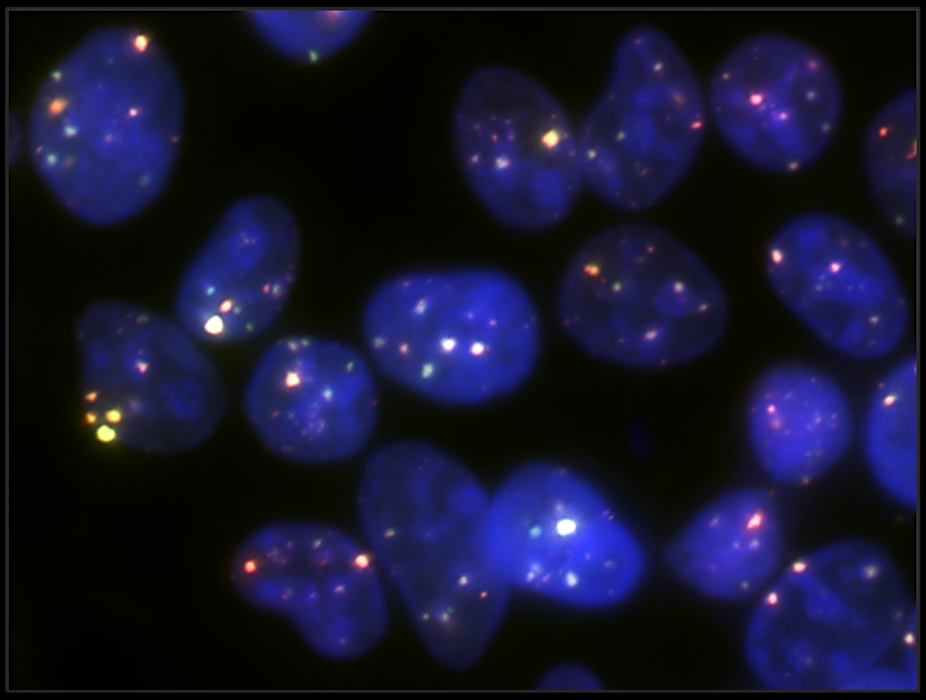
Sequence-specific DNA binders for Nucleotide Resolution Analysis of Genomic 5-Methylcytosine by Cell Imaging

Álvaro Muñoz López



Dortmund 2020

Sequence-specific DNA binders for Nucleotide Resolution Analysis of Genomic 5-Methylcytosine by Cell Imaging

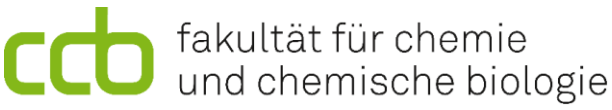
Dissertation Submitted for the degree of Doctor of Natural Sciences

(Dr. rer. nat.)

Presented by

Álvaro Muñoz-López

at the



of the

Dortmund 2020

All we have to decide is what to do with the time that is given us.

J.R.R. Tolkien, *The Fellowship of the Ring.*

This work was prepared from September 2016 to October 2020 in the group of Prof. Dr. Daniel Summerer at the Chemistry and Chemical Biology Faculty of the Technical University of Dortmund. PhD Program in Chemical and Molecular Biology was organized by the International Max Planck Research School funded by the Max Planck Institute of Molecular Physiology. The work presented here was funded by the European Research Council (ERC), the Deutsche Forschungsgemeinschaft (DFG), the Volkswagenstiftung and the Technical University of Dortmund.

List of Publications

Parts of this work are published in:

[1] **Muñoz-Lopez A**, Buchmuller B, Wolffgramm J, Jung A, Hussong M, Kanne J, Schweiger MR, Summerer D. *Designer Receptors for Nucleotide Resolution Analysis of Genomic 5-Methylcytosine by Cellular Imaging.*
Angewandte Chemie International Edition. 2020 Jun 2;59(23):8927-8931.

[2] **Munoz-Lopez A**, Jung A, Buchmuller B, Wolffgramm J, Maurer S, Witte A, Summerer D. *Engineered TALE Repeats for Enhanced Imaging-based Analysis of Cellular 5-Methylcytosine.*
ChemBioChem. 2020 Sep 29 (Accepted, Online ahead of print).

[3] **Muñoz-López Á**, Summerer D. *Recognition of Oxidized 5-Methylcytosine Derivatives in DNA by Natural and Engineered Protein Scaffolds.*
Chemical Records. 2018 Jan;18(1):105-116.

Other Publications:

[4] Buchmuller B, Jung A, Muñoz-López Á, Summerer D. *Programmable tools for targeted analysis of epigenetic DNA modifications.*
Current Opinion in Chemical Biology. 2021 vol 63:1-10.

[5] Witte A, **Muñoz-López Á**, Metz M, Schweiger MR, Janning P, Summerer D. *Encoded, Click-Reactive DNA-Binding Domains for Programmable Capture of Specific Chromatin Segments.*
Chemical Science. 2020 (Accepted, Online ahead of print).

[6] Kanne J, Hussong M, Isensee J, **Muñoz-López Á**, Wolfgramm J, Bessonov S, Grimm C, Meder L, Heß F, Reinhardt H, Odenthal M, Hucho T, Büttner R, Summerer D, Schweiger MR. *Pericentromeric Satellite III transcripts induce etoposide resistance.*
Cell Death and Disease. 2020 (*Currently under revision*).

[7] Wolffgramm J, Palei S, **Muñoz-López Á**, Buchmuller B, Kanne J, Schweiger MR, Summerer D. *Light-Control of DNA-Methyltransferase Catalysis.*
Angewandte Chemie International Edition. 2020 (*Currently under revision*).

[8] Buchmuller B., **Muñoz-López Á.**, Gieß M., Summerer D. *Design and Application of DNA Modification-Specific Transcription-Activator-Like Effectors.*
In: Ruzov A., Gering M. (eds) DNA Modifications.
Methods in Molecular Biology. 2021 vol 2198. Humana, New York, NY.

[9] Palei S, Buchmuller B, Wolffgramm J, **Muñoz-Lopez Á**, Jung S, Czodrowski P, Summerer D. *Light-Activatable TET-Dioxygenases Reveal Dynamics of 5-Methylcytosine Oxidation and Transcriptome Reorganization.*
Journal of the American Chemical Society. 2020 142 (16), 7289-7294.

[10] Gieß M, **Muñoz-López Á**, Buchmuller B, Kubik G, Summerer D. *Programmable Protein-DNA Cross-Linking for the Direct Capture and Quantification of 5-Formylcytosine.*
Journal of the American Chemical Society. 2019 Jun 19;141(24):9453-9457.

[11] **Muñoz-López A**, Romero-Moya D, Prieto C, Ramos-Mejía V, Agraz-Doblas A, Varela I, Buschbeck M, Palau A, Carvajal-Vergara X, Giorgetti A, Ford A, Lako M, Granada I, Ruiz-Xivillé N, Rodríguez-Perales S, Torres-Ruíz R, Stam RW, Fuster JL, Fraga MF, Nakanishi M, Cazzaniga G, Bardini M, Cobo I, Bayon GF, Fernandez AF, Bueno C, Menendez P. *Development Refractoriness of MLL-Rearranged Human B Cell Acute Leukemias to Reprogramming into Pluripotency.*

Stem Cell Reports. 2016 Oct 11;7(4):602-618.

[12] Bueno C, van Roon EH, **Muñoz-López A**, Sanjuan-Pla A, Juan M, Navarro A, Stam RW, Menendez P. *Immunophenotypic Analysis and Quantification of B-1 and B-2 B Cells During Human Fetal Hematopoietic Development.*
Leukemia. 2016 Jul;30(7):1603-6.

[13] **Muñoz-López Á**, van Roon EH, Romero-Moya D, López-Millan B, Stam RW, Colomer D, Nakanishi M, Bueno C, Menendez P. *Cellular Ontogeny and Hierarchy Influence the Reprogramming Efficiency of Human B Cells into Induced Pluripotent Stem Cells.*
Stem Cells. 2016 Mar;34(3):581-7.

[14] Bueno C, Sardina JL, Di Stefano B, Romero-Moya D, **Muñoz-López A**, Ariza L, Chillón MC, Balanzategui A, Castaño J, Herreros A, Fraga MF, Fernández A, Granada I, Quintana-Bustamante O, Segovia JC, Nishimura K, Ohtaka M, Nakanishi M, Graf T, Menendez P. *Reprogramming Human B Cells into Induced Pluripotent Stem Cells and Its Enhancement by C/EBPα.*
Leukemia. 2016 Mar;30(3):674-82.

Acknowledgments

PhD is a long learning process full of ups and downs in which we develop ourselves not only as scientists but also as individuals. Now that mine is coming to an end, I see the importance to thank and remember all the people and institutions whose help and support led me to be here today.

First of all, I would like to thank Prof. Dr. Daniel Summerer for inviting me and giving me the opportunity to do my PhD at his lab. Undoubtedly, part of my success is directly derived from his fantastic supervision and good decision-making while managing projects together with fruitful and frequent discussions. Moreover, I would like to highlight his effort to create a good working atmosphere, encouraging team spirit, allowing for creativity and research freedom and being understanding at scientific and personal level.

I would also like to thank to Prof. Dr. Daniel Rauh for his time and dedication as second evaluator of my doctoral thesis.

High School education and adolescence play an important role in the career choices we make and the Bachelor we decide to study. My secondary education at I.E.S. Jaranda sparked my interest in Science and that is why I ended up studying a B.Sc. in Biotechnology. I want to thank all of the teachers, classmates and friends for all of the many unforgettable moments we shared.

I extend my thanks to Educational Institutions to the Pablo de Olavide University where I was well trained by professors who were highly motivated for teaching, approachable and really caring for the future of the students. It was a pleasure to study and learn when most of the professors are easily available and willing to help. I would like to mention three of them who had special importance for the development of my career: Carlos Santos Ocaña, Ángel Manuel Carrión Rodríguez and "Manolo" (Manuel Jesús) Muñoz Ruíz.

I thank Carlos Santos Ocaña, who I consider my first "Scientific Father", for being very open and always willing to help offering his lab and expertise every time we needed. I feel grateful for his full support, his advices and for being such a good professor.

Acknowledgments

I would like to acknowledge Ángel Manuel Carrión for the opportunity of working in his lab, where I could have my first touch with the field of Neuroscience and using mice as animal model. My experience in his group led me to be awarded with a summer research fellowship.

I extend my thanks to Manuel "Manolo" Muñoz Ruiz, for his interesting, engaging and amusing lectures. It was very inspiring to hear about his innovative ideas. I feel very grateful for the project we started together with David Caballero Pradas, for your constant support, for offering your lab and for all of the lunches and beer meetings we had while discussing and dreaming about start-ups.

During this stage of my career, there were particularly important agencies and institutions that funded my studies and additional trainings and opportunities. First of all, I would like to thank the Spanish Ministry of Education and Culture for funding my Bachelor and Master studies for six years through a Scholarship. Also, I would like to highlight the importance of Pablo de Olavide University for granting me an Atlánticus Scholarship to study a semester abroad in the U.S.A. Thanks to this, I had the opportunity to work at the lab of Professor Geoffrey B. Smith and had my first touch with deeper experimental work and training in virology techniques. I extend my thanks to the Spanish Research Council (CSIC) for the JAE-Intro Fellowship and the National Center for Cardiovascular Research (CNIC) for the Cicerone Fellowship; both of them are Summer Research Programs. The first one allowed me to gain more experience in the field of Cell Therapies for Neuropathologies under the supervision of Manuel Álvarez Dolado and the second one provided me a strong practical knowledge and skills in the design of Viral Vectors at Juan Carlos Ramirez's lab. I would like recognize Juan Carlos and express my gratitude for his help and full support during all these years. These experiences were essential in the development of my career.

I thank Jordi Surrallés and his team for the opportunity of doing my Master thesis at his lab, where I could learn and apply cutting-edge techniques (like CRISPR and TALENs). Special thanks to my Master thesis supervisor Miriam Aza Carmona from who I learned a lot. Also, thanks to Jordi for recommendation letters and further support over these years.

I thank Pablo Menéndez for my stay at his lab and all the highly valuable experience I got in the field of Stem Cells and Blood Cancer. Moreover, I learnt from him a set of useful skills like goal-oriented project planning, ambition, networking and productive collaboration. I thank the members of his group as well, especially Alessandra Giorgetti, Julio Castaño, Silvia Cufí and Lorena Ariza. In addition, I would like to mention the labor of the Spanish Ministry of Economy and Competitiveness for the FPI Fellowship to fund this important and productive period of my career.

I would like to recognize the labor and fundamental role of the people I spent most of my time with in the last four years: the members of the Summerer and Dehmelt labs: Sarah Flade, Preeti Rathi, Sara Maurer, Mario Gieß, Anna Witte, Jan Wolffgramm, Benjamin Buchmuller, Shubhendu Palei, Tzu Chen Lin, Sudakshina Banerjee, Katharina Kuhr, Simone Eppmann, Nadine Schmidt, Damian Schiller, Brinja Kosel, Dominic Kamps and Suchet Nanda. Thanks to Moritz Schmidt and Grzegorz Kubik as well for his help and good feedback about the Summerer's Lab before I joined. Also, I would like to mention and thank my students Philip Berninger, Leonie Fleige and Frederik Götz for their fruitful work under my supervision.

Research cannot be conducted without economic support. I want to express my gratitude to the European Research Council (ERC), the Deutsche Forschungsgemeinschaft (DFG) and the Technical University of Dortmund for funding my work. Further thanks to the International Max Planck Research School (IMPRS) for funding my attendance to workshops and conferences.

Additional thanks to our collaborator Michal R. Schweiger and her group, to the coordinators of the International Max Planck Research School Christa Hornemann and Lucia Sironi, to the members of my Thesis Advisory Committee; Andrea Musacchio and Christian Schröter and finally, to Stefano Maffini and Leif Dehmelt for their time, expertise and fruitful discussions.

Finally, I want to express my deepest gratitude and appreciation to my girlfriend Anne and to all my family, especially to my brother Eduardo, my mother Carmen and my father Jacinto. I am immensely thankful for the constant support to overcome any adversity and, of course, for the limitless love.

