718	0.09593659	ABR	6	8.02E-04
MYPN	655	0.08751876	ALS2CL	6	8.02E-04
PEAK1	588	0.07856646	ANKRD53	6	8.02E-04
STMN4	535	0.07148479	MSLNL	6	8.02E-04
WDFY4	488	0.06520481	DIMT1	6	8.02E-04
GPX3	478	0.06386865	USH2A	5	6.68E-04

CKM	413	0.05518358
KLK3	391	0.05224402
ATP10A	382	0.05104147
PODXL	371	0.04957169
PCBP1	330	0.04409342
TPM1	308	0.04115386
ADGRG1	298	0.03981769
CCNO	296	0.03955046
MOV10L1	268	0.0358092
RIN1	265	0.03540835
IER2	256	0.0342058
CD151	238	0.03180071
PLEKHM2	234	0.03126624
PC	204	0.02725775
MCMBP	203	0.02712413
MAP3K15	198	0.02645605
SURF4	195	0.0260552
FUNDC2	180	0.02405096
NA	151	0.02017608
CKB	150	0.02004246
EEF1A1	149	0.01990885
TPM3	144	0.01924076
PNPLA7	140	0.0187063
TTN	133	0.01777098
BCKDHA	130	0.01737013
CHMP6	130	0.01737013
PPP2CA	129	0.01723652
TSPO	127	0.01696929
BCL10	123	0.01643482
MARK4	121	0.01616759
ZNF837	102	0.01362888
KRT78	100	0.01336164
ZDHHC4	99	0.01322803
SPTA1	5	6.68E-04
MT-ND4	5	6.68E-04
NA	5	6.68E-04
TRIM23	5	6.68E-04
HTT	5	6.68E-04
NUBP1	5	6.68E-04
RAD23A	5	6.68E-04
NA	5	6.68E-04
STIM1	5	6.68E-04
ITIH4	5	6.68E-04
DENND2C	5	6.68E-04
CEP85	5	6.68E-04
BEND2	5	6.68E-04
SORCS1	5	6.68E-04
GAL3ST3	5	6.68E-04
CREB3L1	5	6.68E-04
WNK3	5	6.68E-04
PLEKHG2	5	6.68E-04
ANKEF1	5	6.68E-04
BIN2	5	6.68E-04
STK17A	5	6.68E-04
RNF150	5	6.68E-04
DNM3	5	6.68E-04
RIMS2	5	6.68E-04
MAPK8IP1	5	6.68E-04
TACC3	5	6.68E-04
CAMTA1	5	6.68E-04
PRSS51	4	5.34E-04
TCERG1	4	5.34E-04
ZMYM6	4	5.34E-04
IGHA2	4	5.34E-04
TUBB	4	5.34E-04
PDIA4	4	5.34E-04

DHX38	98	0.01309441
TRIM56	91	0.01215909
DNAJB9	91	0.01215909
CTSD	89	0.01189186
MAL2	88	0.01175825
ABTB1	82	0.01095655
IFI16	79	0.0105557
EIF4G3	77	0.01028846
SRRT	77	0.01028846
BAG6	75	0.01002123
HNRNPA0	73	0.009754
STMN1	71	0.00948677
KIAA0408	71	0.00948677
ESYT2	68	0.00908592
TRIM7	67	0.0089523
SSBP3	65	0.00868507
TMPRSS2	63	0.00841783
DSPP	62	0.00828422
DEGS1	60	0.00801699
CELSR1	55	0.0073489
TTLL6	54	0.00721529
HERPUD1	53	0.00708167
WDR11	52	0.00694805
RCOR3	52	0.00694805
ISYNA1	51	0.00681444
PRDM8	51	0.00681444
SLC25A13	49	0.0065472
CALM1	46	0.00614636
E4F1	46	0.00614636
NIPBL	46	0.00614636
NLRX1	45	0.00601274
NCOR2	45	0.00601274
MGA	43	0.00574551
MYH11	4	5.34E-04
RPL3	4	5.34E-04
SMARCA2	4	5.34E-04
PCBP3	4	5.34E-04
MST1R	4	5.34E-04
EEF1A2	4	5.34E-04
RTN1	4	5.34E-04
KHDRBS2	4	5.34E-04
MFSD5	4	5.34E-04
ACAD11	4	5.34E-04
TRPM8	4	5.34E-04
NOP9	4	5.34E-04
CMYA5	4	5.34E-04
NA	4	5.34E-04
GOLGA6L1	4	5.34E-04
TTC9	4	5.34E-04
C19orf44	4	5.34E-04
ETNK2	4	5.34E-04
MBD2	4	5.34E-04
GABBR1	4	5.34E-04
UST	4	5.34E-04
PCDHB11	4	5.34E-04
N6AMT1	4	5.34E-04
DDTL	3	4.01E-04
ZNF726	3	4.01E-04
C4orf54	3	4.01E-04
TNFRSF10B	3	4.01E-04
PPM1G	3	4.01E-04
SIPA1L1	3	4.01E-04
CACNA1G	3	4.01E-04
MYO1B	3	4.01E-04
MBD3	3	4.01E-04
EDN1	3	4.01E-04

ATP6V0B	43	0.00574551
HDAC3	41	0.00547827
ABCF1	41	0.00547827
SLC2A1	40	0.00534466
PYHIN1	40	0.00534466
NA	35	0.00467657
DNAJB4	35	0.00467657
ASB4	35	0.00467657
DLGAP3	34	0.00454296
GAPDH	34	0.00454296
TPM2	32	0.00427573
C16orf89	32	0.00427573
NA	30	0.00400849
HNRNPDL	30	0.00400849
CPLX4	28	0.00374126
UCP3	27	0.00360764
MYO9B	27	0.00360764
PDE8A	26	0.00347403
KLHL41	26	0.00347403
VARS1	25	0.00334041
RPL4	25	0.00334041
DZANK1	25	0.00334041
GTPBP1	23	0.00307318
ADRA2B	22	0.00293956
HIBADH	22	0.00293956
CCND1	21	0.00280594
ZNF76	20	0.00267233
FKBP1A	20	0.00267233
CELSR3	20	0.00267233
SERINC2	19	0.00253871
SLC22A23	18	0.0024051
NCKAP1L	17	0.00227148
ALDH6A1	17	0.00227148
NA	3	4.01E-04
PSG2	3	4.01E-04
SRC	3	4.01E-04
B4GALT1	3	4.01E-04
KLK2	3	4.01E-04
CAD	3	4.01E-04
COIL	3	4.01E-04
CLK3	3	4.01E-04
CDH15	3	4.01E-04
CLDN2	3	4.01E-04
GNAS	3	4.01E-04
NA	3	4.01E-04
TAP1	3	4.01E-04
BPTF	3	4.01E-04
TCIRG1	3	4.01E-04
SCN5A	3	4.01E-04
ITPR1	3	4.01E-04
HYDIN	3	4.01E-04
PABPC1L	3	4.01E-04
XKR4	3	4.01E-04
ARFGEF3	3	4.01E-04
B4GALNT3	3	4.01E-04
ZNF280D	3	4.01E-04
PKN3	3	4.01E-04
ZNF574	3	4.01E-04
FGD5	3	4.01E-04
TMPRSS7	3	4.01E-04
TMC3	3	4.01E-04
FRAS1	3	4.01E-04
TTLL11	3	4.01E-04
MYCBPAP	3	4.01E-04
PREX1	3	4.01E-04
ADAMTS16	3	4.01E-04

SLC22A17	17	0.00227148
PKP2	17	0.00227148
HCN2	16	0.00213786
ZNF318	15	0.00200425
RHOT2	15	0.00200425
NA	14	0.00187063
TUBB4B	14	0.00187063
TCF25	14	0.00187063
NFE2L3	14	0.00187063
CD2	13	0.00173701
FTH1	12	0.0016034
TLE3	12	0.0016034
ESPL1	12	0.0016034
CROCC	12	0.0016034
TWNK	12	0.0016034
C14orf93	12	0.0016034
CAPZB	11	0.00146978
TUT4	11	0.00146978
USP51	11	0.00146978
SIPA1L2	11	0.00146978
SRSF10	10	0.00133616
CACNA1H	10	0.00133616
CKMT1A	10	0.00133616
CKMT1B	10	0.00133616
SNX33	3	4.01E-04
SLC9A6	3	4.01E-04
BRF1	3	4.01E-04
MTARC2	3	4.01E-04
ARHGEF26	3	4.01E-04
SGSM3	3	4.01E-04
COG3	3	4.01E-04
NA	3	4.01E-04
RCHY1	3	4.01E-04
PIK3R4	3	4.01E-04
OSMR	3	4.01E-04
TXN2	3	4.01E-04
TIMM29	3	4.01E-04
SPON2	3	4.01E-04
KLHL4	3	4.01E-04
TRPV6	3	4.01E-04
RAB3GAP2	3	4.01E-04
TTC12	3	4.01E-04
ZDBF2	3	4.01E-04
IFT46	3	4.01E-04
BIRC6	3	4.01E-04
ENAM	3	4.01E-04
LRRC4B	3	4.01E-04
ZBTB47	3	4.01E-04

Table S7: Raw values of survival rate determination of *E. coli* BL21(DE3) Tuner cells harboring the display libraries. The fraction of viable cells was determined via colony counting after O/N incubation of FACS sorted regular-shaped, single cells.

Library	c(arab) / %	Replicate	Sorted cells	Counted colonies	Survival fraction	Average fraction
Fragm. prostate	0	1	286	269	0.941	0.956±0.005
		2	287	273	0.951	
	0.05	1	288	249	0.865	0.870±0.005
		2	288	252	0.875	
Full-length thyroid	0	1	378	282	0.746	0.770±0.024
		2	199	158	0.794	
	0.05	1	384	293	0.763	0.758±0.005
		2	384	289	0.753	

Table S8: Mapping of in silico translated Illumina sequence reads of the Myc-tag-enriched fragmented human prostate CDS display library against the human proteome. Total reads, and mapped and quality-passed reads, as well as their fraction throughout the ORF enrichment process.

Library	Total reads	Mapped reads	Fraction of proteome-mapped peptides
Parent	13,735,080	748,411	5.45%
1st sorting	756,711	166,364	21.99%
2nd sorting	782,014	150,097	19.19%
3rd sorting	800,632	227,938	28.47%

Table S9: Proteome-mapped peptides of the fragmented human prostate CDS display library after three rounds of Myc-tag enrichment for ORFs. Values represent the sum or the mean per HGNC gene locus (gene symbol).