Table of Contents

List of Figures

Main Figures

Supplementary Figures

Abbreviations

5'-azadC	5-aza-2'-Deoxycytidin
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A	Adenine
AD	Activation domain
Ala	Alanine
AM-AR	Active Modification – Active Repair
Arg	Arginine
Asn	Asparagine
Asp	Aspartic Acid
BER	Base excision repair
BFP	Blue Fluorescent Protein
bp	Base pair
BS-seq	Bisulfite sequencing
C	Cytosine
CENP-A	Centromeric Protein A
CRD	Central repeat domain
CTR	C – terminal region
Cys	Cysteine
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DSBs	Double strand breaks
E. coli	*Escherichia coli*
eGFP	Enhanced Green Fluorescent Protein

EMSA	Electromobility shift assay
ES cells	Embryonic Stem cells
FI	Fluorescence Intensity
FISH	Fluorescent *in situ* hybridization
FRET	Förster Resonance Energy Transfer
G	Guanine
G*	Glycine 12 + deletion at position 13 Repeat Variable Diresidue
Gln	Glutamine
Glu	Glutamine
Gly	Glycine
H	Hydrogen
H1	Histone 1
H2A	Histone 2A
H2B	Histone 2B
H3	Histone 3
H4	Histone 4
HATs	Histone acetyltransferases
H-bond	Hydrogen bond
HD	Histidine 12 + Aspartic Acid 13 Repeat Variable Diresidue
HDACs	Histone deacetylases
HDR	Homology Directed Repair
His	Histidine
HP1	Heterochromatin Protein 1
HPCE	High Performance Electrophoresis
HPLC	High Performance Liquid Chromatography
HSF1	Heat Shock Factor 1
Ile	Isoleucine
KO	Knockout
Leu	Leucine

Abbreviations

LINEs	Long Interspersed Nuclear Elements
Lys	Lysine
MBD-seq	Methyl-CpG binding domain protein-enriched sequencing
MeDIP-seq	Methylation-dependent immunoprecipitation sequencing
Met	Methionine
MGEs	Mobile Genetic Elements
MRE-seq	Methylation-sensitive restriction enzyme sequencing
MS	Mass spectrometry
NG	Asparagine 12 + Glycine 13 Repeat Variable Diresidue
NH*	Asparagine 11 + Histidine 12 + deletion at position 13 Repeat Variable Diresidue
NHEJ	Non-homologous End Joining
NI	Asparagine 12 + Isoleucine 13 Repeat Variable Diresidue
NLS	Nuclear localization signal
NN	Asparagine 12 + Asparagine 13 Repeat Variable Diresidue
nSB	Nuclear Stress Body
NTR	N – terminal region
NY*	Asparagine 11 + Tyrosine 12 + deletion at position 13 Repeat Variable Diresidue
PD	Passive dilution
Phe	Phenylalanine
Pro	Proline
PyAOP	(7-Azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate
RNA	Ribonucleic acid
RVD	Repeat Variable Diresidue
SAM	S-adenosylmethionine
Sat2	Satellite 2
Sat3	Satellite 3
SatIII	Satellite 3
Ser	Serine
SINEs	Short Interspersed Nuclear Elements

SNP	Single nucleotide polymorphism
S-phase	DNA Synthesis Phase
SSB	Single stranded DNA binding proteins
SSRs	Simple Sequence Repeats
SV40	Simian vacuolating virus 40
T	Thymine
T3SS	Type III secretion system
TALE	Transcription Activator-Like Effector
TALENs	Transcription Activator-Like Effector Nucleases
TDG	Thymine DNA Glycosylase
TET	Ten-eleven translocation dioxygenase
Thr	Threonine
Trp	Tryptophane
TSS	Transcription Start Site
Tyr	Tyrosine
U	Uracil
Val	Valine
VP16	Viral Protein 16 transcription activator
VP64	For tandem repeats of viral protein 16 transcription activator
WT	Wild type
ZFP	Zinc Finger Proteins
α – Sat	Alpha satellites

1. Abstract

5-methylcytosine (5mC) is a fundamental epigenetic modification in mammalian genomes involved in development, cell differentiation and genomic imprinting. In addition, aberrant DNA methylation patterns are responsible for the pathogenesis of many diseases including neurodegenerative disorders, cardiovascular affections and cancer. In this thesis, we describe the development of a novel method for image-based analysis of 5mC using pairs of fluorescent Transcription Activator Like-Effectors (TALEs). These DNA binding proteins can recognize specific sequences of canonical or epigenetically modified DNA via modular repeats that interact with nucleobases in a one-to-one correspondence. We employed fluorescent TALE pairs that differ only in the repeat responsible for recognizing cytosine (C) at CpG dinucleotides and the fluorophore fused to them (either eGFP or mCherry). By using the 5mC selective repeat HD in one of the TALEs, we can detect differences in methylation level, while the universal binder repeat G* in the other TALE is not responsive to 5mC and allows to detect local changes in chromatin compaction. This way it is possible to analyze 5mC independently of potential differences in target accessibility. We applied our method using recombinantly expressed and purified TALE pairs in cellular stains to image SatIII DNA. This pericentromeric DNA is the origin of nuclear stress bodies (nSBs), exhibits aberrant methylation in several cancers and remains challenging to study by conventional methods due to its highly repetitive nature. We proved the applicability of our method to study 5mC differences in user-defined repetitive sequences with single nucleotide and strand resolution. Furthermore, we correlated the methylation status of SatIII with the presence of heat shock factor 1 (HSF1) at its recognition sequence after stress, revealing a role for 5mC in HSF1 recruitment as initial step of nSB formation in a subpopulation of cells. Finally, we constructed and screened a library of size-reduced TALE repeats to identify potential 5mC binders. We found that RVD NH* binds selectively to 5mC, but not C and its application in combination with HD TALEs allows for improved imaging with higher dynamic range. These studies offer a promising imaging tool for studying 5mC function in repetitive sequences and its interplay with other imageable chromatin-interacting proteins with nucleotide, strand, locus and cell resolution.

1. Zusammenfassung

5-Methylcytosin (5mC) ist eine grundlegende epigenetische Modifikation in Säugetiergenomen, die an der Entwicklung, Zelldifferenzierung und genomischen Prägung beteiligt ist. Darüber hinaus sind aberrante DNA-Methylierungsmuster für die Pathogenese vieler Krankheiten verantwortlich, einschließlich neurodegenerativer Störungen, kardiovaskulärer Erkrankungen und Krebs. In dieser Arbeit beschreiben wir die Entwicklung einer neuartigen Methode zur bildbasierten Analyse von 5mC unter Verwendung von Paaren fluoreszenzmarkierter „Transcription Activator Like-Effectors" (TALEs). Diese DNA bindenden Proteine können spezifische Sequenzen kanonischer oder epigenetisch modifizierter DNA anhand modularer Wiederholungseinheiten erkennen, welche mit den Nukleobasen in einer eins-zu-eins-Korrespondenz interagieren. Wir verwendeten fluoreszierende TALE-Paare, die sich nur in der Wiederholungseinheit unterscheiden, welche für die Erkennung von Cytosin (C) innerhalb von CpG Dinukleotiden verantwortlich ist, und dem mit ihnen fusionierten Fluorophor (entweder eGFP oder mCherry). Durch Verwendung der 5mC-sensitiven Wiederholungseinheit HD in einem der TALEs können Unterschiede im Methylierungsgrad festgestellt werden, während die universelle Wiederholungseinheit G* des anderen TALEs nicht empfänglich 5mC gegenüber ist und somit die Detektion lokaler Veränderungen der Chromatinkompaktierung erlaubt. Auf diese Weise ist es möglich 5mC unabhängig von möglichen Unterschieden in der Zielsequenzzugänglichkeit zu analysierten. Wir nutzten unsere Methode mit rekombinant exprimierten und gereinigten TALE-Paaren in Zellfärbungen zur Abbildung von SatIII DNA. Diese perizentromerische DNA ist Ursprung nukleärer Stresskörper (nSBs), weist bei mehreren Krebsarten eine aberrante Methylierung auf und ist aufgrund ihrer sich stark wiederholenden Natur nach wie vor schwierig mit herkömmlichen Methoden zu untersuchen. Wir haben die Anwendbarkeit unserer Methode zur Untersuchung von 5mC-Unterschieden in benutzerdefinierten repetitiven Sequenzen mit Einzelnukleotid- und Strangauflösung bewiesen. Darüber hinaus korrelierten wir den Methylierungsstatus von SatIII mit dem Vorhandensein von Hitzeschockfaktor 1 (HSF1) an seiner Erkennungssequenz nach Stress, was eine Rolle für 5mC bei der HSF1-Rekrutierung als ersten Schritt der nSB-Bildung in einer Unterpopulation von Zellen enthüllte. Schließlich haben wir eine Bibliothek mit verkleinerten TALE-Wiederholungseinheiten erstellt und getestet, um potenzielle 5mC-Binder zu identifizieren. Wir fanden heraus, dass RVD NH* selektiv an 5mC bindet, nicht jedoch an C, und dass seine Anwendung in Kombination mit HD TALEs eine verbesserte Bildgebung mit höherem Dynamikbereich ermöglicht. Diese Studien bieten ein vielversprechendes Bildgebungsinstrument zur Untersuchung der 5mC-Funktion in repetitiven Sequenzen und ihres Zusammenspiels mit anderen abbildbaren Chromatin-interagierenden Proteinen mit Nukleotid-, Strang-, Positions- und Zellauflösung.

2. Introduction

2.1 The Composition and Organization of the Eukaryotic Genome

2.1.1 Genetic Information is stored on deoxyribonucleic acid molecules

From the most inhospitable deserts to the vast Amazon rainforest, our planet is full of life. Each species is different and although members of the same species have common characteristics, each individual is unique. Every living organism contains a set of essential instructions to define itself, develop, function and reproduce. This set of instructions is what we call Genetic Information.

There are more than 10 million of different living species populating the Earth and, despite the stunning diversity, all of them store their Genetic Information in the same way; double-stranded molecules of a long polymer called deoxyribonucleic acid (DNA) (*1*, *3*). DNA is composed by linear strands of four different types of monomers called nucleotides, which consist of the sugar deoxyribose attached to one or more phosphate groups and a nitrogen-containing base (*4*). Four different types of bases (also known as nucleobases) can be attached to the deoxyribose, resulting in the four different nucleotides. Two of the possible nucleobases, cytosine (C) and thymine (T), are derivatives of the six-membered heterocyclic compound pyrimidine (**Figure 1a**). The other two, adenine (A) and guanine (G) are derivatives of purine, a two ring-containing heterocyclic compound (**Figure 1b**) (*5*, *6*).

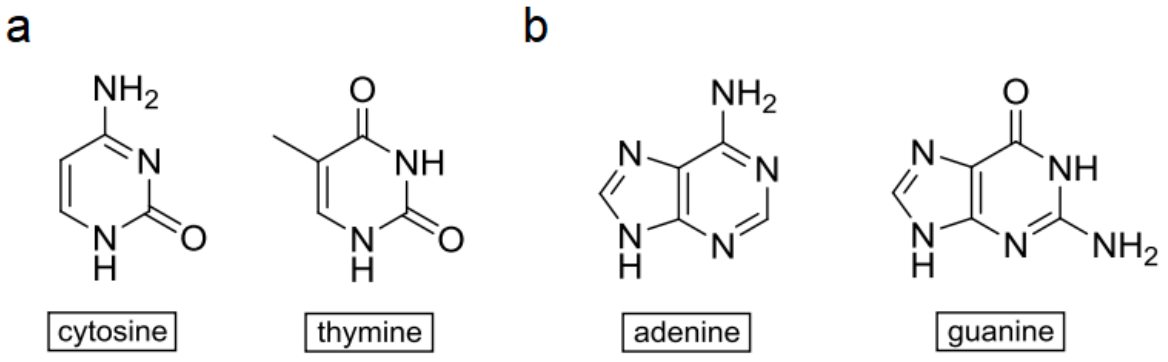


Figure 1. Molecular structure of the four nitrogen-containing bases that constitute DNA. a) Chemical structure of pyrimidine derived nucleobases cytosine and thymine. b) Chemical structure of purine derived nucleobases adenine and guanine.

Nucleotides are covalently linked together by a phosphodiester bond between the phosphate group attached to the 5' carbon of the deoxyribose of a nucleotide and the 3' carbon of the sugar of the following one (**Figure 2**). Each sugar is linked to the next via the phosphate group, creating a polymer chain composed of a repetitive sugar-phosphate backbone with a series of nucleobases protruding from it (*6*). This linear sequence of nucleotides is encoding the genetic information and can be read, interpreted and replicated by the cellular machinery (*1*). The sugar-phosphate units are asymmetric giving the backbone a definite directionality with two clearly distinguishable termini: a 5' phosphate end and a 3' hydroxyl end (*4*).

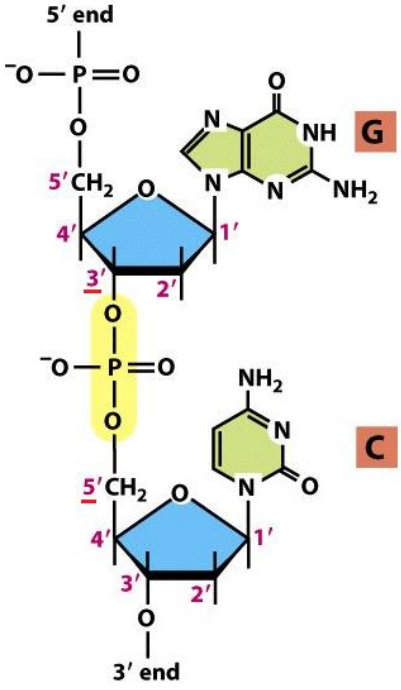


Figure 2. Schematic representation of a phosphodiester bond between two nucleotides. Adapted from (*1*).

In nature, DNA is present as dimers of two linear strands which are held together in an antiparallel orientation by hydrogen bonds (H-bonds) between the bases on different strands (**Figure 3a**). Due to the chemical structure of the bases, only interactions between G and C (three H-bonds, **Figure 3b**) and between A and T (two H-bonds, **Figure 3c**) are energetically favorable and therefore allowed to confer stability to the dimer (*4*). These interactions are responsible for the three-dimensional structure of DNA: a right-handed double helix in which each turn is composed by 10.4 base pairs (bp, **Figure 3d**). In this double helix, the bases are packed in the inner side (**Figure 3e**), while the sugar-phosphate backbone are positioned on the outside leading to an overall negatively charged surface. In addition, the coiling of the two strands around each other creates two grooves of different sizes known as major and minor grooves (**Figure 3f**) (*4*). Both provide different and unique chemical information because the edge of each base pair presents a distinctive pattern of H-bond donors, H-bond acceptors, and hydrophobic patches. The information displayed in the major groove can be read by transcription factors, a specific type of proteins that plays an essential role in control of gene expression (*7*).

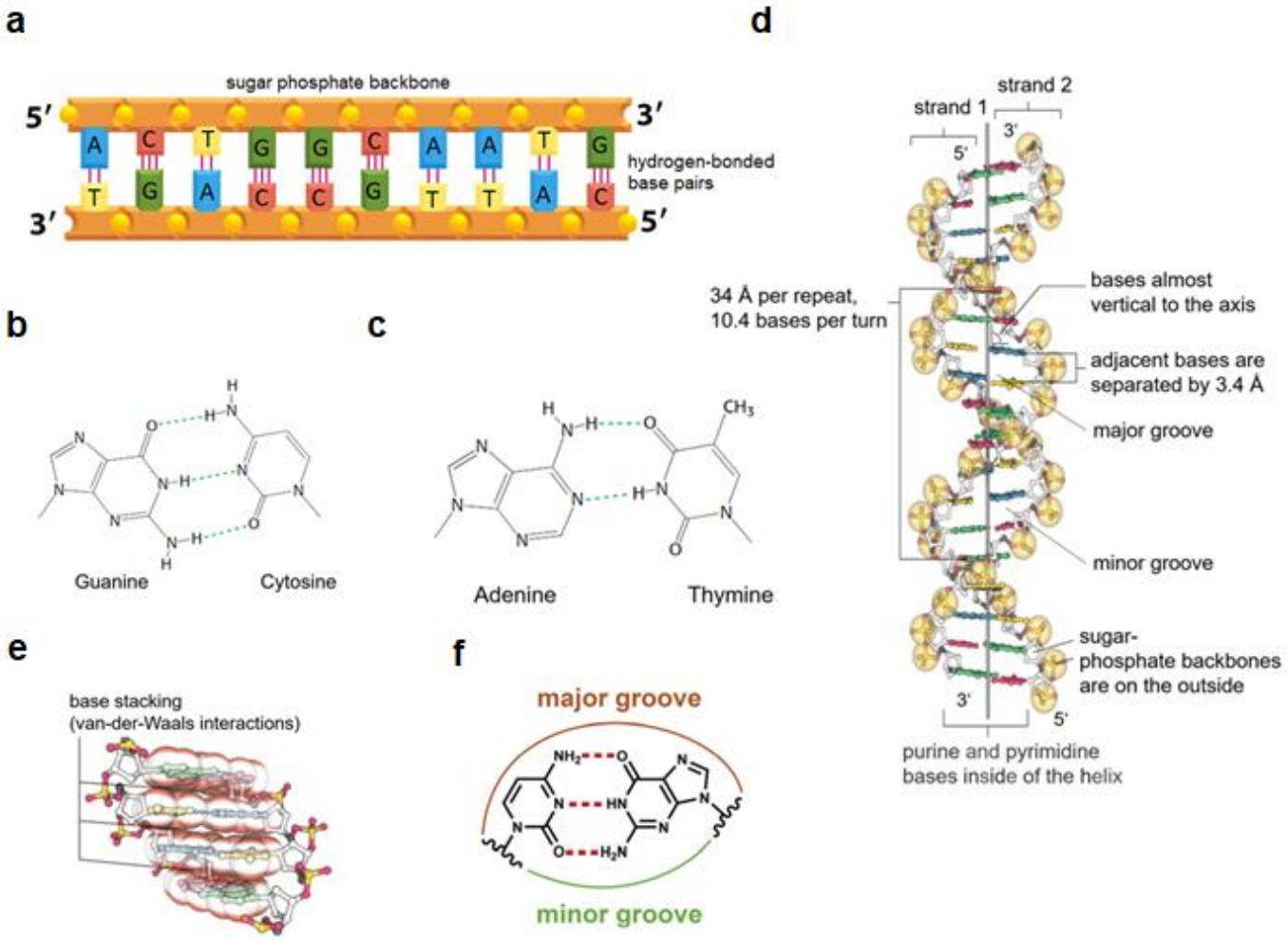


Figure 3. Structure of DNA. a) Schematic representation of the linear structure of DNA with antiparallel orientation of the two strands. Hydrogen bonds between base pairs are represented as red lines. **b)** Schematic representation of the three hydrogen bonds established between G and C. **c)** Schematic representation of the two hydrogen bonds formed between A and T. **d)** Three-dimensional structure of the DNA, a right-handed double helix in which the base pairs (represented in blue, yellow, green and red) stacked in the inner side and the sugar-phosphate backbone (yellow balls) remains outside providing negative charge. **e)** Detailed view of the base stacking. **f)** Chemical information displayed by the major and minor groove. Adapted from (*1*) and (*8*).

The elucidation of the structure of DNA led to hypothetical answers to some of its most fundamental functions. On one hand, the mechanism of heredity, how the genetic material can be replicated and transmitted to the progeny (*4*). On the other hand, how the DNA could store and encode the instructions needed to produce every RNA and every protein within cells (*9*).

2.1.2 DNA replication

The discovery of the structure of DNA with its specific base-pairing immediately suggested a possible copying mechanism for the genetic material. Since each strand contains a nucleotide sequence which is exactly complementary to the other strand, each of them acts as template to generate a new complementary strand (*4*). DNA replication is a tightly controlled and regulated process occurring at a specific time during the cell cycle, the so-called DNA Synthesis phase (S-phase) (*10, 11*). The process starts at the same time at multiple origins of replication, A-T-rich sequences located at many different positions in the genome (*12*). Multiple proteins with different functions are recruited at these sites to form a multienzyme complex known generically as replication machinery (*1*). Firstly, DNA helicases hydrolyze ATP to open the double helix by breaking the H-bonds between complementary base pairs generating a replication fork. Then, Single-strand DNA-binding (SSB) proteins bind tightly to the exposed single-stranded DNA to stabilize it without covering the bases, so it can be used as template by DNA polymerase (*1*). This enzyme can read the bases in the single-stranded DNA and add complementary bases to generate an entire new complementary strand which will then pair with the one used as a template (*10, 11*). However, DNA polymerase cannot synthesize DNA *de novo*. The enzymatic addition of nucleotides by this enzyme requires a free 3'-OH group, this means that it is not able to create a completely new strand, but only adding nucleotides at the 3' end of an existing one (*10*). To solve this, cells count on an enzyme called DNA primase, which uses ribonucleoside triphosphates to synthesize short RNA primers that provide the hydroxyl group for DNA polymerases to incorporate new nucleotides (*13–15*). This requirement of a 3'-OH group involves that polymerization can only proceed in 5' to 3' direction. Consequently, the two newly generated strands at the replication fork are synthesized at different rate because of the antiparallel orientation of DNA. One of the strands, the leading strand, can be synthesized continuously and therefore faster because the replication fork orientation is favorable. The other one, the lagging strand, is synthesized discontinuously as short fragments known as Okazaki fragments and requires of multiple RNA primers in order to proceed in the appropriate direction (*15, 16*). After the synthesis of each fragment is finished, RNA primers are removed and substituted by DNA. Finally, Okazaki fragments are joined into a long and continuous DNA chain by DNA ligase. (*15, 17*) Once the complete DNA is copied

and the cell divides, each of the daughter cells will have a genome composed by an original template strand and a newly synthesized strand. This is what we know as semiconservative replication of DNA (*18*).

2.1.3 Gene Expression. The interpretation of the Genetic Information

The genetic information contained in DNA is the set of instructions to build up all the proteins needed by the cell. However, the DNA does not direct protein synthesis itself, it requires the use of RNA as an intermediary (*9, 19*). The specific nucleotide sequence encoding a protein is firstly used as a template to produce an RNA molecule, known as messenger RNA (mRNA), containing the same information in a process called **transcription (*1*)**. Then, this mRNA undergoes several processing steps in the nucleus, including RNA splicing and finally it is transported to the cytosol. There, it associates with ribonucleoprotein complexes and is used as template to produce the final protein. This process of converting the information carried by an mRNA into a protein is called **translation (*1*)**. The genetic information always (with few exceptions) flows this way: from DNA to RNA and from RNA to protein in every living cell, a principle so fundamental that it is known as the *central dogma of molecular biology (1).*

2.1.3.1 Transcription. From DNA to RNA

Transcription is the process by which a gene sequence encoded in a strand of DNA is copied into a complementary RNA molecule. These are also polynucleotide chains which differ from DNA in containing ribose as sugar (**Figure 4a**) and replacing T by the nitrogenated base uracil (U, **Figure 4b**) (*9*). RNAs are classified into different groups according to their function. Messenger RNAs (mRNAs) contain the instructions to produce a protein, ribosomal RNAs (rRNAs) associate to specific proteins to form ribosomes and catalyze protein synthesis and transfer RNAs (tRNAs) carry the appropriate amino acids and hold them in place for incorporation into proteins. Small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) play a fundamental role in post-transcriptional modification and processing of RNA. Several other types of RNAs such as microRNAS (miRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs) and long noncoding RNAs (lncRNAs) are involved in the control of gene expression and other cell processes (*1*).

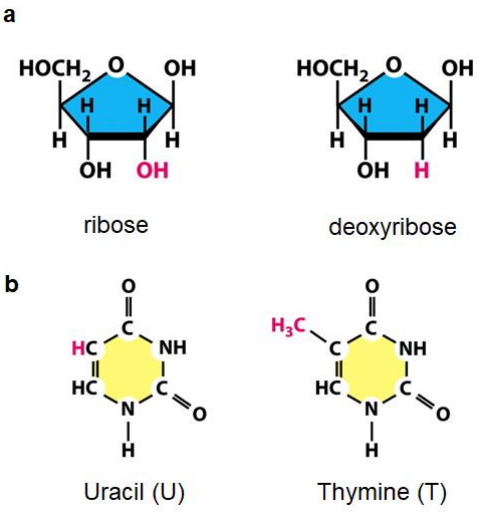


Figure 4. Differences in the constituents of DNA and RNA. a) Chemical structures of the sugar ribose and deoxyribose that are present in RNA and DNA, respectively. **b)** Chemical structure of the nitrogen-containing bases Uracil and Thymine that are part of RNA and DNA, respectively. Adapted from (*1*).

Transcription is triggered by the binding of specific DNA binding proteins called transcription factors to the promoter and regulatory sequences of the genes. This leads to the recruitment of a set of general transcription factors to the TATA box, an A-T rich sequence near the transcription start site (TSS). These factors recruit RNA polymerase II at the TSS, stabilize it and open up the double helix to make it accessible (*1, 20, 21*). Shortly after initiation, RNA polymerase II is phosphorylated and undergoes a series of conformational changes that enhance its interaction with DNA. This allows it to enter in the elongation phase and transcribe long RNA molecules without dissociating from DNA. During this phase, most of the general transcription factors separate from the initiation complex and they are substituted by elongation factors. In addition, chromatin remodeling complexes and histone-modifying enzymes are also recruited to facilitate access to DNA during transcription (*20, 21*). As the nascent mRNA molecule is growing, proteins involved in RNA processing bound to it and modify the 5'-end by the addition of a methyl-guanosine cap (**Figure 5**). At the same time as the RNA is synthesized, introns are removed by a process known as RNA splicing. Once the mRNA is completely synthesized, the 3'-end is modified by the enzymatic addition of a poly-A tail. This reaction is catalyzed by poly-A polymerase (PAP), a template independent polymerase which adds approximately 200 A to the 3'-end. Finally, mature mRNAs are selectively exported from the nucleus to the cytosol, where they will associate with ribosomes to be translated into proteins (*20*).

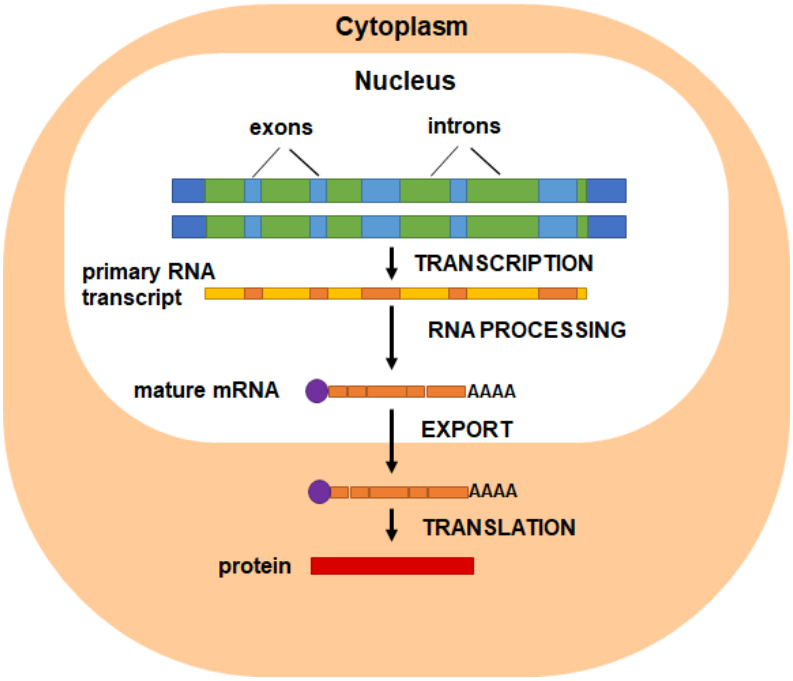


Figure 5. Overview of the Transcription process. DNA is transcribed into long mRNA transcripts by RNA polymerase (not shown). Simultaneously, mRNA molecules undergo maturation, a process in which introns are removed by RNA splicing, the 5' end is protected by a methyl-guanosine cap and the 3'end is polyadenylated. Mature mRNAs are exported to the cytosol where translation takes place. Adapted from (*1*).

2.1.3.2 Translation. From mRNA to protein

Many mRNA molecules are produced from the same gene simultaneously, so it is possible to generate large amounts of a specific protein when needed. RNA is essentially "written" in the same code as DNA (only with U substituting T) and therefore the coding sequence is not altered. However, converting the information encoded in a mRNA molecule into a fully functional protein requires a specific correlation between the linear sequence of the four different nucleotides present in RNA and the 20 canonical amino acids that constitute proteins. RNA is read as triplets or codons, subsets of three consecutive nucleotides (*22*). Each codon specifies either one amino acid or a stop signal to terminate translation. There are 64 possible combinations of groups of three nucleotides and thus, 64 different codons (**Figure 6**). Since only 20 amino acids are commonly found in proteins, this implies that each amino acid could be specified by more than one codon. This property of the genetic code is known as redundancy. Some amino acids like arginine (Arg), leucine (Leu) or serine (Ser) are coded by six different codons, while others like methionine (Met) or tryptophan (Trp) are coded by only one codon. Three special codons: UAA, UAG and UGA do not correspond to any amino acid, but to stop signals which indicate the end of the protein and stop translation (*23–26*). Strikingly,

the genetic code is universal, the same set of codons specifies the same amino acid in all living organisms (with few exceptions) (*1*).

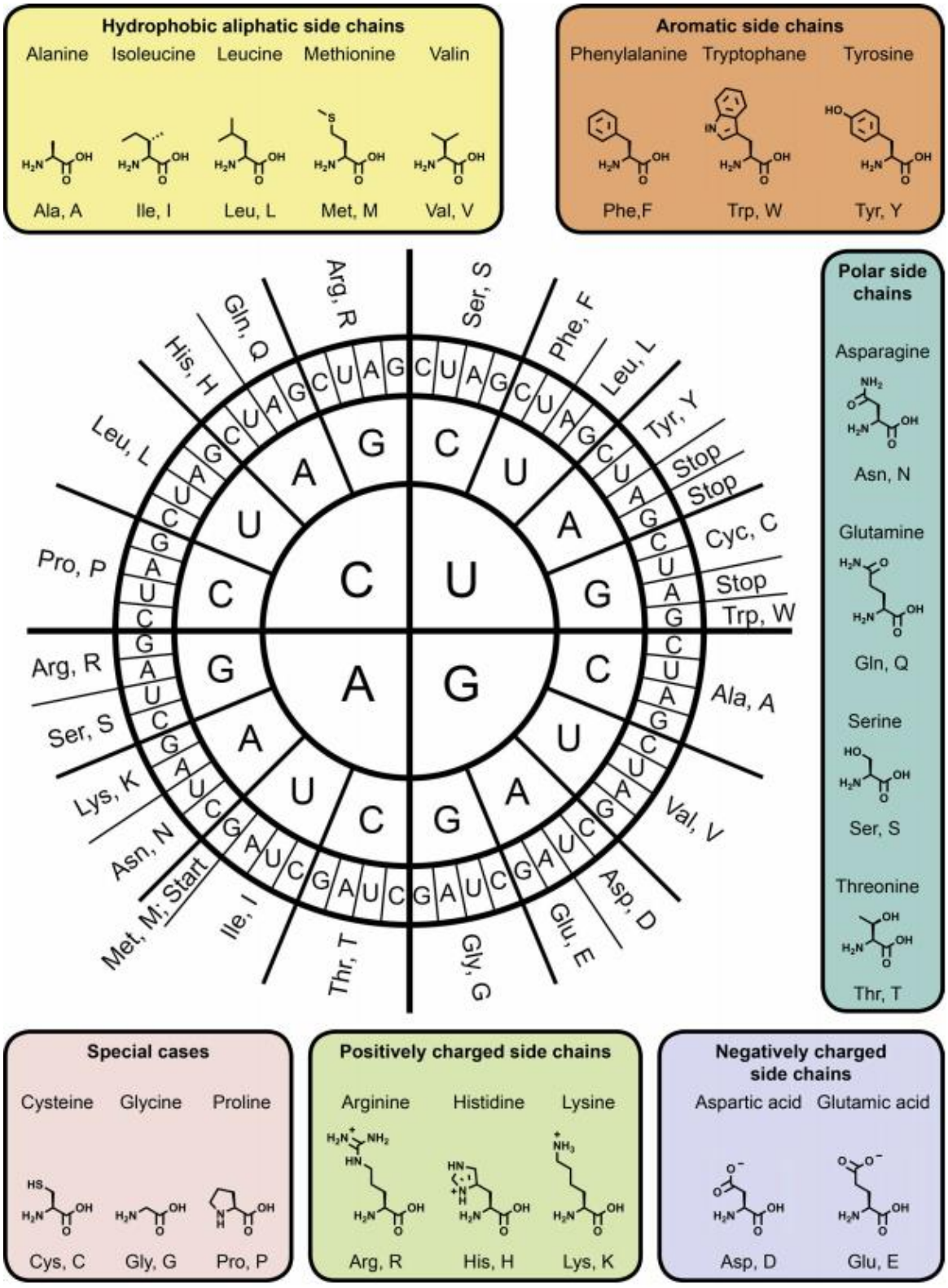


Figure 6. Genetic Code and amino acid structures. The Genetic Code expresses the correspondence between three consecutive RNA nucleotides and one of the twenty canonical amino acids. Codon sequences are read from

the inner side of the diagram (1st nucleotide) to the outer side (3rd nucleotide). Both three letter and one letter codes for each amino acid are shown. Amino acids are grouped according to their chemical properties.

Translation takes place in the cytosol, where mRNAs are decoded by ribosomes, ribonucleoprotein complexes composed by more than 50 different proteins and rRNAs. Ribosomes bind to mRNAs and search for the first start codon (AUG), which is recognized by a unique initiator tRNA. This type of RNA, composed by around 80 nucleotides, has a three-dimensional structure with three loops that resembles a cloverleaf (**Figure 7a**). The middle loop, known as the anticodon loop, recognizes a specific codon in the mRNA by complementary base-pair interactions. The 3'-end of each tRNA is loaded with the specific amino acid which corresponds to the codon recognized by the tRNA (*27*). The ribosome pulls through the mRNA reading its codons with the help of aminoacyl-tRNAs and uses a peptidyl transferase to transfer the growing polypeptide chain to the amino acid of the next tRNA by a peptide bond using GTP as energy source (**Figure 7b**). The protein synthesis will continue until the ribosome finds a stop codon in the mRNA sequence. These codons are recognized by a release factor, which forces the peptidyl transferase to catalyze the addition of a water molecule to the last peptidyl-tRNA releasing the polypeptide chain. Additional proteins contribute to disassemble the ribosome and release the mRNA (*28*). The newly synthesized protein can undergo further processing, post-translational modifications and it is exported to its final cellular destination (*27*).