HGNC symbol	Mean subject start	SD$_{\text{subject start}}$	Mean subject end	SD$_{\text{subject end}}$	Mean e-value	SD$_{\text{e-value}}$	Mean Alignment length	SD$_{\text{alignment length}}$	Total counts
CIRBP	1	0	25.9915254	6.09959472	4.37E-05	4.60E-04	25.9915254	6.09959472	118
TPM1	1	0	21	0	3.83E-10	0	21	0	1
RBM3	1.01111111	0.1490712	39.9	1.34164079	4.79E-13	6.42E-12	39.8888889	1.34906916	180
BCL10	1.48387097	0.62046877	18.8870968	0.48264431	3.63E-04	0.00134904	18.4032258	0.79876293	62
NOP53	2	0	17.1410256	2.80872647	0.00569709	0.00280915	16.2307692	3.0913498	78
UBXN2B	4	0	21	0	0.002	0	20	0	1
TRABD	4.67577231	28.3474087	32.0890501	28.71094	1.30E-06	8.34E-05	28.4134578	8.6842874	77754
TMEM74	17	0	39	0	0.008	0	24	0	1
PDE8A	18	0	35	0	0.008	0	18	0	1
IFT74	23	0	36	0	0.002	0	21	0	1
DPP6	23	0	41	0	0.007	0	28	0	2
ADRA2C	27	0	44	0	0.009	0	26	0	1
NEB	29	0	46	0	0.009	0	18	0	1
NA	30	0	54	0	1.55E-04	0	25	0	1
ARPP21	41	0	55	0	0.003	0.00244949	15	0	6
FBXO46	41	0	63.5	0.70710678	0.007	0.00282843	27.5	0.70710678	2
STMN1	41.6666667	8.03326418	65.952381	8.1576724	2.49E-05	1.14E-04	25.4761905	1.28914885	21
FBXO31	51	0	65	0	0.008	0	15	0	1
TRIM7	61	0	86	0	0.005	0	29	0	1
ONECUT2	65	0	87.5	3.53553391	0.007	0.00424264	23.5	3.53553391	2
CCNO	67	0	88	1.73205081	0.0042	0.00432666	22	1.73205081	3
ZNF318	67	0	89.5	0.70710678	5.36E-04	6.57E-04	24	1.41421356	2
STMN3	71.5123001	20.462102	110.63286	23.0876718	4.72E-06	9.60E-05	40.1795234	6.88032501	31138
TLE7	81	0	99.1666667	0.40824829	0.00783333	0.00348807	19.1666667	0.40824829	6
STMN2	85.0270936	10.267743	109.812808	10.2261736	5.55E-08	1.93E-06	26.1863711	1.09929901	1218
KDM2A	86	0	108	0	0.006	0	23	0	1
PPP2CB	88	0	130	0	1.07E-22	0	43	0	1
STMN4	90.2049123	0.91881305	114.897544	1.79829847	6.73E-06	2.38E-04	25.6926316	1.74041771	1425

ZNF195	97	0	119	0	2.99E-07	0	23	0	1
MT-CO1	100	82.0243866	118.5	84.145707	3.59E-07	5.05E-07	19.5	2.12132034	2
PODXL	110	0	124.428571	3.25868802	0.00414286	0.00323669	18.4285714	3.25868802	7
SLC25A40	110.088235	0.28790224	130.088235	0.3788057	2.13E-08	3.26E-08	21	0.42640143	34
FGFBP1	113	0	128	0	0.007	0	18	0	1
TOP3A	125	0	147	0	0.002	0	23	0	1
MAGT1	134	0	147	0	0.004	0	15	0	1
USP17L8	142	0	191	0	8.66E-47	0	50	0	1
TLE4	145	0	180	0	4.13E-05	0	36	0	1
SLC25A39	147.63078	9.15879959	165.730159	4.58585172	5.56E-04	4.79E-04	19.1042098	7.5877584	1449
MBD2	151	0	176	0	4.02E-19	8.20E-19	26	0	11
ASPDH	151	0	170	0	4.09E-06	0	20	0	1
TLE1	155.535211	5.66292079	181.48169	2.35460301	2.24E-04	7.41E-04	29.3070423	2.67734395	355
TLE2	156.131377	11.8266584	197.886598	14.4533011	2.52E-06	1.11E-04	42.7575998	7.28833451	7566
CREB3L1	158	0	173	0	0.001	0	16	0	1
EIF4G3	160	0	184	7.21110255	0.00733333	0.00251661	29.3333333	5.50757055	3
DZANK1	161	0	178	0	0.01	0	22	0	1
PYHIN1	166	0	189	0	0.006	0	26	0	1
WFIKKN2	184	0	198	0	3.60E-04	0	15	0	1
MICAL1	186	0	204.6	0.89442719	0.0064	0.00350714	19.6	0.89442719	5
BRD3	187	0	206	0	0.007	0	20	0	1
DLGAP3	194	0	221	0	0.007	0	38	0	1
NA	195	0	244	0	7.19E-25	0	50	0	2
MT-CO3	202.5	14.8492424	239	0	1.66E-15	2.34E-15	37.5	14.8492424	2
SLC4A10	219	0	233	0	0.009	0	18	0	1
RSPH1	220	0	241	0	0.004	0	22	0	1
LRATD2	230	0	247	0	0.002	0	23	0	1
MT-ND1	264	0	283	0	2.24E-05	0	20	0	1
MT-ND5	267.5	30.4055916	284	31.1126984	5.30E-04	6.65E-04	17.5	0.70710678	2
RABL6	271	0	300	0	6.49E-12	0	30	0	1
RECK	272	0	287	0	0.009	0	18	0	1
MCAM	276	0	290	0	0.01	0	15	0	1
GLYR1	284	0	302	0	0.006	0	22	0	1

ERI1	286	0	303	0	0.004	0	21	0	1
CHAF1A	301	0	334	0	0.002	0	34	0	1
TTC38	303	0	323	0	0.008	0	21	0	2
DENND2C	333	0	355	0	0.007	0	23	0	2
NLRX1	333	0	358	0	0.004	0	26	0	1
ASNSD1	360	0	377	0	0.002	0	18	0	1
MYC	423.000843	0.0290173	438.004875	0.12134132	1.06E-04	2.72E-04	19.0022872	0.09307351	16614
GSN	424.043295	0.96636181	446.626413	3.87377214	3.86E-06	1.68E-04	23.5842069	3.65648035	7345
B4GALNT3	436	0	454	0	0.007	0	20	0	1
MTMR1	441	0	460	0	4.02E-15	0	20	0	1
MARK4	463	0	493.666667	2.30940108	2.31E-04	3.05E-04	31.6666667	2.30940108	3
SERAC1	468	0	501	0	5.59E-29	0	34	0	1
GLP2R	507	0	526	0	0.01	0	20	0	1
NAV2	509	0	525	0	3.45E-04	0	17	0	1
CPEB4	511	0	556	0	4.03E-37	0	46	0	1
FOXK1	529.030928	0.20121951	544.041237	0.40508493	5.01E-04	7.74E-04	16.0103093	0.35248475	194
PIK3R5	573	0	590	0	0.007	0	24	0	1
DPP8	585	0	598	0	0.008	0	16	0	1
ZC3H3	585	0	600	0	0.006	0	16	0	1
CELSR1	613	0	629	0	0.0085	7.07E-04	22	0	2
MCMBP	616	0	631	0	0.0062	0.00168655	16	0	10
RNF123	647	0	672	0	0.005	0	26	0	1
KIAA0408	654	0	678	0	0.004	0	25	0	1
ZAN	744.5	14.8492424	756.5	14.8492424	0.009	0	15	0	2
ANO9	755	0	774	0	0.01	0	25	0	1
DNM1	810	0	836.734694	1.56492159	0.00462038	0.00302582	28.1020408	2.45157176	49
NA	810.666667	417.662942	846	405.4479	6.77E-06	1.17E-05	36.3333333	16.2583312	3
MYPN	841.666667	0.57735027	859.333333	0.57735027	0.00466667	0.0046188	18.6666667	0.57735027	3
SPECC1L	858	0	877	0	6.97E-04	0	21	0	1
CLASP2	897	0	916	0	0.009	0	21	0	1
PEAK1	1147	0	1161	0	0.0085	0.00212132	15	0	2
TNKS2	1147	0	1166	0	0.01	0	20	0	1
BAZ1B	1191	0	1213	0	0.003	0	23	0	1

MYO18B	1217	0	1232	0	0.003	0	16	0	1
SLIT2	1344	0	1366	0	0.006	0	23	0	1
TAF1L	1565	0	1581	0	0.009	0	17	0	1
NCOR2	1739.6	6.40230222	1767.95	0.2236068	0.00224564	0.00226084	29.35	6.3765607	20
DOCK4	1787	0	1801.75	0.5	0.0095	5.77E-04	15.75	0.5	4
MIA3	1857	0	1877	0	0.001	0	21	0	1
MYO9B	2030	0	2063	0	0.007	0	35	0	1
DNAH8	2128	0	2150	0	0.01	0	24	0	1
SPTA1	2253	0	2270	0	2.53E-04	0	18	0	1
SPTAN1	2272.23605	10.561539	2302.88568	11.927478	5.02E-06	1.03E-04	31.6512706	8.67196967	82164
TTN	21099	0	21112	0	0.009	0.0011547	23	0	4

Table S10. Enrichment and depletion for general DNA binding properties. EF of proteins per round and continuously monitored EF. The peptide counts were summed for each protein, normalized on the read number and these fractions were used to calculate the EF between two sub libraries.

	EF per round			**Continuous EF**		
HGNC symbol	**EF1**	**EF2**	**EF3**	**1 round**	**2 rounds**	**3 rounds**
FUNDC2	6.02789786	1.25677054	NA	6.02789786	7.57568445	NA
MARK4	3.4481284	0.53936402	NA	3.4481284	1.85979641	NA
ZNF318	2.39213908	NA	NA	2.39213908	NA	NA
TRIM56	2.1494583	NA	NA	2.1494583	NA	NA
NCOR2	2.01443291	NA	NA	2.01443291	NA	NA
CKM	1.91817213	NA	NA	1.91817213	NA	NA
MOV10L1	1.78278779	0.87646654	NA	1.78278779	1.56255384	NA
STMN1	1.59475939	NA	NA	1.59475939	NA	NA
ZNF837	1.57684074	NA	NA	1.57684074	NA	NA
NFE2L3	0.91129108	NA	NA	0.91129108	NA	NA
PC	0.70878195	NA	NA	0.70878195	NA	NA
CCNO	0.69337365	NA	NA	0.69337365	NA	NA
ZFYVE28	0.63790376	NA	NA	0.63790376	NA	NA
KIAA0408	0.39868985	NA	NA	0.39868985	NA	NA
ATP10A	0.35502722	NA	NA	0.35502722	NA	NA
DTD1	0.33382214	NA	NA	0.33382214	NA	NA
TLE1	0.22389479	NA	NA	0.22389479	NA	NA
GPX3	0.21590589	NA	NA	0.21590589	NA	NA
ALDH3A1	0.19164491	NA	NA	0.19164491	NA	NA
IER2	0.16612077	NA	NA	0.16612077	NA	NA
PGD	0.15341363	NA	NA	0.15341363	NA	NA
GPAA1	0.1357242	NA	NA	0.1357242	NA	NA
TTC38	0.10180398	NA	NA	0.10180398	NA	NA
TCF25	23.9213908	4.02725137	2.28645572	23.9213908	96.3374539	220.271323
DHX38	13.6693662	2.58894731	3.25184814	13.6693662	35.3892688	115.080528
CHMP6	12.917551	1.86446823	1.7333512	12.917551	24.0843635	41.7466603
HNRNPA0	6.37903755	2.3372441	2.25127948	6.37903755	14.9093679	33.565154