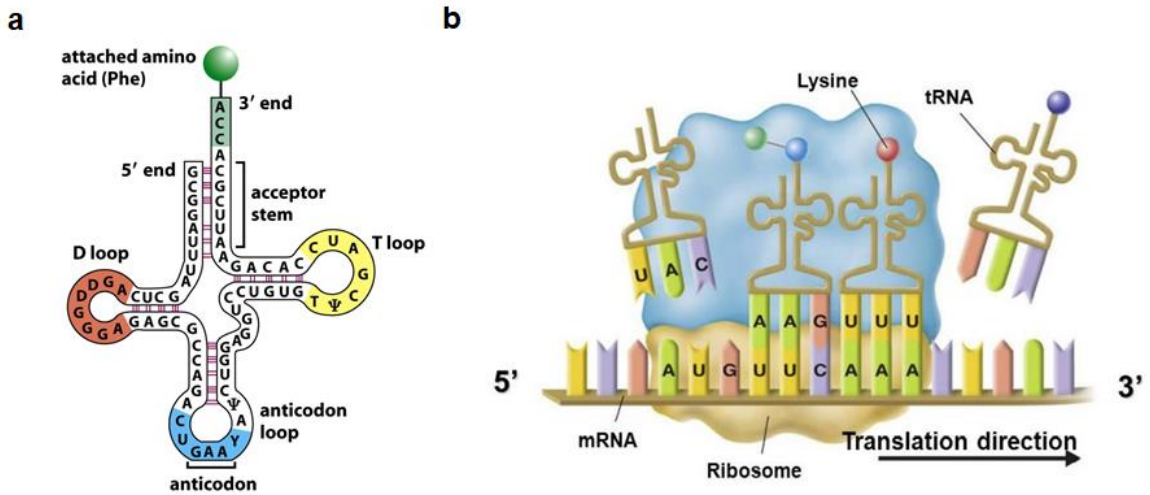


Figure 7. a) Schematic structure of a tRNA. The approximately 80 bp that compose tRNAs fold generating a cloverleaf-like structure with three loops: a D-loop (red), a T loop (yellow) and the anticodon loop (blue). The

anticodon is responsible for recognition of the complementary codon in the mRNA. The corresponding amino acid is loaded on to the 3'end of the tRNA by the enzyme aminoacyl tRNA synthetase. **b)** Scheme depicting protein synthesis, a process known as Translation. Ribosome subunits associate with a mRNA molecule and aminoacyl-tRNAs pairs to direct protein synthesis by sequential addition of amino acids until a stop codon is found. Adapted from (*1*).

2.1.4 Organization of the DNA in the Eukaryotic nucleus

The nuclear DNA of eukaryotes is organized in separate units known as chromosomes. The human genome is composed by 23 pairs of chromosomes, of which two of them (X and Y) are known as sex chromosomes and are responsible for sexual differentiation. The rest of them are called autologous chromosomes (*1*).

The approximately 3 billion of base pairs that composed the human genome would reach a length of 2 meters if we could arrange them in a linear fashion, and they have to fit into a nucleus of about only 6 μm in diameter (*1*). This implies that DNA must be tightly packed and still remain accessible to transcription factors, regulators, replication machinery and many other interacting proteins (*29*). The most basic level of chromosome packing is called **nucleosome** and consists of an octamer of proteins tightly bound to DNA (**Figure 8a**) (*30*). Each nucleosome is composed of eight histone proteins, containing pairs of histones H2A, H2B, H3 and H4, associated to 147 bp of DNA. The space between two nucleosomes can be variable, but on average they are separated by 200 bp generating a structure which resembles "beads on a string" (**Figure 8b**) (*30*). This complex of DNA and tightly bound histone proteins is known as chromatin. DNA is further compacted with the aid of H1 histones, stacking the nucleosomes to form a 30-nm chromatin fiber (*31*).

Different chromatin states with variable levels of DNA condensation can be found in the interphase nuclei. The highly condensed form, known as **heterochromatin**, is highly concentrated in specialized regions of chromosomes (like telomeres and centromeres) and generally associates to inactive genes. The rest of the chromatin, called **euchromatin**, is lightly packed and therefore more accessible facilitating gene transcription (*1, 32*).

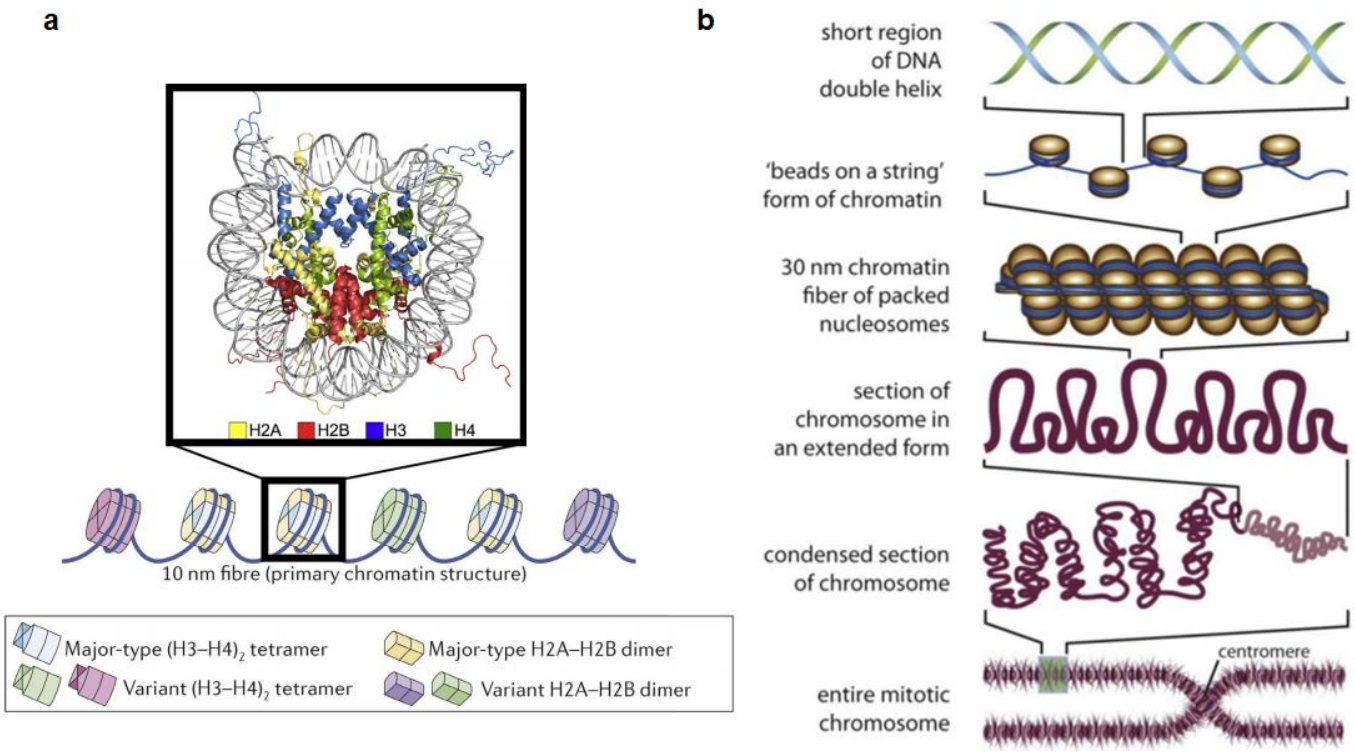


Figure 8. Organization of DNA in the eukaryotic nucleus. a) Top: Crystal structure of the nucleosome core particle wrapped by DNA helix (white). Nucleosome are composed by dimers of four different type of histone proteins: (blue: H3, green: H4, yellow: H2A, red: H2B). Bottom: Schematic representation of nucleosomes in the primary chromatin structure. **b)** DNA packaging levels. Different levels of chromatin condensation promoted by nucleosomes and architectural proteins. Adapted from (*30*) and (*33*).

Heterochromatin is typically associated to histones with specific chemical modifications and to specialized proteins, like Heterochromatin Protein 1 (HP1), that modify the local status of the chromatin to confer a denser structure (*34*). An important part of the heterochromatin is constituted by two specialized DNA regions present in every chromosome in human genome: telomeres and centromeres (*1*). **Telomeres** are tandemly repeated DNA sequences (GGGTTA) located at the end of chromosomes whose main function is preventing genetic information loss during replication (*35*). DNA polymerase cannot finalize the DNA replication at the end of the lagging strand due to the linear nature of human chromosomes and the inability to replace the very last RNA primer. As a consequence, telomeres are progressively getting shorter after each cell division, a usual marker of aging. To overcome this problem, a specialized enzyme called telomerase binds to the tips of the chromosomes and replenishes the telomeric sequence (*35*). In addition, telomeres are associating with a protein complex known as Shelterin that is

responsible for the characteristic T-loop structure of the chromosome ends (*35*). The other specialized regions, the **centromeres**, are constituted by hundreds of thousands of nucleotide pairs (*1*). They mainly contain different variants of short, repeated DNA sequences known as satellite DNA. The centromeric heterochromatin is characterized by presenting a centromere-specific variant H3 histone known as Centromere Protein A (CENP-A), plus additional proteins that pack nucleosomes in a particularly condensed manner (*36*). Centromeres are essential for the assembly of the kinetochore, a specialized structure required for the attachment of the mitotic spindle, whose function is segregating chromosomes during cell division (*36*). Apart from telomeres and centromeres, heterochromatin is present at many other locations along the genome, accounting for more than 10% of the total genetic material (*1*). In addition, these regions that are highly compacted as heterochromatin can dynamically vary to satisfy the needs of the cells according to their physiological state (*1*).

2.1.5 Arrangement and Composition of DNA elements in the Human Genome

The multiple massive sequencing projects of the human genome over the last three decades have allowed us to elucidate the sequence of most of the 3.2 billion nucleotide pairs that compose our genetic material (*37, 38*). Further analysis and characterization of the different DNA elements revealed unexpected properties of the functional composition of our genome. All the instructions to build proteins, the so-called protein-coding genes, are contained in just 1.5% of the total DNA (**Figure 9**) (*39*). This is particularly striking considering that previously sequenced genomes from other organism have a much higher content of coding DNA. For instance, 89 % of the genome from the model bacteria *Escherichia coli* (*E. coli*) code for proteins by contrast to the less than 2% in humans (*1*). Since the solely function of the DNA was long thought to be storing the genetic information in the form of instructions for synthesizing proteins, the remaining 98% of the human genome was long considered to be just "junk" DNA (*40*).

We are just starting to unveil the importance of the major components different from protein-coding genes present in our genetic material (*40*). Human genes are composed by the alternation of protein-coding sequences, called exons, and long non-coding regions known as

introns. Introns, which are transcribed and then spliced out from the mRNA, constitute about 26% of the genome (*1*, *41*). Other non-coding sequences closely related to genes include large regulatory regions like promoters and enhancers which are usually located upstream of the gene they control (*41*).

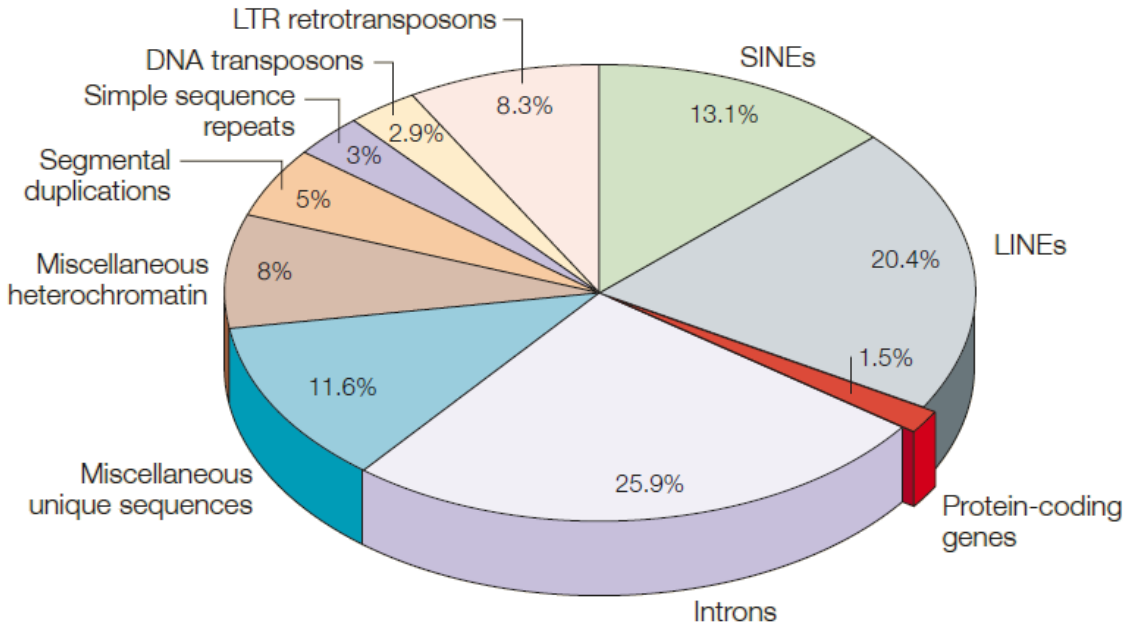


Figure 9. Functional composition of the Human Genome. Only 1.5% of the Human Genome code for proteins, while the non-coding regions of genes, the introns, constitute ~26% of the whole DNA content. Strikingly, a large majority of the Genome (>50%) is composed by repetitive elements of diverse nature. Adapted from (*42*).

Other sequences similar to protein-coding genes contribute to the 98% of non-coding DNA. Apart from the 21000 genes which encode for proteins, there are around 9000 **non-coding genes** whose final product is an RNA itself which can have structural, regulatory or even enzymatic functions (*1*, *41*). In addition, there are more than 20000 **pseudogenes**, DNA sequences which closely resemble to a functional gene, but containing numerous mutations that hamper their proper expression (*1*).

Altogether all these DNA regions mentioned above account for nearly half of the genome (*1*, *42*, *43*). Strikingly, the other more than 50% is entirely composed by repetitive sequences of diverse nature. Some of them like the long interspersed nuclear elements (LINEs), the short interspersed nuclear elements (SINEs), retroviral-like elements and DNA-only transposon are mobile genetic elements (MGEs) (*42*). These mobile elements have multiplied and spread in

our genome by replicating themselves and inserting the new copies at different locations. Other repeated sequences are segmental duplications which are long stretches of DNA (1000 – 200000 bp) that are present at least twice in the genome. Finally, simple sequence repeats (SSRs) are short nucleotide sequences that are tandemly repeated comprising large blocks of DNA (*42*). Within this group we find microsatellites, minisatellites and satellites like those presents in telomeres and centromeres. Satellite DNA arrays are most abundantly found at centromeric and pericentromeric regions (*43*). Due to the current limitations of genome sequencing and the challenge of mapping and assembling long tandem repeats, these regions are largely uncharacterized. It is estimated that DNA sequences at these locations that remain unknown account for 10% of the genome (*44*). Therefore, methods to study centromeres and pericentromeric regions in depth are needed.

Pericentromeric regions are mainly composed by three families of satellite DNAs: alpha satellite (α-sat), satellite 2 (Sat2) and satellite 3 (Sat3). These families are also the most abundant satellites in the whole genome and their repetitive sequences seem to have some degree of heterogeneity, presenting sequence variability (*45*). In addition, the copy number and repeat structure of the different satellite families in centromeric and pericentromeric regions seems to vary extensively between individuals (*45*). These differences, together with dysregulation in expression, are suggested to be involved in oncogenesis (*46–52*) and infertility (*53*). α-Sats, which are highly divergent AT-rich repeats of around 171 bp, constitute a median of 3.1% of the genome, ranging between 1% and 5% based on data from the 1000 Genome Consortium (*45, 54*). Human Sat2 and Sat3 are generally defined as pentameric repeats (GAATG)n (*55*) and the median abundance in the genome is 2.1% with a range between 1% and 7% depending on individuals (*45, 56*). In addition, Sat2 and Sat3 are the major components of the constitutive heterochromatin at pericentromeric regions and their dysregulation lead to aberrant overexpression of lncRNAs with suspected implications in several cancers (*50, 57*) and Parkinson Disease (*56*). Therefore, it is highly relevant to develop methods to study the control of expression and copy number variation of these largely uncharted regions of the genome (*58*).

2.2 Control of Gene Expression

The nearly 200 different cell types that constitute the human body differ drastically in function, structure and morphology (*1*). However, all of them have the same genetic information, containing approximately 21.000 protein-coding genes (*41*). The reason why this is possible was long debated and two main hypotheses were suggested. One of the hypotheses proposed that only the zygote and germ cells contain the complete genetic information, while mature cells lose the genes that are irrelevant to their specific cell type during the differentiation process. The other hypothesis supported that every cell retains the whole genome, but some genes become selectively inaccessible during differentiation. Therefore, the different cell types would arise from which set of genes are accessible and expressed, giving a unique expression pattern for each specific cell type (*59*). A set of classical experiments in frogs conducted in the late 50's by James Gurdon proved the second hypothesis to be the right one (*60*). In these experiments, an isolated nucleus from a fully differentiated frog cell was transferred into an enucleated egg cell, and this fully differentiated cells was capable of direct the development of a complete and normal tadpole (**Figure 10**). This means that a fully differentiated cell retains absolutely all the genetic information needed to develop a complete multicellular organism (*60*). Therefore, the strategy to have multiple different cell types with the same genome relies on the control of the gene expression. Cells have mechanisms to strictly control which genes are activated and when. The specific set of active genes defines the phenotype giving rise to each different cell type.

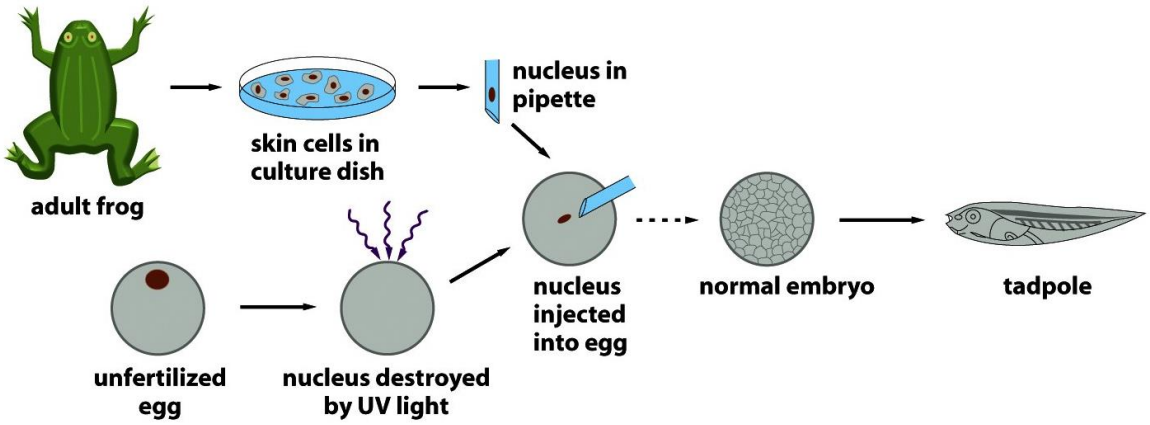


Figure 10. Schematic representation of the experiment performed by James Gurdon. The nucleus of a fully differentiated skin cell was transferred into an enucleated egg cell. This fully differentiated nucleus was able to dedifferentiate and direct the development of a complete and normal tadpole, confirming that every cell in a multicellular organism retains the whole Genetic Information. Adapted from (*1*).

Gene expression can be controlled at many levels in the long pathway that leads from DNA to protein (**Figure 11**). A cell can control the protein synthesis by regulating when and how often a gene is transcribed, it can control the maturation process of the mRNA as well as its transport from the nucleus to the cytosol. There, it can regulate the rate of degradation of the mRNA or the rate of translation to control the final amount of protein. Finally, even when the protein has already been produced, its function can be regulated by further posttranslational modifications that will make it active. From all of them, transcriptional control is the most common and extended mechanism used by cells to control gene expression (*1*). Transcription is controlled mainly by regulating DNA accessibility and chromatin state via chemical modification of DNA and the associated histone proteins, a mechanism known as **Epigenetic regulation** (*32*). Promoters and regulatory sequences of genes that must be repressed are hardly accessible, while the chromatin state of the active regions is loose, facilitating the binding of transcription factors. Interestingly, the resulting chromatin states derived from epigenetic regulation, either repressive or permissive, are heritable even though the modifications do not involve changes in the genetic sequence (*32*). This way, specialized cells can maintain their identity after cell division because they inherit the specific expression patterns through epigenetic modifications (*59*).

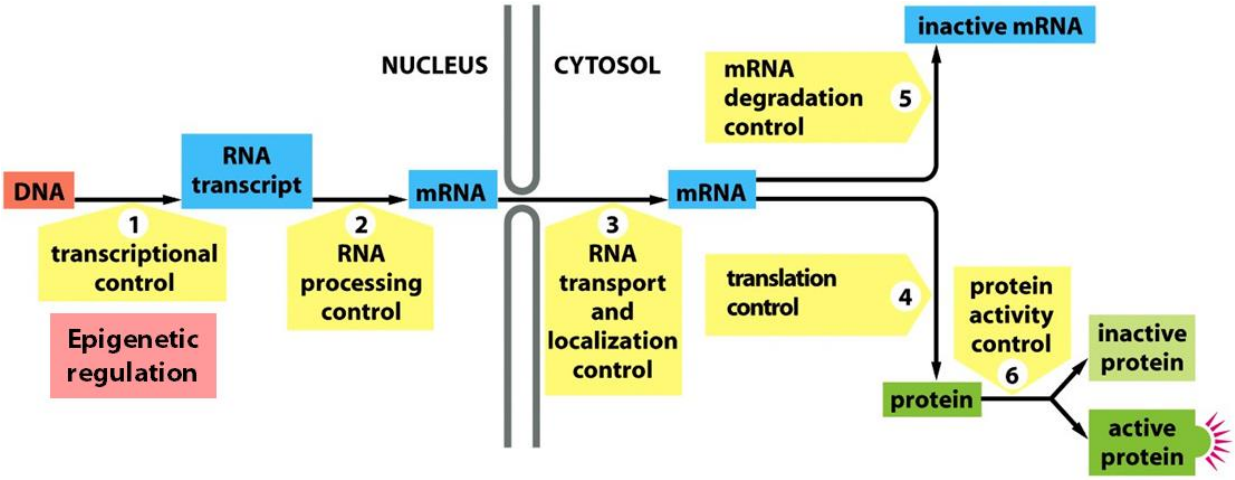


Figure 11. Levels of control of gene expression. Gene expression can be regulated at many levels in the pathway from DNA to protein: by controlling transcription (1), RNA maturation (2), the selective transport to cytosol (3) and then the rate of translation (4) or the rate of mRNA degradation (5). Even when the protein has already been synthesized, its activity can be controlled by additional post-translational modifications (6). Transcriptional control (1) is the most extended mechanism of control of gene expression, an effect mainly achieved by Epigenetic regulation. Adapted from (*1*).

2.2.1 Epigenetic regulation

Epigenetics is defined as the study of the mechanisms of spatiotemporal control of gene expression that involve heritable changes without alteration of the genetic sequence. Epigenetic modifications provide an additional layer of information superimposed on DNA (*epi*, from ancient Greek "on top of") that regulates transcription by shaping chromatin structure (*1, 61*). The different chromatin states are associated to specific sets of genes, conferring a unique signature of expression that defines the phenotype of each specialized cell type. In addition, the heritability of these signatures provides an "epigenetic memory" that allow cells to perpetuate their specialization after cell division (*62*). This is achieved mainly thanks to two mechanisms of epigenetic regulation: histone modification and DNA methylation (*32*). Both mechanisms contribute to modify chromatin state and accessibility, switching from transcriptionally active euchromatin to transcriptionally inactive heterochromatin and vice versa.

2.2.1.1 Histone modification

The four pairs of histones that compose the nucleosome core particles (H2A, H2B, H3 and H4) are subjected to post-translational chemical modifications (*63*). These covalent modifications occur mainly on the "histone tails", the unstructured N-terminal regions that protrude from nucleosomes. More than 30 residues from the histone tails can be either phosphorylated, methylated or acetylated generating a complex code of modifications that contribute together to shape the chromatin structure (**Figure 12**) (*63, 64*). In addition, more than 20 specific side-chain modifications on the nucleosome's globular core have been reported (*63*). All of the above mentioned types of modifications are reversible and each of them are generated or removed by specific enzymes (*64*). For example, there is a set of enzymes, known as acetyl transferases (HATs) that are involved in the addition of acetyl groups. Histone acetylation, like the one that takes place on the ninth lysine of H3 (H3K9ac), loosens chromatin structure by neutralizing the positive charge of the nucleosome. This leads to lower affinity of the histone tails for adjacent nucleosomes. The chromatin structure can be reverted to a more condensed state by removal of acetyl groups by histone deacetylase complexes (HDACs) (*64*). Histone methylation by histone methylases also contributes to chromatin condensation. Some residues from the histone

tails can be mono-, di- and trimethylated leading to the recruitment of specific proteins. For example, trimethylation of lysine nine from H3 (H3K9me3) attracts the heterochromatin – specific protein HP1 and contributes to the establishment and spread of heterochromatin at the recruited sites (*63–66*).

The complex crosstalk between the different histone modifications and DNA methylation plays a fundamental role in shaping chromatin structure to control gene expression according to the type and physiological state of the cell.

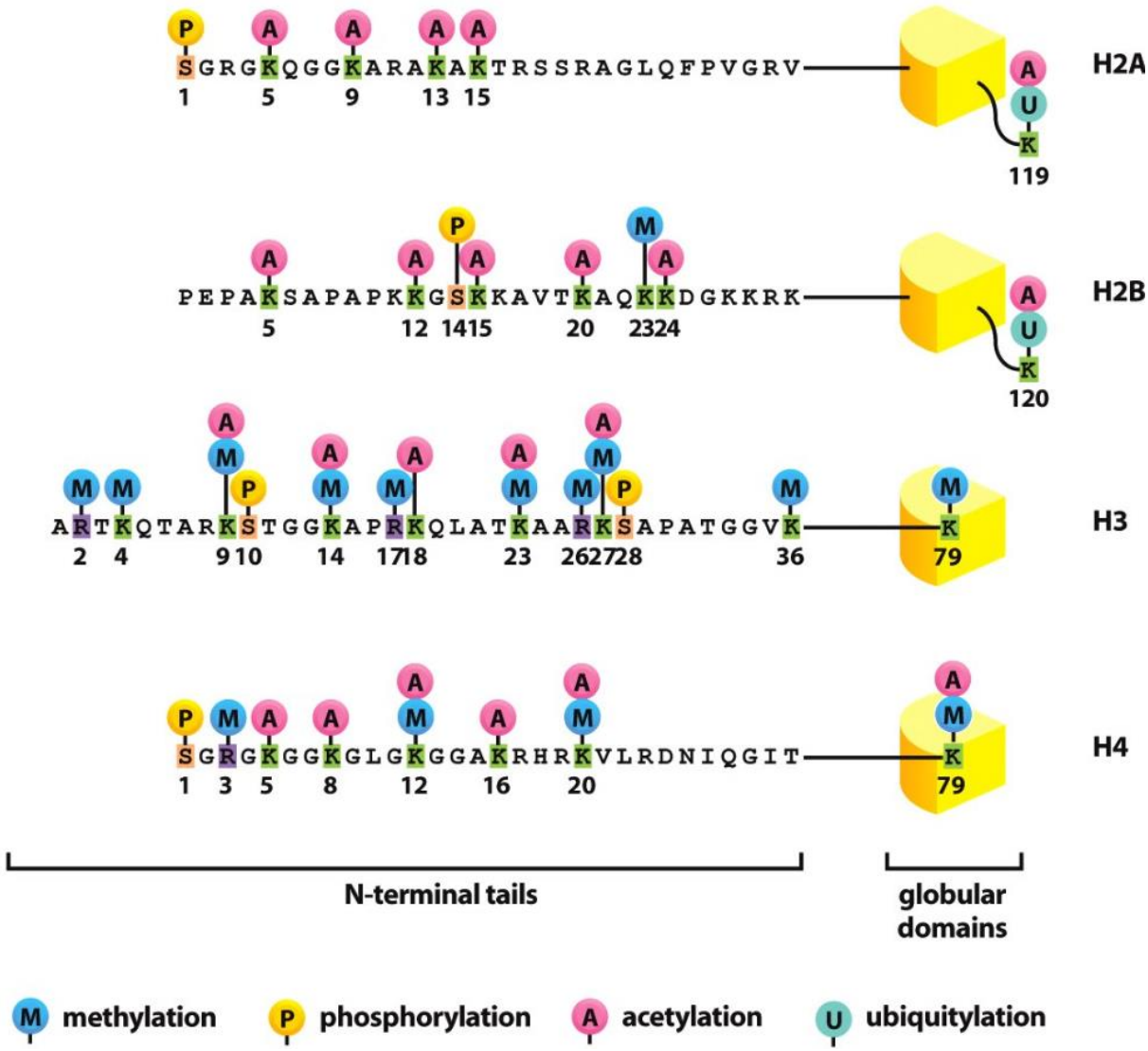


Figure 12. Potential modifications of canonical histones. The N-terminal tails of histone proteins can be decorated with many different post-translational modifications that have a direct impact in chromatin state and affect the recruitment of specific chromatin-remodeling complexes and architectural proteins. Adapted from (*1*).

2.2.1.2 DNA Methylation

DNA methylation is a key epigenetic modification in eukaryotes that involves the enzymatic addition of a methyl group to the fifth carbon of cytosine (C) to form 5-methylcytosine (5mC, **Figure 13**) (*67*). The presence of 5mC in DNA is directly involved in the regulation of gene expression, usually acting as a repressive mark. This is achieved by the presence of this modification in the major groove of DNA leading to impairment of transcription factor binding (*68*). 5mC plays a crucial role in fundamental processes such as development (*69*), cell differentiation, X chromosome inactivation and genomic imprinting. In addition, aberrant DNA methylation patterns are responsible for the pathogenesis of many diseases including neurodegenerative disorders, cardiovascular affections and cancer (*70*). Therefore, techniques that allow us to study DNA methylation are of major interest for the scientific community.

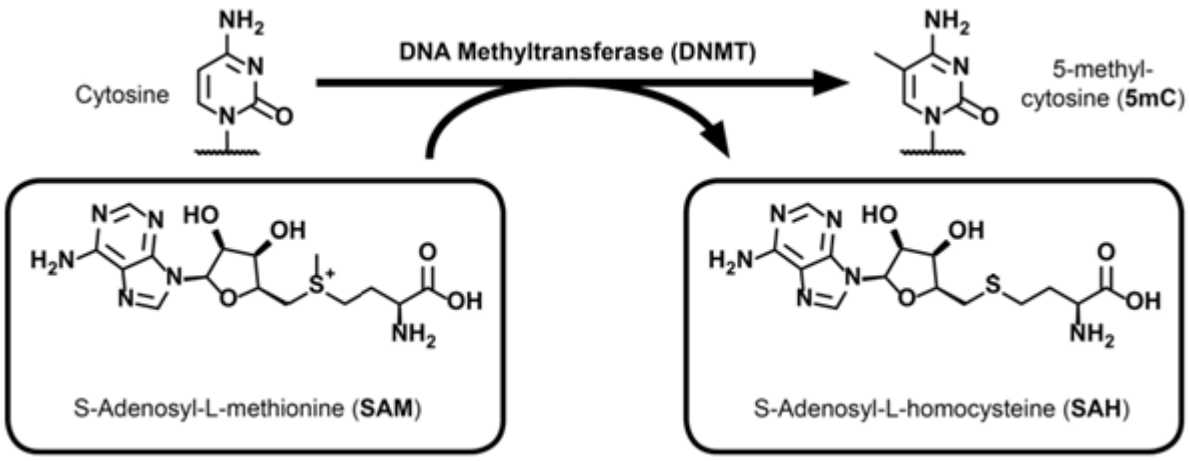


Figure 13. Cytosine methylation reaction in mammalian cells. DNA methyltransferases (DNMTs) catalyze the enzymatic addition of a methyl group to the fifth carbon of Cytosine using S-Adenosylmethionine (SAM) as cofactor.