ESYT2	12.7580751	1.6275546	1.40377747	12.7580751	20.7644638	29.1486864
RCOR3	4.04005711	1.13550321	2.92666333	4.04005711	4.58749781	13.4260616
HNRNPDL	8.42032956	0.98066186	1.28041521	8.42032956	8.25749605	10.5730235
TPM1	1.45806573	NA	10.5176963	1.45806573	NA	10.5176963
PYHIN1	2.8186445	3.40650962	1.0365266	2.8186445	9.60173959	9.95245845
RBM3	2.64078436	3.10075941	0.86866952	2.64078436	8.18843694	7.11304562
EEF1A1	4.05580782	2.72300284	0.54874937	4.05580782	11.0439762	6.06037502
MYPN	2.96254927	1.39190716	1.45189939	2.96254927	4.12359353	5.98704291
CKB	0.51721926	4.31491219	1.82916458	0.51721926	2.23175569	4.08224846
CD2	5.21921254	NA	3.65832916	5.21921254	NA	3.65832916
STMN2	1.41617956	1.31598495	1.93469331	1.41617956	1.86367098	3.60563178
BCKDHA	1.89647062	1.76519135	1.01620254	1.89647062	3.34763353	3.40187371
STMN3	1.07398628	1.22375352	1.27068516	1.07398628	1.31429449	1.6700545
CIRBP	0.68449402	2.57486601	0.9395904	0.68449402	1.7624804	1.65600966
STMN4	0.85475937	0.63924625	2.40077851	0.85475937	0.54640172	1.31178952
FOXK1	0.49765541	1.02969495	2.52598918	0.49765541	0.51243327	1.29440089
SPTAN1	0.9854262	1.09907751	1.08288606	0.9854262	1.08305977	1.17283033
PPP2CB	0.63254322	1.38954799	1.29966957	0.63254322	0.87894916	1.14234348
MYC	0.91129108	1.34841006	0.91458229	0.91129108	1.22879406	1.12383328
PEAK1	1.63338678	0.75802511	0.77387732	1.63338678	1.2381482	0.95817481
SLC25A39	0.8508984	1.04449259	1.01923056	0.8508984	0.88875708	0.90584838
SLC25A40	0.68559749	1.03590519	1.13782679	0.68559749	0.710214	0.80810052
TRABD	0.64735618	0.90474546	1.27997848	0.64735618	0.58569257	0.74967388
ABHD8	0.61552117	0.94143539	1.21944305	0.61552117	0.57947341	0.70663482
GSN	0.58279191	0.6488002	1.5206308	0.58279191	0.37811551	0.57497409
DEGS2	0.76375525	0.56989406	1.13233998	0.76375525	0.43525958	0.49286182
NOP53	0.36637883	0.95298674	0.8383671	0.36637883	0.34915417	0.29271937
TLE2	0.25900764	0.29480767	0.33897106	0.25900764	0.07635744	0.02588296

Table S11: Enrichment and depletion for general DNA binding properties (continuation of Table S10). Values represent the mean per HGNC gene locus (gene symbol). The peptide counts represent the counts in the library after the final enrichment step. *Protein not detected in the final library; the other values are derived from latest selection step in which the peptides of this protein were detected.

HGNC symbol	Mean subject start	$SD_{subject\ start}$	Mean subject end	$SD_{subject\ end}$	Mean e-value	$SD_{e\text{-}value}$	Mean Alignment length	$SD_{alignment\ length}$	Total counts
FUNDC2*	41	0	81.4368932	10.8724061	5.35E-09	5.43E-08	41.4466019	10.8805711	0*
MARK4	466.125	4.43290436	500.225	9.63829834	3.38E-04	9.94E-04	35.15	6.77684595	40
ZNF318	68.5	1.04880885	90.8333333	5.0365332	0.00566667	0.00258199	24.1666667	3.65604522	6
TRIM56	730	0	753.387097	2.23125405	5.61E-08	3.13E-07	24.3870968	2.23125405	31
NCOR2	1747	5.40370243	1780.5	6.12372436	1.02E-04	1.72E-04	34.5	0.83666003	6
CKM	74.5116279	25.2813552	117.523256	28.4748503	2.73E-17	1.64E-16	44.0116279	5.46538667	86
MOV10L1	965	0	1007.8125	4.63552535	1.85E-22	9.95E-22	43.8125	4.63552535	64
STMN1	42.2	5.76965241	66.8	6.01479657	5.94E-07	1.87E-06	25.7	0.48304589	10
ZNF837	105.818182	0.39477102	146.272727	7.47810657	2.97E-09	1.21E-08	41.4545455	7.08941043	22
NFE2L3	145	0	166.6	5.17687164	0.0052	0.00216795	23.2	4.96990946	5
PC	278.055556	0.23570226	320.944444	4.47834295	3.33E-16	1.41E-15	43.8888889	4.71404521	18
CCNO	67.12	0.6	86.88	1.50886271	0.00411516	0.00263771	20.76	1.69016765	25
ZFYVE28	371.857143	0.53452248	387	0	0.00635714	0.00115073	16.2142857	0.80178373	14
KIAA0408	654	0	677	1.73205081	0.008	0.00122475	24	1.73205081	5
ATP10A	1462	0	1476	0	0.00709677	5.39E-04	17	0	31
DTD1	105.818182	60.3757036	146.2	64.185899	7.27E-05	5.39E-04	41.3818182	8.34996321	55
TLE1	156.070175	6.26743183	180.561404	3.28968922	7.47E-04	0.00192553	27.4736842	4.1192287	57
GPX3	17	0	36.1818182	2.48266719	3.65E-13	4.39E-13	20.1818182	2.48266719	11
ALDH3A1	311.904762	2.30010352	340.619048	2.83683257	3.63E-06	7.60E-06	29.7142857	2.83095138	21
IER2	1	0	32.4	10.4067286	7.00E-17	7.27E-17	32.4	10.4067286	5
PGD	366.129032	53.6723032	408.290323	53.6502212	1.73E-21	9.35E-21	43.1612903	5.71020592	31
GPAA1	1	0	36.2	8.49705831	6.12E-07	1.37E-06	36.2	8.49705831	5
TTC38	303	0	323	0	0.00494	7.40E-04	21	0	50
TCF25	513.657143	0.48159399	542.285714	8.9264031	7.29E-14	4.24E-13	29.6285714	9.40302115	35
DHX38	1178.02679	0.28283705	1210.99554	0.06681531	1.89E-12	1.70E-12	33.96875	0.29020829	224
CHMP6	1.01005025	0.09999746	32.9145729	7.80881001	2.31E-06	2.66E-05	32.9045226	7.81515813	199
HNRNPA0	63.03125	0.47412908	92.125	9.82344135	1.01E-05	5.62E-05	30.09375	9.77607143	32

ESYT2	76.2575758	1.26868322	101.969697	0.52534815	0.00237879	0.00168034	32.7272727	1.14415511	66
RCOR3	424.125	0.5	446.125	0.5	0.00596656	0.00287075	26.1875	1.64189931	16
HNRNPDL	17	0	36	0	0.00407862	0.00248165	32	0	21
TPM1	1.08695652	0.28810407	34.2608696	8.15875097	3.06E-04	0.00145927	34.173913	8.31565978	23
PYHIN1	166	0.35355339	181.352941	3.48104109	0.00611765	0.00321874	18.4705882	3.8909775	17
RBM3	1.02114165	0.20475587	39.6997886	2.41875177	6.67E-08	1.31E-06	39.6786469	2.56076754	473
EEF1A1	360	0	404	0	3.19E-40	2.76E-40	45	0	39
MYPN	841.755906	0.62635929	862.007874	4.85748655	0.00488189	0.00256538	21.2598425	4.74126197	127
CKB	1.05555556	0.23570226	31.5555556	9.71488986	9.98E-13	3.56E-12	31.5	9.781435	18
CD2	281.833333	0.38924947	301.25	4.33012702	0.00575	0.00148477	20.4166667	4.29499356	12
STMN2	78.0772727	7.40100153	102.586364	7.89409325	4.10E-09	2.17E-08	25.9	1.35787909	220
BCKDHA	143	0	182	0	5.92E-20	0	40	0	10
STMN3	60.980898	40.5989484	99.9823865	42.4471761	5.98E-06	1.75E-04	40.0052096	7.93595494	12093
CIRBP	1	0	25.7110266	8.05989933	5.54E-05	4.57E-04	25.7110266	8.05989933	263
STMN4	92.1428571	1.38873015	117.142857	1.38873015	4.65E-10	1.31E-09	26	0	21
FOXK1	529.172414	0.50045352	544.844828	2.15049341	7.51E-04	0.00162941	16.8448276	2.88858445	58
SPTAN1	2281.00286	22.7773752	2316.29243	24.8386708	1.15E-05	2.15E-04	36.3002064	8.91832769	6299
PPP2CB	6.18518519	17.7548196	36.2592593	20.7486656	3.08E-16	1.37E-15	31.0740741	6.77623634	27
MYC	423	0	439.2	3.27108545	0.0021974	0.00436589	21.4	3.28633535	5
PEAK1	1147	0	1161.09091	0.30151135	0.00437166	0.00248561	15.0909091	0.30151135	11
SLC25A39	128.721925	2.94774267	161.28984	7.43569531	4.51E-05	4.27E-04	33.597861	7.99362673	935
SLC25A40	109.962025	0.73913425	129.886076	0.66748586	1.68E-06	1.39E-05	20.9746835	0.85174506	158
TRABD	109.420529	116.044599	142.197896	117.76762	2.32E-06	7.82E-05	33.7847597	10.3163839	3517
ABHD8	52	0	79	0	2.45E-10	0	28	0	16
GSN	424.289855	1.25100193	458.550725	8.27014017	1.42E-04	9.44E-04	35.2608696	8.49414083	138
DEGS2	95.2307692	28.8420633	127.576923	29.5853654	1.10E-06	3.89E-06	33.6538462	7.96463336	26
NOP53	2	0	18.3939394	3.3160537	0.00459	0.00290227	17.3939394	3.3160537	33
TLE2	155.471698	5.14259652	187.754717	14.4033993	4.44E-11	1.45E-10	33.9622642	14.207206	53

Table S12: *DP-bind*[309] and *DRNApred*[310] binding prediction of the residues of the proteins enriched for dsDNA binding. Patches of at least 4 binding residues are depicted and combined when connect via small gaps (max. 2 residues). Residues covered by the fragment are highlighted in bold, partially or potentially covered residues are highlighted in oblique.