5mC is introduced into the genome by S-adenosylmethionine (SAM)-dependent DNA methyltransferases (DNMTs, **Figure 13**). DNA methylation can occur *de novo* by the action of DNMT3 and can also be inherited and maintained through cell cycle by the maintenance enzyme DNMT1 that recognizes hemimethylated DNA and selectively methylates it (*70*). 5mC is introduced into CpG dinucleotides and remains stable accounting for 60-80% of all CpGs in somatic cells of mammals (*70, 71*). CpGs are paired with exactly the same sequence in opposite orientation in the complementary strand. This has two implications: on one hand, the

symmetry of the sequence enables the potential to contain the modification in both strands and on the other hand, it facilitates a mechanism to preserve DNA methylation after replication. Due to the semiconservative replication, the two resulting daughter DNA copies will be hemimethylated. The original strands that served as a template for replication will be methylated, while the newly synthesized strands will lack the modification as only cytosines and not methyl cytosines are available for incorporation by DNA polymerases. As previously stated, these hemimethylated sequences can be recognized by DNMT1, which catalyzes the addition of a methyl group to the cytosine of the unmethylated strand (*70*). This way, DNA methylation patterns can be stable and inherited by daughter cells allowing them to keep the identity of their progenitor after cell division (*71*).

Despite the stability of mC in the genome, it can be reverted back to C via passive or active mechanisms (**Figure 14**). Passive removal of 5mC involves DNA replication and absence of maintenance methylation leading to a progressive passive dilution (PD) (*72*). Active demethylation consists of the sequential oxidation of 5mC to 5-hydroxymethyl-(5hmC), 5-formyl- (5fC) and 5-carboxylcytosine (5caC) by Ten-Eleven Translocation (TET) dioxygenases (*73–75*). TET protein family members (TET1, TET2 and TET3) use α - ketoglutarate and Fe(II) as cofactors for the oxidation of 5mC and present different genomic distribution and activity during development (*76*). 5fC and 5caC are substrates for the Base Excision Repair pathway, leading to the excision of these bases by thymine DNA glycosylase (TDG) generating an abasic site (*77*). The generated abasic site is then repaired by active modification – active removal (AM-AR) restoring C (*77*). In addition, some studies suggest that hemi-modified CpGs reduce DNMT1 maintenance methylation compared to hemi-methylated CpGs, leading to the restoration of C by passive dilution (*77*).

The abundance of 5hmC, 5fC and 5caC in the Genome is cell type dependent. 5hmC is 10- to 100- fold more abundant than 5fC and 5caC and it is mainly enriched in neurons (reaching up to 0.7% nucleotides) and stem cells and it is much decreased in cancer cells (*78*). These differences are due to different oxidation kinetics and preference of TET1 and TET2 for 5mC over 5hmC and 5fC (*78*).

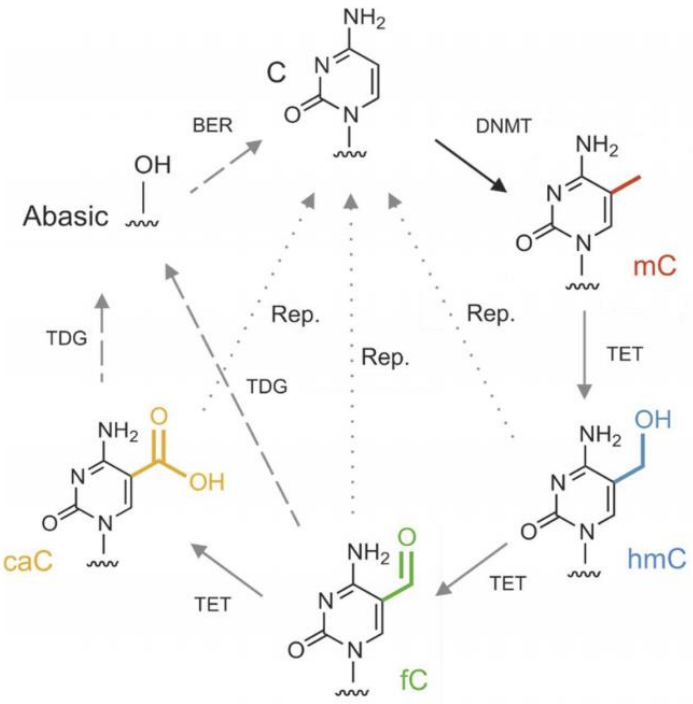


Figure 14. Cytosine methylation and active demethylation pathway. Cytosine is methylated to 5-methylcytosine (5mC) by DNA methyltransferases (DNMT) and can be further oxidized with different kinetics to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by Ten-Eleven Translocation (TET) dioxygenases. 5fC and 5caC can be actively removed by Thymine DNA Glycosylase (TDG) generating an abasic site that is repaired by Base Excision Repair (BER pathway). Also, 5hmC, 5fC and 5caC can lead to C restoration by passive dilution during replication and absence of maintenance methylation. Adapted from (*67*)

2.2.2 Methods for 5mC analysis

Bisulfite Conversion followed by DNA sequencing (BS-seq) is considered the gold standard technique and the most widely used to quantify methylation levels in a sequence-specific manner (*67*). This technique exploits the different chemical properties of C and 5mC. Bisulfite treatment of DNA at stable temperature and pH leads to deamination of C to U, but it does not affect 5mC. Therefore, after PCR amplification followed by DNA sequencing, C can be detected as T and 5mC as C, allowing to quantify the methylation level by mapping and comparison with the original sequence (*79*). BS-seq provides strand and nucleotide resolution, but the harsh treatment conditions can lead to extensive DNA degradation due to decomposition of pyrmidines and formation of abasic sites (*80*).

Other procedures include Restriction Endonuclease-based methods (MRE-seq), Affinity capture-based methods (MeDIP-seq or MBD-seq) and conventional analytical methods such as high performance liquid chromatography (HPLC), high performance capillary electrophoresis (HPCE) and mass spectroscopy (MS) (*81*). Each of these methods have their own limitations including low resolution, not providing sequence specific information, technical difficulties and bias towards highly methylated regions (*81*).

Studying DNA methylation from repetitive sequences contained at centromeric and pericentromeric regions supposes an additional challenge with the current techniques considering that the exact sequences are not fully known. It is estimated that 5 – 10 % of the genome at these regions is still uncharted and exploring their DNA methylation and transcription regulation is highly relevant due to their potential implications in disease (*44, 45*). In fact, anomalous methylation patterns in Satellite DNAs have been associated to several cancers and Parkinson Disease (*51, 52*). One strategy for the *in situ* analysis of cellular 5mC at repetitive sequences could be the application of imaging-based methods that allow to correlate methylation levels with specific genomic locations. Fluorescent immunostaining using 5mC specific antibodies has been used to study global cellular methylation, but it does not provide sequence resolution. Other approaches have included fluorescence in situ hybridization (FISH) probes (*82*) or DNA binding proteins in co-stains with antibodies or methyl-CpG-binding domains (*83–86*) in fluorescent complementation designs (*87–89*). These methods provide locus information but using two different receptor molecules (one for sequence specificity and another for 5mC detection) does not guarantee nucleotide resolution. A potential strategy for sequence-specific DNA methylation by imaging is the use of FISH probes with long chelator linkers for OsO4-mediated crosslinking. However, it requires harsh oxidative staining conditions, and nucleotide/strand resolution has not been demonstrated (*90, 91*).

To date, a cellular imaging method offering sequence-specific 5mC detection with nucleotide and strand resolution has not been reported yet. Having sequence specificity and 5mC selectivity in a single scaffold would be highly relevant for analysis of this epigenetic modification, especially at repetitive sequences difficult to study by other techniques. In this work we propose the use of fluorescent Transcription Activator-Like Effectors (TALEs) as programmable DNA binders that can detect 5mC for its study by imaging techniques (*2*).

2.3 Transcription Activator-Like Effectors

2.3.1 Origin and Discovery

Transcription Activator-Like Effectors (TALEs) are a family of DNA binding proteins firstly discovered in the 90's from different species of the plant pathogenic genus *Xanthomonas* (*92, 93*). TALEs are translocated into plant cells through a type III secretion system (T3SS) and then exported into the nucleus. There, they act as virulence factors activating the transcription of genes that support bacterial infection, proliferation and dissemination (*94, 95*). Preliminary comparative studies of the protein sequence from different *Xanthomonas* species revealed different degrees of homology. While N-terminal and C-terminal domains presented high level of similarity, interesting differences were found in the most characteristic structural feature of TALEs, the Central Repeat Domain (CRD). This region is composed by a variable number of repeats of 34 amino acids that are highly conserved (91 – 100% homology), except the two at positions 12 and 13 (**Figure 15a**). Surprisingly, the amino acids at these two positions were highly variable among the repeats of the same protein, as well as among homologs of different bacteria species. For this reason, this part of the repeats receives the name of Repeat Variable Di-residue (RVD) and it was hypothesized to be responsible for the specificity of TALEs (*92, 96*). Almost two decades later, that hypothesis was proven and the code governing the DNA binding specificity of TALEs was fully unraveled. RVDs can recognize specific nucleobases in a one-to-one basis independently of the sequence context. RVDs HD, NG, NI and NN preferentially bind to C, T, A and G/A nucleobases, respectively (**Figure 15b**) (*97, 98*). This discovery was the major breakthrough that put TALEs on the spotlight of the scientific community because of their evident potential applications. It is possible to target any desired DNA sequence by combining only four different TALE repeats. Previous approaches were based on Zinc Finger Proteins (ZFPs) in which each domain recognizes three nucleobases. This implies that at least 64 different modules are needed for covering all possible combinations. The simplicity of their code made TALEs perfect candidates to replace ZFPs-based approaches for DNA targeting (*99, 100*).

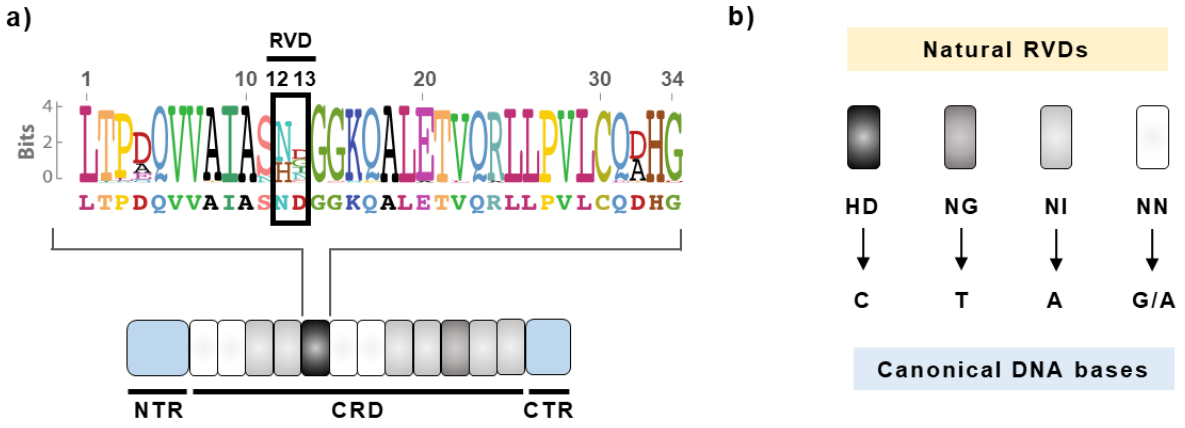


Figure 15. Schematic representation of TALEs and TALE code. a) Schematic representation of a TALE showing its N-terminal region (NTR), Central Repeat Domain (CRD) and C-terminal region (CTR). CRD is composed by a variable number of 34 amino acids modules whose consensus sequence is shown above. Amino acids are highly conserved except the ones at position 12 and 13, known as Repeat Variable Diresidue (RVD). **b)** RVDs are responsible for specific nucleobase recognition by TALEs. HD recognizes C, NG accommodates 5-mehtylpyrimidines like T, NI binds to A and NN recognizes purines like G and A.

2.3.2 Structure and Nucleobase Recognition

Natural TALEs contain an N-terminal bacterial type III secretion system signal (T3SS), followed by a series of amino acids with some similarity to the canonical repeats (**Figure 16a**) (*101*). The central part of the protein is occupied by the CRD containing an array of 11 – 30 repeats followed by an incomplete repeat known as "half repeat" (*102, 103*). The C-terminal domain contains a plant Nuclear Localization Signals (NLS) and an acidic Activation Domain (AD) involved in induction of transcription of target genes (*101*). This last part of the protein is not involved in DNA binding and therefore it has been modified for further applications of TALEs (**Figure 16b**). The NLS is substituted for that of the target organism, typically SV40 NLS for mammalians. The AD is replaced by the effector domain needed for the desired application, for example, FokI nuclease domain for Genome Editing or VP64 for transcription activation. The secretion signal in the N-terminal region was deleted (*99, 100*).

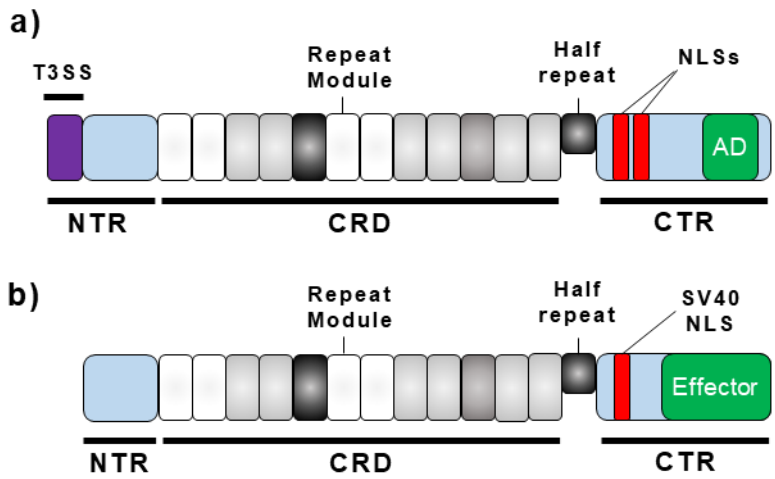


Figure 16. Schematic representation of natural and engineered TALEs. a) Schematic representation of a natural TALE containing the T3SS, plant NLSs and activation domain (AD). b) Schematic representation of a TALE modified for targeted applications.

Crystal Structure studies of TALE proteins bound to their target sequences revealed the major interactions of this protein-DNA complexes (*101–105*). TALEs form a right-handed superhelical structure that wraps the DNA duplex with RVDs interacting with the major groove (**Figure 17a-b**). Each repeat comprises two left-handed helices connected by an interhelical loop (**Figure 17c**). The first helix is composed by amino acids 3 – 11 and the second helix by amino acids 14 – 33, locating the RVD (positions 12 and 13) in a loop between them (*105*). The two hypervariable residues in the RVD loops, play different structural roles. The amino acid at position 12, Histidine (H) or Asparagine (N), orients away from DNA and interacts with the protein backbone to stabilize the interhelical loop (*102, 103, 105*). This interaction is mediated by a hydrogen bond between the side chain of the amino acid at position 12 and the carbonyl oxygen atom of the Alanine at position 8 of the first helix of each TALE repeat (*103*). By contrast to amino acid 12, which does not directly contact DNA, amino acid at position 13 is the responsible for the specific interaction of the repeat with the corresponding nucleobase (*101–103, 105*). In RVD HD, the Asparagine (D) 13 accepts a hydrogen bond from the amine group of cytosine, which cannot be established with other nucleobases (**Figure 17d**) (*105*). In the case of RVD NG, the reduced size of Glycine at position 13 allows sufficient space to accommodate the 5-methyl group of thymine and form nonpolar interactions (**Figure 17e**) (*103*). The isoleucine residue of the RVD NI establishes nonpolar van der Waals interactions with the C8 (and N7) of the adenine purine ring (**Figure 17f**) (*103, 105*). Finally, the second asparagine residue of NN interacts with the N7 nitrogen of the purine ring via hydrogen bond (**Figure 17g**). This explains the specificity of NN for both, guanine and adenine (*105*). Within the TALE

repeat, two additional nonspecific interactions with DNA are mediated by conserved amino acids Lysine 16 and Glutamine 17 through positive electrostatic interactions with the negatively charged DNA phosphates (*105*).

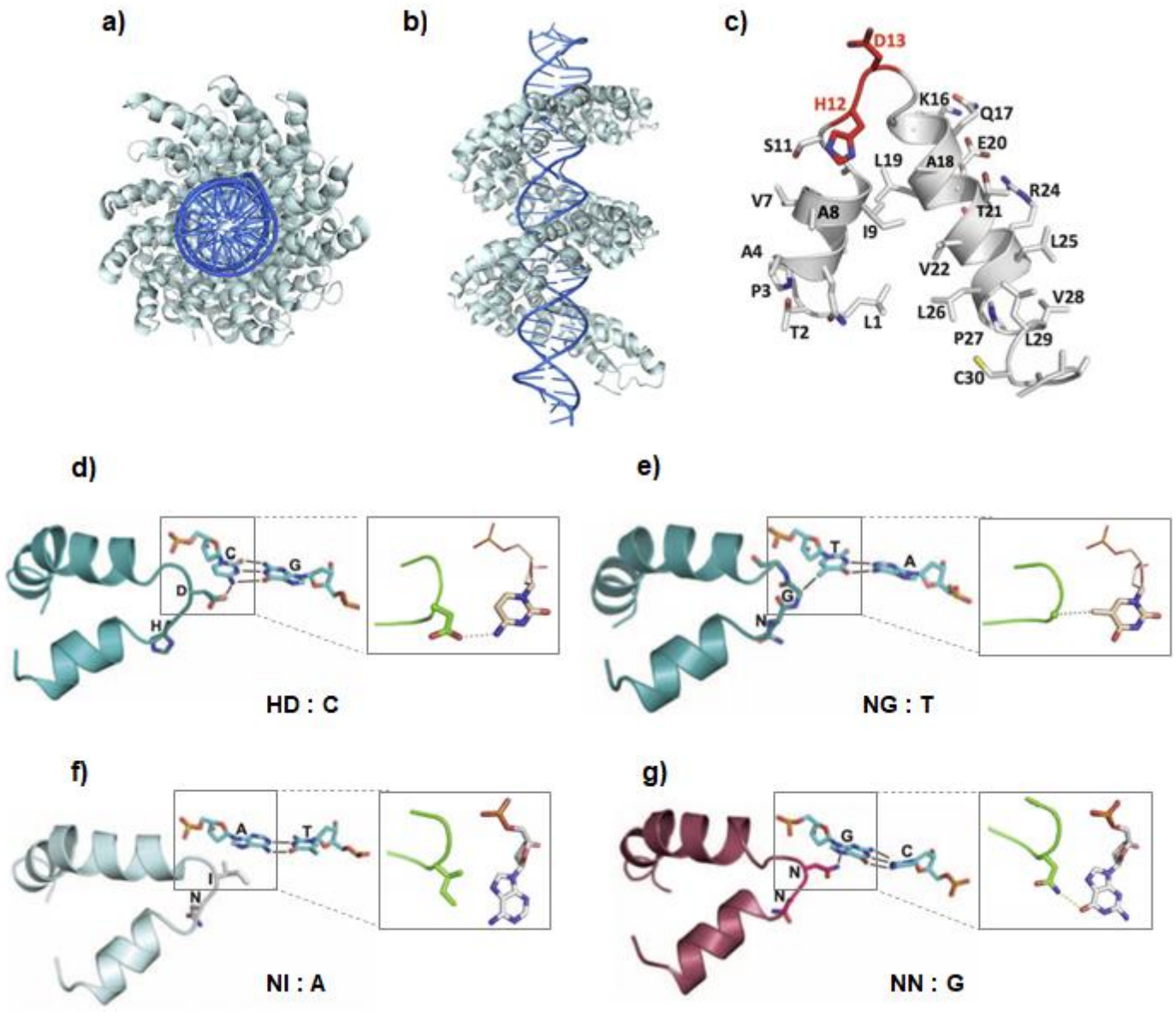


Figure 17. Crystal Structure of TALE PthXo1 bound to its target DNA and structural basis for nucleobase recognition. a) Top view of TALE PthXo1 bound to its target DNA sequence. **b)** Side view of the previously mentioned TALE. **c)** Structure of a single TALE repeat. Two helices with an interhelical loop where the RVD is positioned (red amino acids). **d)** Interaction between RVD HD and C. Hydrogen bond with the amine group of C. **e)** Interaction between NG and T. **f)** Interaction of NI with A. **g)** Interaction of NN with G. The asparagine 13 of NN interacts with the N7 of the purine ring of G or A via hydrogen bond. Adapted from (*105, 106*).

The majority of target sequences of natural TALEs are preceded by a 5' thymine base, suggesting that the N-terminal region could be involved in specific interactions with T (*103, 105, 107*). The N-terminal region bears some similarity to the canonical repeats of the CRD. Structurally, this region forms four cryptic repeats (named -3 to 0) and contains a large number of basic residues, which are thought to be involved in nonspecific direct contacts with DNA, mainly amino acids Trp232, Lys262, Lys265 and Arg266 (*104*). In particular, the indole ring of Trp232 establishes van der Waals interactions with the methyl group of T, possibly explaining the need of this nucleobase at the beginning of the target sequence (*104*). In order to overcome this constraint, mutational and directed evolution experiments were performed on residues 230 – 233 (KQWS). This led to a mutated motif (SRGA) with higher tolerance for other 5' nucleobases, but with overall reduced affinity (*108*). For this reason, the wild type motif is generally preferred even though the T constraint.

2.3.3 Target sequence Search Mechanism

Finding a specific target sequence is a complex process in the intricate landscape of the genome with varied chromatin states and other DNA binding proteins that can act as obstacles. TALEs show an unconventional search mechanism that alternates **sliding** and **hopping** behavior to diffuse rapidly along DNA (*109*). During the **search state**, TALEs bind nonspecifically to DNA through the NTR and move along the DNA strand followed by a hopping phase. **Hopping** involves a brief period of protein dissociation from DNA, after which the protein is able to rebind at a near distance from where it dissociated (*109*). The alternation of these two phases allow TALEs a quick search along the genome and bypassing potential obstacles at the same time. Once the target sequence has been found, TALEs transition into **recognition state**. During this state, the protein adopts a more compact conformation in which the CRD establishes the energetically favorable interactions through the RVD (*109*). If the TALE protein enters into the recognition state while it is bound to an incorrect sequence, the steric and electrostatic clashes between RVDs and nucleobases would tend to destabilize the complex (*109*).

2.3.4 TALE Assembly

One of the major challenges for the adoption of TALEs as programmable DNA binding proteins was the difficulty of their cloning and assembly due to extensive identical repeat sequences. The groups of Bogdanove and Voytas developed an elegant strategy for TALE assembly based on Golden Gate cloning method (*110*). Golden Gate is a method for assembling multiple DNA fragments in an ordered fashion in a single reaction thanks to the use of Type IIS restriction enzymes (*111, 112*). This type of restriction enzymes cleaves outside its recognition site generating unique 4 bp overhangs. By clever design of the resulting overhangs in each RVD repeat module, it was possible to create modules with different sticky ends for each position, making possible to assemble the repeats in the desired order (*110*). TALE assembly by Golden Gate reaction consists of two steps. In the first step, known as Golden Gate 1 (GG1), two plasmids are generated. One of them contains the first 10 repeats and the other one is composed by a variable number of repeats (1-10) which will constitute repeat 11 up to repeat 20. In the second step, Golden Gate 2, both plasmids are fused together with a plasmid containing the last truncated repeat (known as "half repeat") and inserted in the final entry vector that contains the NTR and the effector domain. This way, it is possible to create functional TALEs efficiently in five days (*110*). Golden Gate assembly supposed an important contribution for the general adoption and application of TALEs.

2.3.5 TALE Applications

2.3.5.1 Genome Editing

TALEs have been used as programmable DNA Binding proteins for multiple applications. They have been fused to homodimeric FokI nucleases to create Transcription Activator-Like Effector Nucleases (TALENs) for Genome Editing (*99, 100, 113*). In this approach, two TALENs binding sequences in a short distance between each other are used to induce double strand breaks (DSBs) in the target gene. Then, the natural DNA repair mechanisms can be exploited for two different purposes. On one hand, the high fidelity Homologous Direct Repair (HDR) mechanism can be used in combination with DNA constructs flanked by homology arms for its site-specific insertion. On the other hand, the error-prone Non Homologous End Joining

(NHEJ) mechanism tends to generate insertions or deletions (indels) at the DSB in order to repair it. These indels can lead to the disruption of the targeted gene either by generation of frameshift mutations, of premature stop codons or by large insertions/deletions. This second repair mechanism has been exploited for the generation of gene knockouts (KO) (*100*). However, DSBs could induce the activation of either HDR or NHEJ and the specific elements triggering one or another are not well understood. In order to avoid the activation of NHEJ for applications in which HDR is the preferred mechanism (like gene correction or cassette insertion), one of the FokI domains has been mutated to generate a nickase (*114*). In this approach, only one of the homodimeric FokI domains is functional, while the other one is catalytically inactive. Therefore, only a single stranded break or nick is induced activating only HDR and not NHEJ (*114, 115*). Another approach consisted of the fusion of a recombinase to the TALE so the integration of the corrected gene or expression cassette of interest is recombined with the homologous endogenous gene without DSB induction, thus preventing potential mutations (*116*). TALENs haven been extensively applied for Genome Editing in order to create new study models as well as for potential clinical applications like Gene Therapy *(117–120).*

2.3.5.2 Targeted Gene Activation and Repression

Further applications of TALEs consist of the fusion with transcriptional activators or repressors for the control of gene expression. Activation domain of VP16, a protein from human herpes simplex virus, has been fused to TALEs to effectively increase the expression of silenced genes like endogenous NTF3 or exogenous gene reporters (*121, 122*). Higher levels of expression were achieved by fusion with VP64, which is composed by a tandem of four copies of VP16 (*123*). Using this approach, it was possible to induce transcription of SOX2, KLF4, c-MYC, OCT4 and NANOG (*123–125*)even achieving reprogramming into pluripotency (*126*). Other reports have demonstrated the activation of gene expression of latent HIV (*127*), CELS1R in endothelial cells (*128*) and it has even been proposed as a therapy for the trinucleotide repeat disorder Friedreich's Ataxia by the targeted activation of frataxin gene (*129*).

Several domains have been tested for transcription repression and gene inactivation. One study screened repressor domains from different species including the PIE-1 repression domain (PIE-1) from *Caenorhabditis elegans*, the QA domain within the Ubx gene (Ubx-QA) from

Drosophila melanogaster, the IAA28 repression domain (IAA28-RD) from *Arabidopsis thaliana*, the SID, Tbx3 repression domain (Tbx3-RD) and the Krüppel-associated box (KRAB) repression domain from Homo sapiens (*130*). Results showed that TALE-mediated gene repression is stronger when TALEs are fused to KRAB or SID domains, while other domains led to slight or no repression (*130*). Other studies using TALE-KRAB fusion for targeted gene repression include a lentiviral system for knocking down genes (c-kit and PU.1) in mouse bone marrow cells (*131*), knockdown of Tcf3 to induce self-renewal of Hematopoietic Stem Cells (*132*) and inactivation of human Hepatitis B virus in HEK293 and Huh7 cells (*133*). Apart from the therapeutic and research applications of TALEs fused to activators and repressors, this method has been extensively used as a screening platform to study the affinity and specificity of TALEs (*130, 134*).

2.3.5.3 Epigenome Editing

Another approach to study the effect of gene activation or repression is via the use of epigenetic modifiers (like DNA Methyltransferases or TET dioxygenases) to induce chromatin changes affecting gene expression (*135*). In addition, it allows for studying the role of modified cytosines at specific genomic sites. TALEs have been fused to the catalytic domain of DNA Methyltransferases DNMT3a, DNMT3L or a fusion of both (consisting of the C-termini of DNMT3a and DNMT3L) (*136*). DNA methylation at specific promoter regions of several genes (CDKN2A, DLK1-MEG3 and Ascl1) using this method was proven to be biologically active leading to gene silencing and reduction of expression) (*136–138*). In order to achieve the opposite effect, TALEs have been fused to the catalytic domain of TET1 to actively demethylate the promoters of genes such as Klf4, RHOXF2 (*139*) and CXCR4 (*140*), reducing 5mC levels and effectively increasing the expression level of the target gene.

2.3.5.4 Genome Visualization

Three papers published weeks apart reported the fusion of TALEs to fluorescent proteins as a tool to label and track repetitive genomic sequences (**Figure 18a**) (*141–143*). This method was proven to be effective to image different repetitive DNA sequences (telomeres, alpha-satellites, major satellites, and minor satellites) from a wide variety of human and murine cell lines (**Figure 18b**). In addition, it seems to be compatible with any fluorescent protein variant and

applicable in both live and fix cells (*141*). Specificity of TALEs for their repetitive target sequence has been proven by colocalization with either DNA-FISH probes or with immunostained chromatin specific proteins (*141–143*). Strikingly, application of this method in live murine ES cells allowed to distinguish the parental origin of chromosomes differing only in a single nucleotide polymorphism (SNP) at the targeted sequence. This result suggests that specificity of TALEs is high enough to discriminate sequences differing by a single nucleotide (*142*). Additional studies aimed to improve the signal-to-noise ratio by combining pair of TALEs with biomolecular fluorescence complementation (BiFC) (*144, 145*) and promote correct TALE localization and prevent protein aggregation in live cells by fusion with proteins known to improve protein solubility like TRX (*146*). The latter study applied this method to study hallmarks of ageing like genome instability, telomere attrition, chromatin decompaction and rDNA mislocalization (*146, 147*).

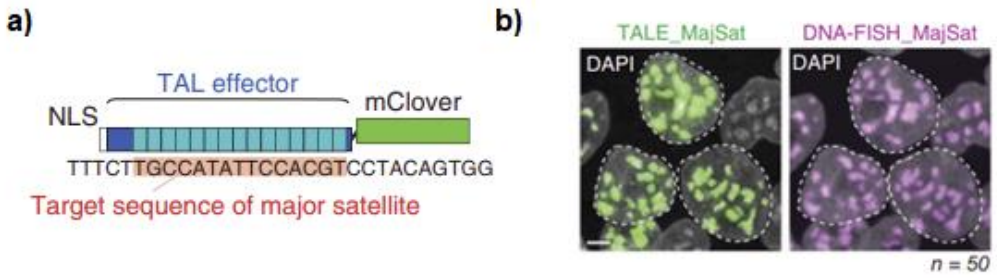


Figure 18. Genome Visualization using fluorescent TALEs. a) Schematic representation of a fluorescent TALE targeting a sequence of murine major satellites. **b)** Visualization of murine major satellites with TALE-mClover fusion (left) and conventional DNA FISH (right). Adapted from (*142*).

2.3.6 Epigenetic Analysis with TALEs

2.3.6.1 Discovery of the sensitivity of TALEs to 5mC

Early studies of gene activation in live cells with TALE-VP64 fusions showed that TALE binding was affected by methylation status of the target sequence (*124*). In particular, TALE-VP64 transcription activation was drastically reduced when targeting heavily methylated promoter as OCT4. Same result was obtained *in vitro*. That sensitivity towards methylated DNA could be overcome by treating cells with the DNMT inhibitor 5-aza-2'-deoxycytidine (5azadC) (*124*). Crystal structure studies of TALEs showed that RVD HD establishes a H-bond between the carboxyl group of aspartate and the N4 atom of C (**Figure 19a**). However, this interaction

cannot take place when a methyl group is present at the 5th carbon of C (*106*). Therefore, 5mC abrogates binding by HD. By contrast, RVD NG, natural binder of T, is able to effectively bind 5mC (**Figure 19b**). The reduced size of Glycine at position 13 and the lack of side chain provide sufficient space to accommodate the 5-methyl group of T (and 5mC) and allows optimal van der Waals interaction between the Cα of Glycine and the 5-methyl group. The only difference between bases T and 5mC is at position 4 (**Figure 19c**), which is not involved in TALE recognition and thus they are undistinguishable by TALE proteins(*106*). Another study found that N*, a natural RVD in which the amino acid at position 13 is not present and natural binder of C and T with the same affinity, is also able to bind 5mC. Due to the missing residue at position 13, N* could avoid steric clashes with the methyl moiety at position 5 of thymine and by extension also of 5mC. Similar results were found with RVD H* (*148*).