Protein	Covered residues	Pot. dsDNA interaction patch (*DP-bind*[309])	Pot. nucleic acid interaction patch (*DRNApred*[310])
TCF25	513.7±0.5 - 542.3±8.9*	1-5, 155-158, 192-196, 207-217, 231-234, 357-361	dsDNA: 151-165 RNA: NA
DHX38	1178.0±0.3 - 1211.0±0.1	28-31, 105-114, 119-126, 159-162, 167-170, 179-182, 192-195, 201-206, 224-230, 234-238, 247-251, 255-289, 364-370, 384-389, 444-448, 471-474, 557-564, 583-598, 610-614, 681-686, 705-709, 804-808, 822-825, 829-836, 845-867, 888-891, 997-1001, 1034-1040, 1075-1078, 1102-1105, 1126-1129	dsDNA: NA RNA: NA
CHMP6	1.0±0.1 - 32.9±7.8*	**1-9**, *54-57*, *65-68*	dsDNA: NA RNA: NA
HNRNPA0	63.0±0.5 - 92.1±9.8*	12-16, 42-49, *135-141*, *184-304*	dsDNA: NA RNA: *almost all residues*
ESYT2	76.3±1.3 - 102.0±0.5	149-162, 426-432, 571-578, 838-847,	dsDNA: 11-14, 28-34, 49-56, 60-67, *70-76*, 115-121, 639-644, 867-879, 875-885, RNA: NA
RCOR3	424.1±0.5 - 446.1±0.5	2-6	dsDNA: NA RNA: NA
HNRNPDL	17.0 - 36.0	**28-33**, 153-157, 221-226, 231-234, 268-276, 322-420	dsDNA: NA

			RNA: *15-28*, *35-44*, 56-98, 105-119, 136-178, 185-204, 209-246, 267-286, 301-356, 369-420,
TPM1	1.1±0.3 - 34.3±8.2*	**3-7**	dsDNA: NA RNA: **1-15**, **19-41**, *47-52*, *61-81*
PYHIN1	166.0±0.4 - 181.4±3.5	104-107, 302-313, 349-352, 400-403	dsDNA: 150-153, 400-407, 445-453 RNA: NA
RBM3	1.0±0.2 - 39.7±2.4*	**9-15**, *43-51*, *75-78*, *90-152*	dsDNA: NA RNA: *all residues*
EEF1A1	360.0 - 404.0*	17-26, 51-56, 93-108, 129-134, 179-185, 192-195, 251-269, 307-314, **381-385**, *429-434*, *449-454*	dsDNA: NA RNA: 1-9, 15-22, 61-122, 162-169, 175-200, 238-255, 313-340, **348-399**, *414-432*, *468-474*
MYPN	841.8±0.6 - 862.0±4.9	NA	dsDNA: NA RNA: NA

Table S13: Numbers of raw reads and QC-passed reads for Illumina sequenced library amplicons.

Sub library name	Number of raw paired-end reads	Number of QC-passed reads
p3494	14,043,974	13,735,080
p3491	763,151	756,711
p3492	794,123	782,014
p3493	808,890	800,632
p3495	10,995,273	10,401,665
p3496	1,681,334	1,630,601
p3497	393,726	377,899
p3498	431,311	413,193
p2716	9,898,774	6,399,000
p2701	730,430	469,955
p2705	598,833	233,970
p2711	299,397	53,543

Table S14: Protein peptide composition of the full-length CDS human thyroid display library as determined via in silico translation and proteome mapping of Illumina sequencing reads. Only the 200 most abundant proteins are shown. The exact proportions of proteins might be inaccurate in certain cases due to differences in clustering efficiency in the sequencing process. Values represent the mean per HGNC gene locus (gene symbol). The total counts represent the sum of all counts of all peptides mapped to an HGNC gene locus product (protein).

HGNC symbol	**Mean subject start**	**$SD_{\text{subject start}}$**	**Mean subject end**	**$SD_{\text{subject end}}$**	**Mean e-value**	**$SD_{\text{e-value}}$**	**Total counts**
RPS12	1.12194075	2.5677071	40.0225558	2.90931188	8.09E-07	3.83E-05	267958
RPS28	1.01548242	0.56137126	39.7662235	1.41241507	6.65E-07	5.14E-05	249315
RPL28	1.19023209	4.24620834	40.0095541	4.16813912	8.34E-07	5.42E-05	204834
MT2A	1.01292748	0.60229554	24.5565421	3.74043259	1.35E-05	1.17E-04	199807
RPS15A	1.02276483	0.69296599	39.9312077	1.21769944	2.65E-07	3.87E-05	138591
RPL22	20.9770673	1.09711056	40.0899529	1.54231092	2.12E-06	8.06E-05	127111
RPL12	2.13135428	6.32682603	40.7482886	6.82138696	7.12E-08	1.03E-05	89841
TIMP1	86.0733522	2.47416332	127.923218	2.64543283	2.36E-09	1.24E-07	74340
ZNF701	15.9211558	1.15763911	36.7860764	1.20964535	6.69E-05	5.40E-04	29590
NDUFB10	1.01910675	1.07120993	39.9455674	1.75592886	1.13E-08	9.35E-07	20778
POLR2L	1.02888157	0.79661224	39.8261291	1.92684416	6.72E-09	1.77E-07	20636
ATOX1	1.17152223	2.51084347	37.7659811	2.56161624	2.68E-07	5.11E-06	18913
NDUFA11	110.950644	9.21815443	140.147605	7.02380484	1.43E-08	6.15E-07	15135
RPS15	67.0100373	0.45051626	108.869244	1.50905887	3.74E-08	2.23E-06	14745
UQCRQ	1.00067512	0.02871847	39.7703848	1.08506011	1.29E-09	7.63E-08	13331
RPLP1	1.02128655	0.70174837	40.0661988	3.03600248	4.72E-07	3.75E-05	12825
HMGN2	1.13852814	2.91468006	29.0872214	2.79674513	1.41E-05	2.49E-04	12474
TOMM7	1.02860956	0.69347902	38.777129	4.97260168	2.67E-05	1.64E-04	11989
CRIP1	1.03141614	0.91185316	39.962204	1.55145759	4.85E-07	4.92E-05	10345
FABP5	1.01700515	0.88668063	39.8433583	1.5789398	6.45E-12	4.87E-10	10291
ADIRF	37.0285959	0.42031796	75.9104637	0.59715368	4.46E-08	2.91E-06	9337
NTHL1	256.013472	0.32853535	304.887733	1.03989465	2.40E-09	7.63E-08	9130
NDUFB8	47.3132677	5.35836884	91.9846771	5.48053398	9.86E-06	2.79E-04	8223
TBL3	793.002553	0.06067475	807.998156	0.04290189	8.47E-05	5.32E-04	7051
Q96NF6	8.21307987	9.04138199	35.9048092	8.99022933	0.00110866	0.00205188	7049
PFDN5	128.47376	14.0151232	152.615299	12.3812791	1.48E-07	1.64E-06	6288

CDYL2	1.00038873	0.01971424	21.5187561	3.00396446	4.75E-05	2.14E-04	5145
RPS3	229.601807	3.84875186	242.864257	1.09630995	2.27E-04	6.72E-04	4980
TG	2425.13	37.9822455	2465.19419	37.1936648	1.51E-05	3.31E-04	4954
GPX3	15.7634195	4.3006979	38.8312979	2.73112276	2.84E-05	1.67E-04	3521
SRSF10	1	0	39.9146188	1.15322506	2.25E-10	4.80E-09	3174
RBBP4	1.00065083	0.02550721	29.2375529	5.92549736	1.17E-08	6.46E-07	3073
MT-ND3	41.9660958	0.4669913	68.9598043	1.93336345	4.27E-06	1.21E-04	2861
EIF1AY	1.12170459	3.19093649	35.9559408	7.94045362	1.87E-05	2.84E-04	2769
MYL10	2.59180518	3.45319753	25.277155	2.83963495	6.17E-04	0.00148845	2587
RPL8	222.029833	0.17477463	256.991249	0.23754906	6.46E-11	3.13E-09	2514
FTH1	2.76866585	8.32468259	40.9836801	9.48922799	4.10E-07	2.02E-05	2451
Q96N38	533.382032	4.32102927	550.43073	3.40294549	0.00450661	0.00426812	2382
PKP2	471.828829	2.13101255	489.224324	4.25093296	3.00E-04	0.00113431	2220
ZNF429	646.421698	2.87960371	670.404902	4.52105916	8.90E-04	0.00231932	2203
RPL13A	149.976768	1.00953797	194.772222	2.14132706	4.55E-06	2.02E-04	1980
C9orf85	148.962539	10.6880255	173.974506	7.98501169	0.00150082	0.00314813	1922
TPM1	3.92154756	17.4491357	42.5024181	15.668255	3.97E-07	1.30E-05	1861
TTC17	1074	0	1114.56108	2.64253982	5.50E-19	1.36E-17	1850
UBE2D2	1.00336889	0.10050375	38.5637282	4.10439726	8.01E-08	1.16E-06	1781
MT-CYB	347.97191	8.48159054	379.601124	8.07653099	1.89E-07	4.30E-06	1780
DAP	1.00057504	0.02398006	39.8688902	1.16580226	1.48E-07	6.14E-06	1739
TMEM175	482.008746	0.21177254	504	0	1.63E-09	4.37E-09	1715
PDIA6	357.177297	0.47617666	399.782329	2.80128633	8.83E-10	2.15E-08	1709
NUP210	1840.03597	4.82354146	1880	0	0.00128317	0.00117814	1696
MT-ATP6	161.327522	45.3132931	202.285803	48.7792907	1.47E-05	2.38E-04	1606
HBA1	110.691892	29.212975	135.382432	25.1294894	3.23E-05	5.94E-05	1480
HBA2	110.691892	29.212975	135.382432	25.1294894	3.23E-05	5.94E-05	1480
BCAS3	862.976744	0.27589603	879.671683	1.44258297	2.46E-06	7.86E-05	1462
RAB14	1.00069735	0.02640739	36.4672245	2.44905661	4.06E-09	1.52E-07	1434
DPP7	435.572259	2.75172518	475.833696	6.15302183	2.75E-06	1.87E-05	1377
BTG1	8.08673469	5.27995802	49.6042274	7.67758849	7.81E-13	6.47E-12	1372
BRDT	405.994152	0.10802888	430.989766	0.12052289	4.97E-08	2.81E-07	1368
MAP2K3	2	0	24.590808	1.38026736	0.00819912	0.00219681	1349

GNAS	1.07937685	2.9143307	25.9992582	4.73162442	1.85E-12	6.78E-11	1348
CALM1	38.8753036	38.8558905	75.8437247	36.71738	0.00119862	0.00140232	1235
VPS28	149.026087	0.39493904	189.007826	0.35759759	2.76E-18	8.91E-17	1150
MT-ND2	267.256343	64.4758191	295.165354	67.7684695	1.86E-04	5.15E-04	1143
RPL14	162.109417	28.1491388	203.695067	28.6680112	5.01E-06	1.50E-04	1115
RPS27A	1.13425926	2.18799751	39.1231481	2.26768026	1.37E-27	3.11E-26	1080
Q9P1C3	5.34575955	5.28997955	39.2357875	4.35368965	6.28E-04	0.00137409	1073
MAN2C1	943	0	983.061205	0.2398194	6.39E-19	6.58E-18	1062
SPTBN1	1960.40571	205.907874	2003.21619	204.953955	2.42E-07	3.50E-06	1050
ATP1A1	1.33199195	5.06756392	19.2344064	6.75632435	2.49E-10	5.48E-09	994
RPL36	56.4914228	5.31508973	93.221998	5.42507836	1.61E-08	1.88E-07	991
PRPF6	1.00101626	0.03187884	39.1412602	3.89973065	2.57E-08	4.64E-07	984
NANOGNB	4.64054336	2.24315484	24.6321839	3.55655573	6.18E-05	4.21E-04	957
MYL4	160.001049	0.03239318	196.997901	0.0457868	3.19E-18	1.85E-17	953
EDF1	1	0	39.8617978	1.43317637	1.38E-10	3.89E-09	890
BOLA2	1	0	39.1804767	3.86215675	3.02E-08	1.46E-07	881
BOLA2B	1	0	39.1804767	3.86215675	3.02E-08	1.46E-07	881
CLYBL	261.002278	0.04770018	302.939636	0.90090584	1.66E-11	4.93E-10	878
ENDOG	142.938356	0.59008109	185.973744	0.43080804	1.76E-18	5.20E-17	876
EEF2	8.07258065	40.2716042	40.2165899	42.7888418	1.62E-05	6.72E-05	868
CRYAB	3.25724218	17.9817803	40.8238702	16.8473159	9.12E-06	1.23E-04	863
RUFY2	4.83739837	14.5577134	42.7770035	14.7901281	1.02E-15	3.01E-14	861
RPS4Y2	245.699415	20.5889215	261.435088	18.6261176	6.10E-10	2.13E-09	855
NFIC	459.581967	1.22391292	502.895785	0.30571889	2.89E-12	8.99E-12	854
PSMA1	1.5657277	9.45179614	39.693662	10.4571784	2.20E-07	5.73E-06	852
BPTF	2973.98146	28.9702972	3014.9382	29.84902	3.71E-06	1.05E-04	809
H4C1	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C11	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C12	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C13	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C14	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C15	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C16	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802

H4C2	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C3	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C4	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C5	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C6	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C8	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
H4C9	82.6546135	0.97875557	102.997506	0.07062246	3.44E-07	6.15E-07	802
VEZF1	1.06733167	1.90680629	39.8004988	2.26686488	3.98E-09	1.11E-07	802
MAP3K3	1	0	39.0369898	1.63998851	1.71E-09	4.79E-08	784
Q96CB5	38.1870027	10.8141755	63.1193634	8.12365075	2.30E-04	6.94E-04	754
KRT8	1.54417952	14.5307142	27.5792426	15.0936111	3.31E-07	3.32E-06	713
RPL3	381.556188	0.50288425	402.998578	0.03771571	1.94E-15	3.21E-14	703
FKBP1A	1.02710414	0.40441319	36.0299572	4.64948091	6.63E-08	1.76E-06	701
APEH	609.077698	1.02967799	657.713669	2.48163867	1.32E-09	1.15E-08	695
GABARAPL2	1	0	39.8661871	1.40529843	3.56E-17	3.63E-16	695
Q8N2A0	136.992733	11.752479	158.603198	9.99407925	5.19E-04	0.00151466	688
PRDX5	48.5438336	7.14298658	92.3298663	5.9390217	2.67E-10	6.17E-09	673
FAM50B	280.183056	2.35193351	320.697428	1.82054172	2.48E-15	3.59E-14	661
COX5A	28.0015198	0.03898406	71.9863222	0.35085653	1.17E-22	4.66E-22	658
SUGT1	110.419753	1.68905338	131.405864	6.27305734	4.99E-05	3.14E-04	648
RPS2	213.163551	34.1096202	254.28972	33.8957315	2.67E-13	4.77E-12	642
TTF1	883.372425	45.7154324	901.301109	45.7883466	2.20E-05	2.68E-04	631
Q8N7I0	165.257552	11.9206691	186.084261	5.61312926	2.76E-05	2.42E-04	629
WDR83	295.6112	0.67048777	314.9888	0.24324726	6.24E-08	1.09E-06	625
CX3CL1	317	0	343.236542	1.97645515	4.93E-07	8.62E-06	613
MRPL2	29.0767974	0.2664875	71.2385621	3.7602386	3.48E-08	2.70E-07	612
BRD9	1.0295082	0.16936495	36.7180328	7.95796651	4.31E-09	1.04E-07	610
WDR91	730.985173	0.36529873	747	0	1.04E-09	5.97E-09	607
IMMT	256.365449	20.0023127	297.363787	19.9206133	3.04E-16	7.46E-15	602
NPHP3	361	0	377	0	0.00594944	4.82E-04	597
A1L3X4	1.7979798	2.4084441	31.6582492	7.26116336	2.61E-04	7.86E-04	594
Q8NHA8	306.6914	4.23317938	326.942664	4.18734632	3.17E-04	0.00128114	593
EIF5A2	3.90625	0.52243161	39.9826389	0.16590405	3.64E-14	4.52E-13	576

IFITM3	74.0296167	30.1754178	113.531359	24.0041785	1.71E-14	5.79E-14	574
TPM2	35.8729875	40.5869124	74.118068	41.7758185	4.09E-06	8.53E-05	559
Q8WTZ3	138.544304	15.167851	168.088608	17.1300788	1.58E-04	0.00116879	553
CTSD	21.9175824	1.30861031	39.992674	0.10466807	5.93E-08	8.07E-07	546
SERF2	1.08867925	0.83045991	37.390566	7.15031069	0.00108868	0.00304994	530
VPS13C	1.00189394	0.04351941	39.8863636	1.42968118	3.94E-12	8.88E-11	528
UBXN6	320.90495	180.732677	341.936634	175.785098	1.43E-04	4.50E-04	505
CSTB	1	0	33.9859438	1.55703221	1.27E-12	8.55E-12	498
TLN1	1	0	39.4783505	3.42783414	1.45E-06	9.51E-06	485
MGP	83	0	103	0	1.02E-06	7.96E-06	483
O00370	830.811966	322.786242	863.094017	320.607319	3.55E-04	0.00118463	468
CDC42EP5	118	0	148	0	1.66E-05	1.32E-05	467
AEBP1	1061.83556	2.07747015	1097.98222	5.73308644	3.16E-04	0.00125852	450
C16orf89	363.582022	7.62733588	384.395506	8.65978473	3.98E-04	0.00110117	445
Q86U02	91.0561798	12.4185269	114.442697	8.01937249	0.00211436	0.00246205	445
WFDC2	19.9977477	0.0474579	52.0022523	0.08226132	2.52E-14	3.87E-13	444
HMGN1	1.612529	2.35376608	30.0928074	1.44535157	2.76E-04	0.00113927	431
RPS13	1.00232558	0.04822428	39.9674419	0.27962757	3.35E-19	5.59E-18	430
Q5TEV5	18.1589242	1.46417172	45.6821516	5.85584167	0.00576285	0.00308713	409
PIEZO1	1388	0	1404	0	0.0014426	0.00125028	403
RPLP2	2.26865672	3.33237167	42.0995025	7.02294421	7.97E-08	5.60E-07	402
GVQW3	146.605985	6.07324988	165.798005	2.97936168	3.29E-04	4.19E-04	401
C19orf33	1	0	40	0	7.99E-19	1.20E-17	386
SUMO2	1	0	39.9350649	0.70595795	3.18E-13	3.60E-12	385
EML2	1	0	18.6484375	3.70573582	0.00250724	0.00136218	384
REPS2	585.989101	0.2087983	601.689373	5.21991472	0.00522071	0.00233594	367
KRT18	412.130556	0.36893239	430	0	2.23E-05	3.73E-04	360
GNG8	1.00277778	0.05270463	40	0	2.75E-19	3.16E-18	360
FARP2	450	0	469.656425	3.04160701	0.0050838	6.97E-04	358
PGAM2	119.45098	8.91570892	164.196078	9.36708835	1.89E-19	9.52E-19	357
DYNLRB1	32.0253521	0.27500298	76.0140845	0.82839709	2.12E-13	4.00E-12	355
CDV3	75	0	108.982906	1.7385523	1.50E-06	2.75E-05	351
PGAM1	7.5389049	34.7474882	27.9538905	36.8108453	1.14E-06	2.65E-06	347

MAP7	453	0	488.270893	5.14150515	2.14E-07	1.73E-06	347
BRD1	678	0	693.787172	1.47638681	0.00722157	0.00324405	343
PITPNM1	1220.02933	0.38235626	1244	0	2.02E-06	6.30E-06	341
COX7C	1.00887574	0.09393117	22.8668639	0.80615009	2.06E-06	1.75E-05	338
ZDHHC4	281	0	323	0	1.97E-41	8.75E-41	335
RAB35	1	0	39.7035928	0.45735789	3.42E-21	4.98E-21	334
NDUFA3	1.94461538	5.97758237	33.5538462	8.72280743	3.19E-07	7.83E-07	325
SF3A2	347.416667	36.6294621	390.87037	36.6346093	9.99E-04	0.00195614	324
RAB11FIP1	187.117284	2.42634101	208.259259	2.1174655	1.76E-04	2.88E-04	324
EIF1	52.1055901	26.9291449	68.5372671	23.610155	7.61E-04	8.61E-04	322
ZNF195	93.9657321	5.62489529	112.214953	3.35883547	1.98E-04	7.81E-04	321
RBM3	1	0	39.0603175	0.99657689	1.82E-19	1.91E-19	315
RPL4	1	0	20.5897436	1.98182025	1.22E-06	6.84E-06	312
NCAPH2	571.993569	0.08006325	605	0	1.99E-12	3.97E-12	311
MYH14	1448.98697	0.22829206	1479.13681	4.00010379	3.54E-05	5.71E-04	307
RPA3	71.2315436	2.02428313	117.852349	1.15689029	4.46E-11	7.70E-10	298
CSNK1D	1	0	39.9222973	1.00036068	1.29E-11	2.22E-10	296
RPS20	1	0	39.8817568	1.39595236	7.28E-06	1.17E-04	296
PTK2B	712.01049	0.13203577	726.51049	0.80670117	0.00468028	0.0028588	286
KMT2E	454.007018	0.08362317	489	0	2.72E-20	2.15E-19	285
VTI1B	2.72280702	0.52775686	43.0631579	1.30662337	1.75E-17	2.68E-16	285
BIRC6	2426.01056	0.17801725	2467.95423	0.77140808	2.38E-19	9.28E-19	284
SSU72	1.5212766	8.75372261	40.5035461	8.87807817	8.83E-24	1.48E-22	282
MID1	91.9927273	0.08512452	116.007273	0.25620372	0.00666393	0.00122643	275
C11orf52	1	0	22.1476015	0.97770696	5.96E-09	2.06E-08	271
CRABP1	1.24344569	3.38571525	39.6067416	3.71630621	2.54E-14	2.63E-13	267
FLACC1	256.392453	4.41553083	276.815094	0.78320346	0.0048679	0.00270631	265
AKAP1	842.231939	2.38862741	884.988593	0.18498792	9.37E-21	6.07E-20	263
MT-CO2	86.45	41.1310341	118.657692	45.2480309	1.17E-05	1.86E-04	260
CLPP	222	0	261	0	2.59E-20	4.04E-20	256
KIFAP3	1	0	39.8740157	1.20926676	3.72E-19	5.78E-18	254
NCL	3.47011952	39.0074109	32.438247	39.2865266	5.67E-06	9.69E-06	251
CAPZB	4.2310757	22.7244888	37.9043825	22.6493007	6.72E-08	5.26E-07	251

NPLOC4	1.01606426	0.2534897	23.1767068	1.66828203	1.56E-10	2.45E-09	249
O60397	57.0201613	0.31750032	81.0685484	1.07950108	7.02E-07	1.10E-05	248
NSA2	1	0	30.9798387	6.380924	2.21E-10	5.25E-10	248
NNMT	245.201613	0.40201595	263.995968	0.06350006	2.16E-13	6.82E-13	248
RAB1A	3.218107	0.46104592	40.5884774	1.54391332	1.79E-21	1.65E-20	243
SPTY2D1	1	0	39.838843	1.37360246	5.51E-13	8.55E-12	242
Q8N769	47.6735537	2.95134173	66.892562	1.29667083	8.96E-04	0.0018155	242
PGLS	221	0	258	0	2.25E-15	1.12E-14	240
BCL10	1	0	39.8291667	1.82267888	3.46E-08	3.78E-07	240

Table S15: Values from the FACS enrichments. Number of screened cells, the resulting (sub-)library coverage and the number of sorted cells. Library coverage was calculated, based on **Equation (1**) based on the determined library diversities.

Library (selection)	Selection round	Screened cells	Library coverage	Sorted cells
Fragmented prostate (Myc-tag)	1	3,002,752	99.99%	4,641
	2	100,773	100.00%	3,890
	3	102,469	100.00%	9,445
Fragmented prostate (30NC)	1	3,106,381	99.98%	63,100
	2	654,834	100.00%	118,845
	3	134,314	100.00%	41,214
Full-length thyroid (8NmC)	1	1,466,611	79.80%	669
	2	50,998	100.00%	2,351
	3	52,523	100.00%	20,958

Table S16: Enrichment of 5mCpG-specific reader proteins. EF of proteins per round and continuously monitored EF. The peptide counts were summed for each protein, normalized on the read number and these fractions were used to calculate the EF between two sub libraries.

	EF per round			Continuous EF		
HGNC symbol	**EF1**	**EF2**	**EF3**	**1 round**	**2 rounds**	**3 rounds**
TPM1*	0.9932752	NA	NA	0.9932752	NA	NA
TMEM167B	6.61514171	0.42187783	NA	6.61514171	2.79078165	NA
Q8WTZ3	0.28515493	1.56094799	NA	0.28515493	0.44511201	NA
Q8N976*	NA	5.46331795	NA	NA	5.46331795	NA
PTK2B	0.52073455	1.83640939	NA	0.52073455	0.95628183	NA
PKP2	NA	5.20315995	NA	NA	5.20315995	NA
PCBD1*	34.0061731	1.01180562	NA	34.0061731	34.407637	NA
NDUFA11	0.39013147	2.16309409	NA	0.39013147	0.84389107	NA
MT1F*	4.06741821	1.20072922	NA	4.06741821	4.8838679	NA
MT1E	2.82599777	2.49751678	NA	2.82599777	7.05797683	NA
KMT5A*	8.76059308	NA	NA	8.76059308	NA	NA
HHATL	1.04292775	3.12189597	NA	1.04292775	3.25591193	NA
FLACC1*	0.33058842	NA	NA	0.33058842	NA	NA
FABP5*	0.01106673	NA	NA	0.01106673	NA	NA
EEF1AKMT2	12.5268294	NA	NA	12.5268294	NA	NA
DZIP1L*	NA	0.67867304	NA	NA	0.67867304	NA
CRYAB*	0.11166457	2.55427852	NA	0.11166457	0.28522241	NA
ATOX1*	0.15748964	NA	NA	0.15748964	NA	NA
CIRBP	83.552522	38.3128454	1.13785356	83.552522	3201.13486	3642.4227
RBM3	54.8440938	28.5482505	1.10328355	54.8440938	1565.70293	1727.41428
RACK1	77.6926281	2.51357401	1.31522227	77.6926281	195.28617	256.84472
BDH1	31.8147854	2.89567162	1.70693707	31.8147854	92.1251712	157.251869
CRIP1	41.0464907	2.92042515	0.97944201	41.0464907	119.873204	117.408852
RPL24	36.4277684	1.6657096	1.02602305	36.4277684	60.6780834	62.2571122
FAM50B	30.0192788	1.48858616	0.87395178	30.0192788	44.686283	39.0536564
ZDHHC4	27.4324243	1.61600525	0.68403224	27.4324243	44.3309418	30.3237933
C19orf33	9.62303488	1.48731836	0.9301962	9.62303488	14.3125165	13.3134484

MT2A	4.15385761	1.35603227	1.67236113	4.15385761	5.63276496	9.42001714
C9orf85	NA	NA	7.28293148	NA	NA	7.28293148
NTHL1	7.03534156	1.19902606	0.77588048	7.03534156	8.43555785	6.54498471
O00370	0.14975373	5.46331795	5.61826142	0.14975373	0.81815223	4.59659311
A1L3X4	1.46009885	1.57671514	1.83529873	1.46009885	2.30215995	4.22515124
UBE2D2	1.07724306	1.39701281	1.33768129	1.07724306	1.50492235	2.01310647
Q96N38	1.978673	1.32303398	0.72829315	1.978673	2.61785161	1.90656339
C16orf89	0.86621595	1.41904362	1.52941561	0.86621595	1.22919821	1.87995493
Q86U02	0.25592744	2.40145844	2.18487944	0.25592744	0.61459911	1.34282495
NUP210	0.41840097	1.11771584	2.5615828	0.41840097	0.46765339	1.19793287
Q8N2A0	0.57300391	1.31813385	1.37992386	0.57300391	0.75529585	1.04225076
RPL28	0.80568681	1.61315083	0.79144082	0.80568681	1.29969435	1.02863116
ZNF701	0.37807629	3.83330688	0.69113533	0.37807629	1.44928243	1.0016503
Q96NF6	3.73465487	0.87475421	0.30619451	3.73465487	3.26690508	1.00030839
RPS27A	0.70571444	1.11239971	0.84575978	0.70571444	0.78503654	0.66395234
RPL22	0.27761302	1.7252583	1.26420697	0.27761302	0.47895417	0.6054972
MT-ND3	1.24626403	0.65966352	0.45730035	1.24626403	0.82211492	0.37595344
ADIRF	0.65584819	2.07679775	0.2725226	0.65584819	1.36206405	0.37119323
POLR2L	0.41476543	1.29413293	0.60421357	0.41476543	0.5367616	0.32431865
MYL10	0.0372503	6.81140939	1.09243972	0.0372503	0.25372704	0.27718149
RPS28	0.6897366	0.79204452	0.44838289	0.6897366	0.5463021	0.24495251
UQCRQ	0.34763737	1.36914908	0.41437369	0.34763737	0.47596738	0.19722836
NDUFB10	0.13239129	3.20143472	0.40712039	0.13239129	0.42384207	0.17255475
HMGN2	0.34553566	0.04441722	11.2365229	0.34553566	0.01534773	0.17245515
RPS12	0.27250369	0.74348692	0.64941001	0.27250369	0.20260293	0.13157237
RPL12	0.22457058	0.83368043	0.50447623	0.22457058	0.1872201	0.09444809
TOMM7	0.4479309	1.1407907	0.17557067	0.4479309	0.5109954	0.08971581
TIMP1	0.19444416	0.6149189	0.40336236	0.19444416	0.11956739	0.04822898
RPS15A	0.12301025	0.27914178	0.25113557	0.12301025	0.0343373	0.00862332

Table S17: Enrichment of 5mCpG-specific reader proteins (continuation of Table S16). Values represent the mean per HGNC gene locus (gene symbol). The peptide counts represent the counts in the library after the final enrichment step. *Protein not detected in the final library; The respective values are derived from latest selection step in which the peptides of this protein were detected.

HGNC symbol	Mean subject start	SD$_{\text{subject start}}$	Mean subject end	SD$_{\text{subject end}}$	Mean e-value	SD$_{\text{e-value}}$	Mean Alignment length	SD$_{\text{alignment length}}$	Total counts
TPM1*	39.971564	0.29137918	3.15E-17	1.07E-16	39.971564	0.29137918	39.971564	0.29137918	211
TMEM167B	40	0	1.97E-21	0	40	0	40	0	1
Q8WTZ3	185.5	9.19238816	1.91E-05	2.70E-05	41.5	6.36396103	185.5	9.19238816	2
Q8N976*	64.8571429	8.11230694	6.12E-06	6.94E-06	29.2857143	9.08688223	64.8571429	8.11230694	7
PTK2B	726.333333	0.57735027	0.00533333	0.00351188	15	0	726.333333	0.57735027	3
PKP2	490.25	2.87228132	9.55E-06	1.90E-05	18.75	2.5	490.25	2.87228132	4
PCBD1*	39.991453	0.09245003	6.41E-21	4.26E-21	39.991453	0.09245003	39.991453	0.09245003	117
NDUFA11	141	0	4.44E-13	5.50E-13	30	0	141	0	2
MT1F*	30.2	13.8636215	1.55E-05	1.52E-05	30.2	13.8636215	30.2	13.8636215	5
MT1E	24	0	2.43E-05	0	24	0	24	0	1
KMT5A*	117.563636	2.21746967	2.40E-13	1.78E-12	32.5636364	2.21746967	117.563636	2.21746967	55
HHATL	244	0	0.00366667	0.00208167	20.3333333	1.15470054	244	0	3
FLACC1*	277	0	0.005	0.00312694	19	0	277	0	10
FABP5*	40	0	5.97E-21	7.36E-21	40	0	40	0	13
EEF1AKMT2	277.5	6.36396103	9.15E-07	1.29E-06	28.5	6.36396103	277.5	6.36396103	2
DZIP1L*	27	0	0.00418949	0.00350167	26	0	27	0	23
CRYAB*	40	0	4.36E-18	1.16E-17	40	0	40	0	11
ATOX1*	37.9647059	0.49652619	2.94E-16	5.33E-15	37.9647059	0.49652619	37.9647059	0.49652619	340
CIRBP	39.3134182	3.03736428	1.77E-09	3.20E-08	39.3114594	3.03659792	39.3134182	3.03736428	2042
RBM3	39.9363057	0.77662589	8.98E-11	6.03E-09	39.9358665	0.77715529	39.9363057	0.77662589	4553
RACK1	23.0081633	0.12777531	4.04E-16	1.37E-15	23.0040816	0.23078553	23.0081633	0.12777531	245
BDH1	336.28	2.28254244	1.70E-08	6.04E-08	28.88	2.61916017	336.28	2.28254244	25
CRIP1	39.9797304	0.55016274	1.35E-08	1.36E-06	39.9797304	0.55016274	39.9797304	0.55016274	10163
RPL24	40	0	2.32E-22	2.46E-21	40	0	40	0	112
FAM50B	320.87037	1.7122953	3.13E-13	4.59E-12	41.8981481	1.71418061	320.87037	1.7122953	216
ZDHHC4	322.705882	2.50154014	9.88E-19	9.11E-18	42.7058824	2.50154014	322.705882	2.50154014	85
C19orf33	40	0	1.58E-19	9.73E-20	40	0	40	0	43

MT2A	27.472157	9.64335129	9.44E-06	3.99E-05	27.4722205	9.64352926	27.472157	9.64335129	15749
C9orf85	165	15.6044865	1.18E-04	2.64E-04	35.2	12.6372465	165	15.6044865	5
NTHL1	304.996	0.06318215	2.71E-23	4.37E-23	49.994	0.07730428	304.996	0.06318215	500
O00370	870.222222	335.236099	1.51E-05	4.45E-05	41.1111111	11.1560529	870.222222	335.236099	18
A1L3X4	27.6190476	2.83683257	7.46E-05	1.74E-04	27.6190476	2.83683257	27.6190476	2.83683257	21
UBE2D2	40	0	7.96E-22	8.01E-22	40	0	40	0	30
Q96N38	552.5	6.26336409	4.39E-04	0.00139331	23.5789474	1.28676378	552.5	6.26336409	38
C16orf89	372.714286	17.969551	8.64E-04	0.00226457	33.8571429	9.04486174	372.714286	17.969551	7
Q86U02	108	8.91627725	8.00E-04	0.00130383	33.8	15.5145093	108	8.91627725	5
NUP210	1880	0	0.00123529	7.52E-04	40.5882353	5.820855	1880	0	17
Q8N2A0	148.5	16.9440255	5.80E-06	6.42E-06	32.1666667	15.6386274	148.5	16.9440255	6
RPL28	39.9461146	0.66358197	4.05E-13	1.70E-11	39.9461146	0.66358197	39.9461146	0.66358197	1763
ZNF701	37.9395161	0.74771647	9.77E-08	1.51E-06	25.9475806	1.29514595	37.9395161	0.74771647	248
Q96NF6	34.2711864	10.9589052	8.47E-04	5.77E-04	19.7288136	3.59995454	34.2711864	10.9589052	59
RPS27A	39	0	1.68E-33	3.53E-33	39	0	39	0	6
RPL22	39.9937888	0.07862697	3.60E-07	2.95E-06	20.0232919	0.79329482	39.9937888	0.07862697	644
MT-ND3	68.7777778	0.66666667	9.17E-09	1.47E-08	27.7777778	0.66666667	68.7777778	0.66666667	9
ADIRF	76	0	3.28E-16	0	40	0	76	0	29
POLR2L	39.6428571	2.67261242	6.54E-07	4.89E-06	39.6428571	2.67261242	39.6428571	2.67261242	56
MYL10	25.5	1.22474487	1.65E-05	2.86E-05	23	4.5607017	25.5	1.22474487	6
RPS28	39.8473581	0.5127663	1.24E-17	2.81E-16	39.8473581	0.5127663	39.8473581	0.5127663	511
UQCRQ	40	0	7.56E-22	1.68E-23	40	0	40	0	22
NDUFB10	40	0	3.98E-21	7.64E-21	40	0	40	0	30
HMGN2	28.6111111	1.41997883	6.30E-07	2.47E-06	28.6111111	1.41997883	28.6111111	1.41997883	18
RPS12	38.9966102	3.78818276	1.91E-09	2.40E-08	38.9966102	3.78818276	38.9966102	3.78818276	295
RPL12	39.6056338	0.66502144	3.00E-33	1.18E-32	39.6056338	0.66502144	39.6056338	0.66502144	71
TOMM7	40	0	3.06E-21	0	40	0	40	0	9
TIMP1	128	0	6.54E-22	2.09E-21	42.8	0.61025715	128	0	30
RPS15A	40	0	2.86E-20	1.90E-22	40	0	40	0	10

Table S18: Sanger sequencing results of the cDNA inserts of 27 single clones of the full-length CDS human thyroid library after the final selection round for 5mCpG readers. Nucleic acid sequences were subjected to nucleotide blast search against the human transcriptome and proteome and the theoretically displayed amino acid sequences were determined by in silico translation of the ORF by the virtual ribosome 2.0 suite.[384]

Clone	Insert length / bp	Transcript (blastN)	Theor. displayed amino acid sequence	Number of displayed residues
1	**824**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
2	**825**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
3	**949**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
4	**263**	Homo sapiens splicing factor 3b subunit 2 (SF3B2), mRNA	MXMXTXRWMTPLWAPRSPXLWXRSCS	**26**
5	**825**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
6	**826**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
7	**821**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGS	**168**

			RDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	
8	**827**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
9	**923**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 1, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYATHNE	**172**
10	**826**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
11	**826**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
12	**611**	unknown	MPGWFLYFW	**9**
13	**920**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 1, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYATHNE	**172**
14	**827**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
15	**997**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 2, non-coding RNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYEAGQRAEARGIPGRPQTASRLASSAVASRVLSILC	**202**

16	**923**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 1, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYATHNE	**172**
17	**400**	Homo sapiens cysteine rich protein 1 (CRIP1), mRNA	MPKCPKCNKEVYFAERVTSLGKDWHRPCLKCEKCGKTLTSGGHAEHEGKPYCNHPCYVAMFGPKGFGRGGAESHTFK	**77**
18	**827**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
19	**824**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
20	**826**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
21	**825**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
22	**825**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
23	**952**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYV	**168**
24	**823**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGR	**168**

			GRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	
25	**950**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MXXXXXXFFXGRVIFDSNELVLDSVFSVTDRSLXWLL	**37**
26	**826**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
27	**825**	Homo sapiens cold inducible RNA binding protein (CIRBP), transcript variant 4, mRNA	MASDEGKLFVGGLSFDTNEQSLEQVFSKYGQISEVVVVKDRETQRSRGFGFVTFENIDDAKDAMMAMNGKSVDGRQIRVDQAGKSSDNRSRGYRGGSAGGRGFFRGGRGRGRGFSRGGGDRGYGGNRFESRSGGYGGSRDYYSSRSQSGGYSDRSSGGSYRDSYDSYG	**168**
average	811.7±154.3			150.3±48.6

14.3 Supplementary Calculations

$$c_{cells} = \frac{n_{harvested\ cells}}{V_{suspension}} = \frac{3.2 * 10^{8}\ cells}{200\ \mu L} = 1.6 * 10^{6}\ \frac{cells}{\mu L}$$

$$\boldsymbol{c_{Myc-tag}} = c_{cells} * \frac{Int'}{cell} = 1.6 * 10^{6}\ \frac{cells}{\mu L} * 3.6 * 10^{4} \frac{Int'}{cell} \triangleq 5.76 * 10^{10}\ \frac{displ.Myc}{\mu L}$$

$$\triangleq \mathbf{95.65\ \boldsymbol{nM}}$$

$$\boldsymbol{c_{antibody}} = \frac{\frac{c_{Stock}}{MW}}{dilution\ factor} = \frac{\frac{0.5\ \frac{mg}{mL}}{6.7 * 10^{5} \frac{g}{mol}}}{500} = \mathbf{14.92\ \boldsymbol{nM}}$$

Equation S1: Calculation of the molar concentration of displayed Myc-tag and of the used anti-Myc-antibody in the FACS screening assay. The amount of Int' molecules per cell was obtained from calculations of Wentzel et al.[242] The antibody stock concentration was obtained from the manufacturers website. The MW of the antibody was taken from a similar product of SantaCruz Biotechnology (c-Myc (9E10)).

$$c_{cells} = \frac{n_{harvested\ cells}}{V_{suspension}} = \frac{3.2 * 10^{8}\ cells}{200\ \mu L} = 1.6 * 10^{6}\ \frac{cells}{\mu L}$$

$$\boldsymbol{c_{displ.proteins}} = c_{cells} * \frac{Int'}{cell} = 1.6 * 10^{6}\ \frac{cells}{\mu L} * 3.6 * 10^{4} \frac{Int'}{cell} = 5.76 * 10^{10}\ \frac{Int'}{\mu L}$$

$$\triangleq \mathbf{95.65\ \boldsymbol{nM}}$$

$$\boldsymbol{c_{probe}} = \mathbf{40.0\ \boldsymbol{nM}}$$

Equation S2: Calculation of the molar concentration of displayed DNA binding proteins after three rounds of FACS selection. The amount of Int' molecules per cell was obtained from calculations of Wentzel et al.[242] The 30NC stock concentration was determined and adjusted via spectrophotometric measurements.

$$c_{cells} = \frac{n_{harvested\ cells}}{V_{suspension}} = \frac{3.2 * 10^{8}\ cells}{200\ \mu L} = 1.6 * 10^{6}\ \frac{cells}{\mu L}$$

$$\boldsymbol{c_{displ.proteins}} = c_{cells} * \frac{Int'}{cell} = 1.6 * 10^{6}\ \frac{cells}{\mu L} * 3.6 * 10^{4}\frac{Int'}{cell} = 5.76 * 10^{10}\ \frac{Int'}{\mu L}$$
$$\triangleq \mathbf{95.65}\ \boldsymbol{nM}$$

$$\boldsymbol{c_{probes}} = \mathbf{6.0}\ \boldsymbol{nM}$$

Equation S3: Calculation of the maximal concentration of displayed 5mC binding proteins after three rounds of FACS selection. The amount of Int' molecules per cell was obtained from calculations of Wentzel et al.[242] The 8NmC and 8NC stock concentration was determined and adjusted via spectrophotometric measurements.

14.4 Vector maps

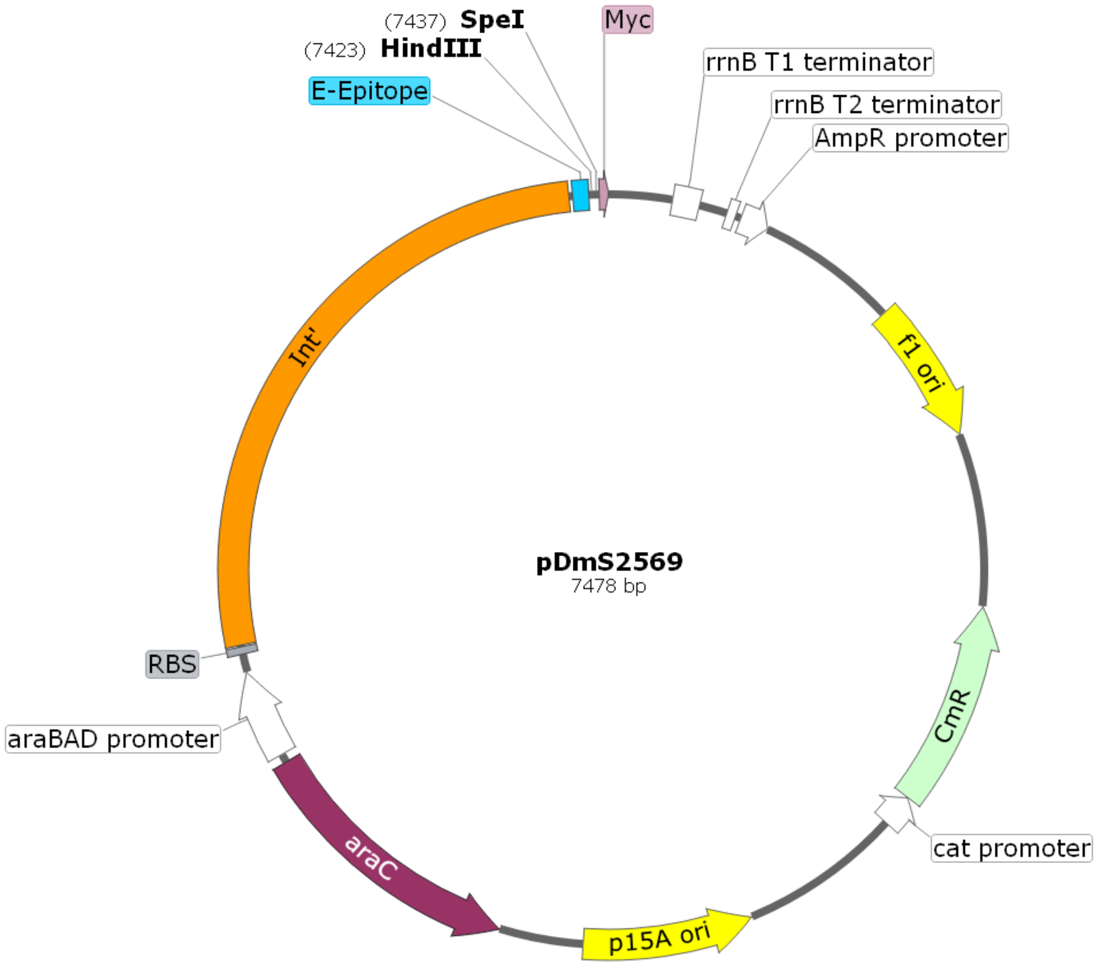

Figure S10: Vector map of pDmS2569. Gibson assembly entry vector for the intimin-mediated bacterial surface display of full-length human CDS containing the N-terminal E-epitope and the C-terminal Myc-tag.

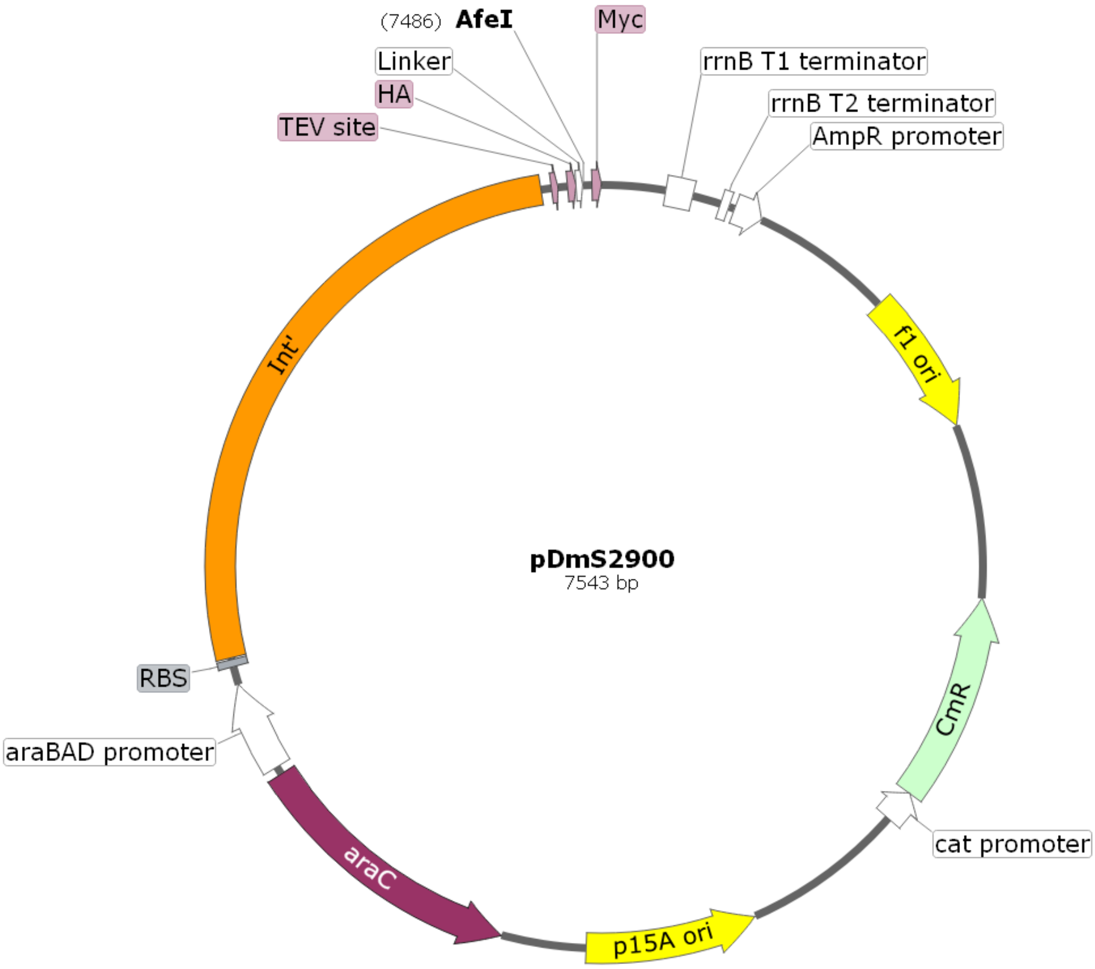

Figure S11: Vector map of pDmS2900. Blunt-end cloning entry vector for the intimin-mediated bacterial surface display of fragmented human CDS containing the N-terminal TEV-site, HA-tag and linker as well as and the C-terminal Myc-tag (out of frame in the empty vector).

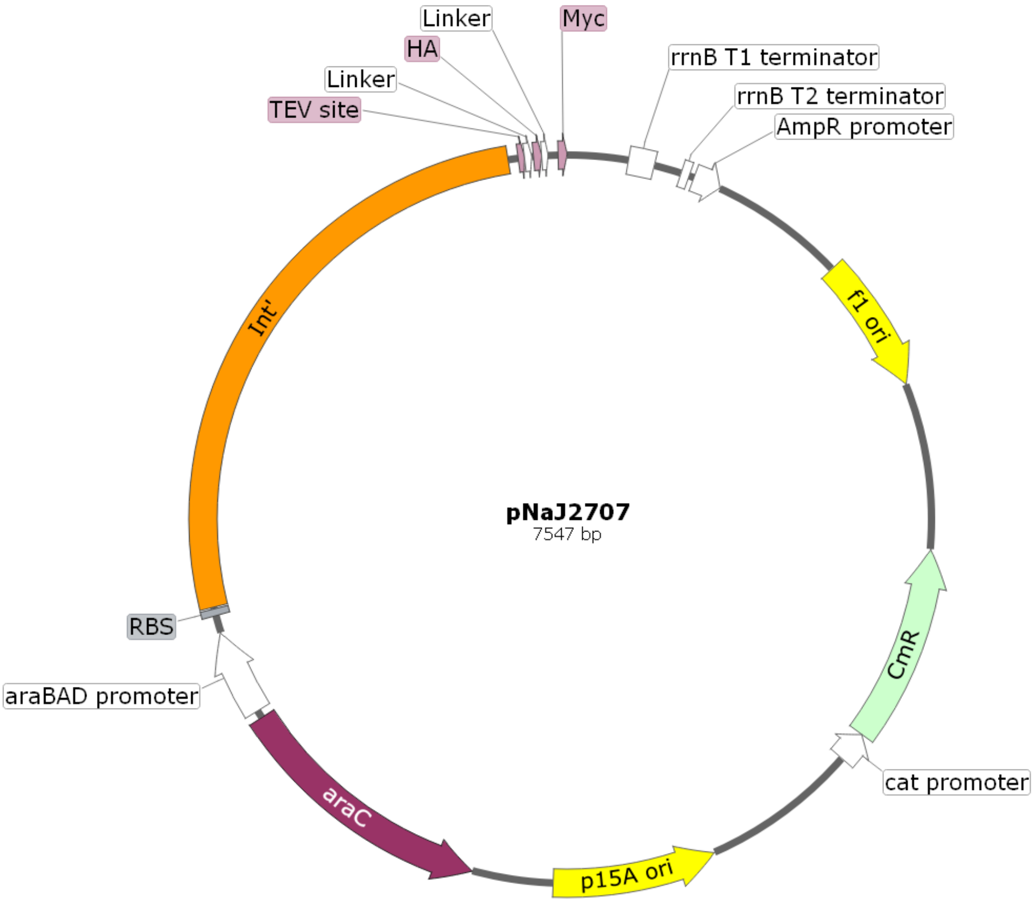

Figure S12: Vector map of pNaJ2707. All features are identical to pDmS2900 except of the removal of the AfeI site and shifting the Myc-tag in-frame to the Int'-ORF to allow for surface staining.

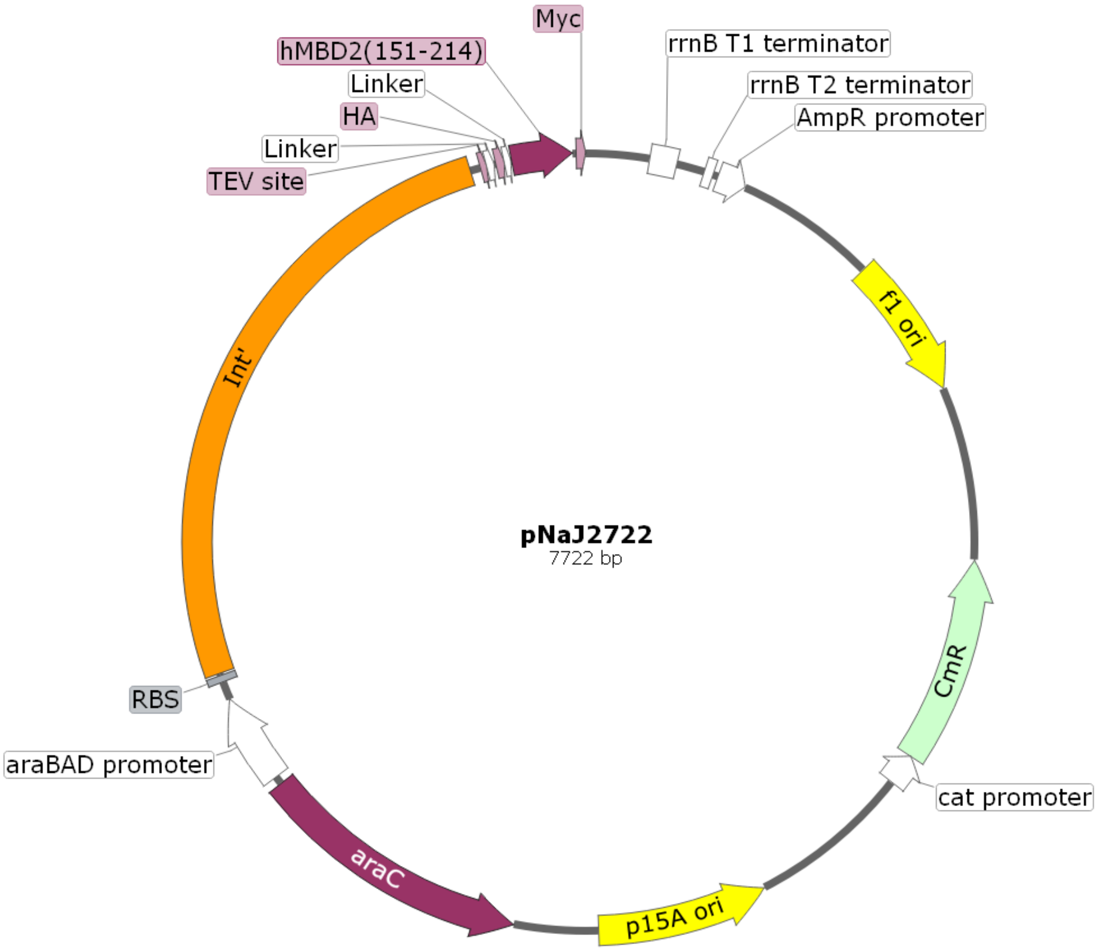

Figure S13: Vector map of pNaJ2722. Vector for the intimin-mediated bacterial surface display of fragmented human MBD2 containing the N-terminal TEV-site, HA-tag, linkers, the sequence coding for the hMBD2 DNA binding domain (residues 151-214) as well as and the C-terminal Myc-tag.

Eidesstattliche Versicherung (Affidavit)

Name, Vorname (Surname, first name)	Matrikel-Nr. (Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum (Place, date)	Unterschrift (Signature)

Titel der Dissertation:
(Title of the thesis):

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum (Place, date)	Unterschrift (Signature)