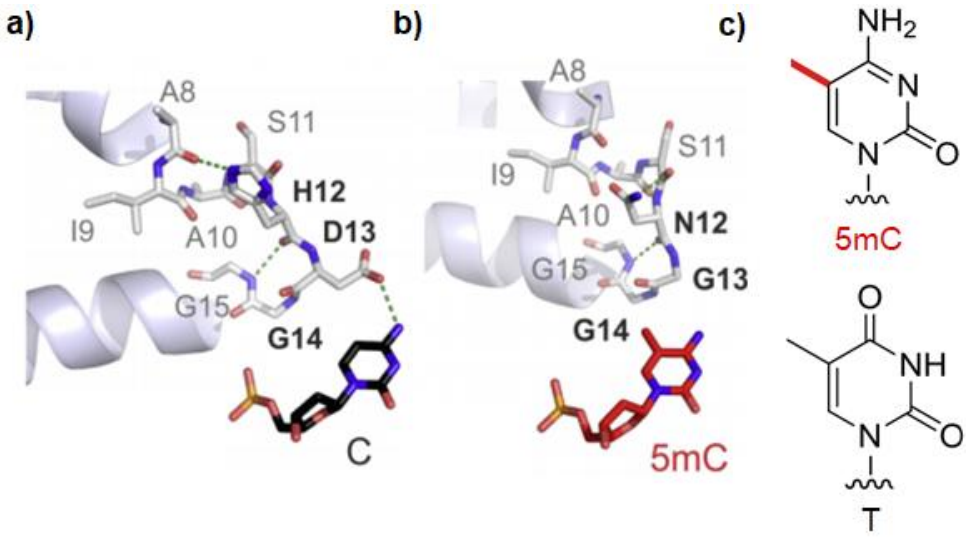


Figure 19. TALE repeat interactions with C and 5mC. a) Detailed interaction of TALE repeat HD with C via hydrogen bond. **b)** Detailed interaction of TALE repeat NG, natural binder of T, with 5mC. The reduced size of Glycine at position 13 allows the accommodation of the methyl group of 5mC. **c)** Chemical structures of 5mC and T for comparison. The interaction surface displayed on the major groove by both of them is the same. Modified from (*149*).

2.3.6.2 Detection of C, 5mC and 5hmC with specific and universal RVDs

The sensitivity to 5mC revealed a limitation of TALEs for Genome Engineering application because special care must be taken when targeting a methylated sequence. However, it also opened a new avenue for using TALEs for epigenetic studies. The sensitivity of HD towards 5mC was exploited to detect methylated DNA *in vitro* using DNA footprinting assay by primer

extension and applied even in genomic DNA context (*150*). This method was further improved to have single nucleotide resolution by using RVD HD opposite to the C to be analyzed and RVD N* (C and mC binder) for the rest of C. This way, single C positions within a target sequence can be interrogated by strategically placing the mC sensitive RVD HD (*151*). Discovery of the progressive oxidation of 5mC by TET to 5hmC, 5fC and 5caC led to the exploration of selective RVDs towards such bases (*152*). A new approach based on the previous one could detect differences in C, 5mC and 5hmC. In this case, RVD HD was used for binding C but not mC or hmC, NG was chosen for its preferential recognition of 5mC and the newly discovered sensitivity of N* to 5hmC was used as a negative sensor (*152*). This same strategy was later applied to an affinity enrichment approach based on solid phase DNA sequence isolation with immobilized TALEs (*153*). It was possible to quantify 5mC levels and detect 5hmC of several cancer biomarkers sequences in Human Genome context. Additionally, 5fC levels could be indirectly detected by N* TALE after chemical reduction of 5fC to 5hmC using sodium borohydride (*153*).

2.3.6.3 Engineered RVDs for recognition of epigenetically modified bases

The binding mode of TALEs offers the potential to recognize the unique chemical information displayed in the major groove and makes these modular proteins good candidates for protein engineering. Several groups have explored the possibility of engineering TALE repeats for improving or even generating selective RVDs for each of the epigenetically modified bases. Previous reports of natural RVDs NG, N* and H* being able to bind 5mC indicated that small residues as well as deletions at position 13 seemed to be beneficial for the accommodation of bigger functional groups. Based on this, seven libraries carrying randomized mutations at either position 11 or 12 together with successive deletions were generated: X*, X**, NX**G, SX***, NX***, X***G, X**** (X = any amino acid, * = deletion) (*149*). The screening showed that RVDs G*, S* and T* worked as universal repeats binding C as well as all the epigenetic bases. P* was found to be a negative sensor of 5caC (*149*). By contrast, R**** did not bind any nucleobase except 5caC, being the first highly selective RVD for this nucleotide variant ever reported (*154*). An independent study screened every possible combination of residues at position 12 and 13 (library XX) as well as the previously tested X* for C, 5mC and 5hmC selectivity (*134*). This

study confirmed the universality of RVD G* and the sensitivity of N* to 5hmC as previously reported (*7, 134, 149*). In addition, HA was found to be a 5mC binder, FS a 5hmC binder and RG can bind to 5mC and 5hmC, but not unmodified C (note that affinity for 5fC and 5caC was not tested) (*155*). A third study constructed a TALE repeat library by randomizing positions 11-14, this include the RVD and the flanking residues (*156*). The most interesting RVD found, ASAA, showed strong preference for 5mC. This can be explained by the small size of alanine and serine residues which reduce the steric hindrance between the RVD loop and the methyl moiety (*135, 156*).

2.3.6.4 Enhanced 5mC selectivity by engineering DNA Backbone Interactions

A different strategy to enhance 5mC selectivity of TALEs consists of reducing nonspecific interactions of TALE repeats. Apart from the specific interactions of TALE repeats with specific nucleotides through the RVDs, TALEs nonspecifically interact with DNA phosphates via basic amino acids (*101*). These positively charged residues are mainly located in the NTR. Conserved residues 16 and 17 (KQ) from each TALE repeat are also involved in nonspecific interactions with the phosphate backbone (*101*). An alternative strategy to increase selectivity of RVDs consists of decreasing the nonselective binding energy (*157*). In order to achieve this, basic amino acids in the NTR as well as KQ residues in the TALE repeat were substituted by alanines (*157*). C over 5mC selectivity was moderately increased with K262A mutation in the NTR and more strongly enhanced by K16A + Q17A substitutions within the HD repeat. Combination of mutations in the NTR and CRD can have synergistic effect in particular TALEs, typically those larger than 25 RVDs, which can tolerate more than one mutation (*157*).

2.3.6.5 Detection with TALEs in combination with chemoselective blockage

Another strategy to quantify 5mC oxidized species using TALEs consists of the use of a universal RVD like G* together with chemoselective nucleobase blockage (*158*). The rationale behind this idea is to selectively modify the functional group at position 5 by a bulkier group which impairs its accommodation by G*. 5hmC can be converted into 5-glucosyl-hmC by uridine-diphosphate (UDP)-glucose-dependent glucosylation with T4-β-glucosyltransferase.

The large hydrophilic glucose moiety cannot be tolerated by G*. Binding to 5fC can be negatively affected by its conversion to oximes using hydroxylamines. In particular, tertbutyloxime strongly abolished TALE binding. Finally, amines can be employed for amide-derivatization of 5caC using (7-azabenzotriazol-1-yloxy)tripyrrolidino-phosphonium hexafluorophosphate (PyAOP) as coupling reagent under nondenaturing conditions. Only benzylamine exhibited effective blocking and therefore binding inhibition (*158*).

A different chemical approach for selective detection of 5fC used G* repeat containing the noncanonical amino acid p-acetyl-L-phenylalanine at position 11 or 12 (*159*). This ketone-bearing amino acid selectively forms diaminooxy-linker-mediated dioxime cross-links to 5fC enabling single CpG resolution and direct quantification of 5 fC levels (*159*).

3. Aim of work

The aim of our work is the development of an imaging-based method for 5mC analysis of single cells that overcomes the current limitations. Current approaches for studying DNA methylation by cell imaging include immunostaining with 5mC specific antibodies or methyl-binding domains (MBDs) in combination with FISH probes or programmable DNA binding proteins to provide sequence resolution. Although they can offer locus information, using two different receptor molecules (one for sequence specificity and another for 5mC detection) does not guarantee nucleotide resolution. An alternative approach tried to tackle this issue by using FISH probes with long chelator linkers for OsO4-mediated crosslinking. However, it requires harsh oxidative staining conditions, and nucleotide/strand resolution has not been demonstrated (87, 88). To date, a cell imaging method offering sequence-specific 5mC detection with nucleotide and strand resolution has not been reported yet. Having sequence specificity and 5mC selectivity in a single scaffold would be highly relevant for analysis of this epigenetic modification, especially at repetitive sequences difficult to study by other techniques.

Our method will exploit the sequence specificity and 5mC sensitivity of TALEs and use them in fluorescent scaffolds to visualize chromatin and detect methylation levels. To achieve this, we will use pairs of fluorescent TALEs targeting the same CpG containing sequence. Each of the TALEs from this pair will be fused to a different fluorophore (either eGFP or mCherry) and will only differ from its counterpart in the RVD interacting with the C at the CpG of interest. One of the TALEs will contain the 5mC sensitive RVD HD and the other TALE will carry the universal RVD G* that can bind C and 5mC with similar affinities. This way, differences in methylation will be detected as differences in the fluorescence intensity of the HD TALE, but not of the G* TALE, which will serve as a control. Therefore, an increase of 5mC will lead to a reduction of binding of HD TALE and its fluorescence intensity will decrease. By contrast, G* TALE is not responsive to methylation and its fluorescence will not be affected by 5mC variations. However, DNA methylation is also associated with chromatin compaction which leads to an overall reduced target accessibility, in which case the fluorescence intensity of G* TALE will decrease. This way, G* TALE will serve as a control for chromatin accessibility.

This method will facilitate the study of DNA methylation of repetitive loci at centromeric and pericentromeric regions, whose exact full sequences are not known according to the current assembly of the Human Genome. In addition, the use of a TALE-based imaging approach will potentially overcome the limitations of current methods and provide new insights into the impact of 5mC in chromatin structure of repetitive sequences offering topological and positional information. We will explore changes in methylation of SatIII, a subfamily of satellite DNA that is present at pericentromeric regions of most of chromosomes as clustered repeats. These sequences are involved in nuclear stress bodies formation and aberrant expression and methylation patterns are potentially involved in the development of several diseases, including cancer. Studying SatIII remains a challenge since its abundance through the genome makes difficult to map and correlate 5mC levels with discrete genomic positions. We aim to demonstrate that our method for 5mC analysis is applicable to study user-defined clustered repetitive DNA with single nucleotide and strand resolution. Furthermore, the imaging properties of our approach will allow to perform single cell studies and chromosome-specific analysis. In addition, we will combine our method with classical immunostaining to reveal the role of 5mC in HSF1 recruitment and prove if our method will be applicable to correlation studies with other imageable chromatin features.

Finally, we will explore the potential of TALE repeat engineering to generate novel RVDs with enhanced 5mC selectivity. In order to find potential 5mC readers, we will construct and screen a library of size-reduced TALE repeats. We expect that a pair of TALEs, one of them with a 5mC sensitive RVD and the other one with a 5mC-binding RVD, will improve imaging-based analysis offering higher dynamic range and response to methylation by both TALEs.

4. Results and Discussion

Fluorescent TALEs have been previously used to visualize repetitive loci in human, murine and plant cells (*141–143, 160*). We designed TALEs fused to eGFP to specifically label the repetitive loci SatIII in Human Genome. This pericentromeric DNA sequences are responsible for nuclear stress bodies (nSBs) formation and their dysregulation has been reported in several types of cancer (*50-57*). Fluorescent TALEs to label SatIII were recombinantly expressed in *E. coli* and purified by His-tag affinity enrichment. We chose to use purified proteins to precisely control the protein concentration, ensure homogenous staining and allow for potential applications in clinical tissue samples and hard-to-transfect cell lines. Previous reports showed that affinity and specificity of TALEs are affected by their length, peaking between 15 and 19 repeats (*161*). In order to ensure minimal excess binding energy, maximal affinity and high signal/nose ratio, we optimized the number of repeats per TALE. We assembled TALEs of variable length targeting the SatIII sequence "TGGAACGAACGGAATGGAATGGAATGGAA" and test their performance by cell staining and electromobility shift assay (EMSA) (**Figure 20a-b, Fig S1-2**). Microscopy analysis of stained HeLa cells revealed that TALEs with 17 repeats presented the highest fluorescence intensity, optimal signal-to-noise ratio and expected foci number and morphology (**Fig S1**). Shorter TALEs showed similar fluorescence intensity, but lower signal-to-noise ratio, while longer TALEs presented a markedly decrease in fluorescence intensity and strongly reduced number of foci. EMSAs showed that TALE affinity increases with length, peaking at 17 repeats and rapidly decreasing in TALEs longer than 20 repeats (**Fig 20a and S2**). Therefore, we proceeded with a 17 repeats TALE, termed TALE_2 (for two CpGs in the target: TGGAACGAACGGAATGGA) for further experiments. This TALE length lay within the range considered as optimal by other studies (*161*).

We aimed to develop an imaging-based method for 5mC analysis with strand and nucleotide resolution. Crystal structures revealed that TALEs bind to DNA in the major groove interacting with only one of the strands, what implies that strand resolution is expected (*101–103*). Furthermore, a cell imaging study of mouse centromeres using fluorescent TALEs reported the capability of TALEs to distinguish single nucleotide polymorphisms(*142*). This suggests that single nucleotide selectivity may be achieved. Regarding 5mC discrimination with TALEs, both

in vitro and live cell studies reported by independent groups have consistently demonstrated sensitivity of RVD HD towards methylated cytosine (*7, 106, 124, 150–153, 156, 157, 162*). By contrast, the engineered repeat G* was characterized as a universal binder able to recognize any nucleobase, including 5mC (*149*). In order, to demonstrate the strand and single nucleotide resolution as well as the discriminatory power of TALE_2 to distinguish methylated and non-methylated DNA, we tested the performance of this TALE *in vitro*. We initially expressed two versions of TALE_2 bearing either the selective RVD HD or the universal RVD G* at positions 5 and 10, which are the repeats involved in the recognition of C from both CpGs present in the target sequence. *In vitro* footprinting assays confirmed the universality of RVD G*, equally binding methylated and non-methylated DNA, and the sensitivity of RVD HD to 5mC (**Fig 20c and SI**). Moreover, binding of HD TALE_2 was inhibited only if the methylation occurred in the target strand, demonstrating strand discrimination. Strikingly, methylation in only one of the CpGs of the target strand was sufficient to reduced binding of HD TALE_2, confirming the single nucleotide resolution capabilities of TALEs.

Next, we studied the actual nucleobase composition of SatIII DNA at the two studied CpGs from the target sequence population. We assembled both HD and G* TALE_2 versions fused to eGFP and mCherry for co-staining of HeLa cells (**Fig 1d**). Co-staining with either both HD or both G* constructs showed perfect colocalization and homogenous fluorescence levels, demonstrating that fluorophores do not influence TALE selectivity (**Fig 20e**). Interestingly, despite similar affinities of HD and G* repeats (**Fig S1c**), the number of foci is significantly more abundant for the G* TALE. In fact, co-staining with HD and G* TALEs showed a not fully overlapping foci pattern (**Fig 1f**), revealing a SatIII population not containing C at the target CpG. This population is visible as additional G* foci that do not colocalize with any HD focus, while every HD focus is colocalizing with G* one because this universal binder covers every possible nucleobase. To reveal the actual nucleobase content at these positions, we performed co-stains with HD TALE_2 and versions bearing NN, NI or NG (binding G, A and T/5mC, respectively; **Fig 20g and Fig S3b, S5**). NN and NI TALEs did not afford any foci, suggesting that G and A are not present at SatIII loci at detectable levels. By contrast, NG TALE resulted in an abundant number of foci partially colocalizing with HD and perfectly colocalizing with G* (**Fig 20g**). This indicates that SatIII DNA presents a population of foci containing 5mC/T at

the target CpGs, suggesting that the presence of differential methylation. This makes this sequence a suitable target for analysis of differential C/5mC levels.

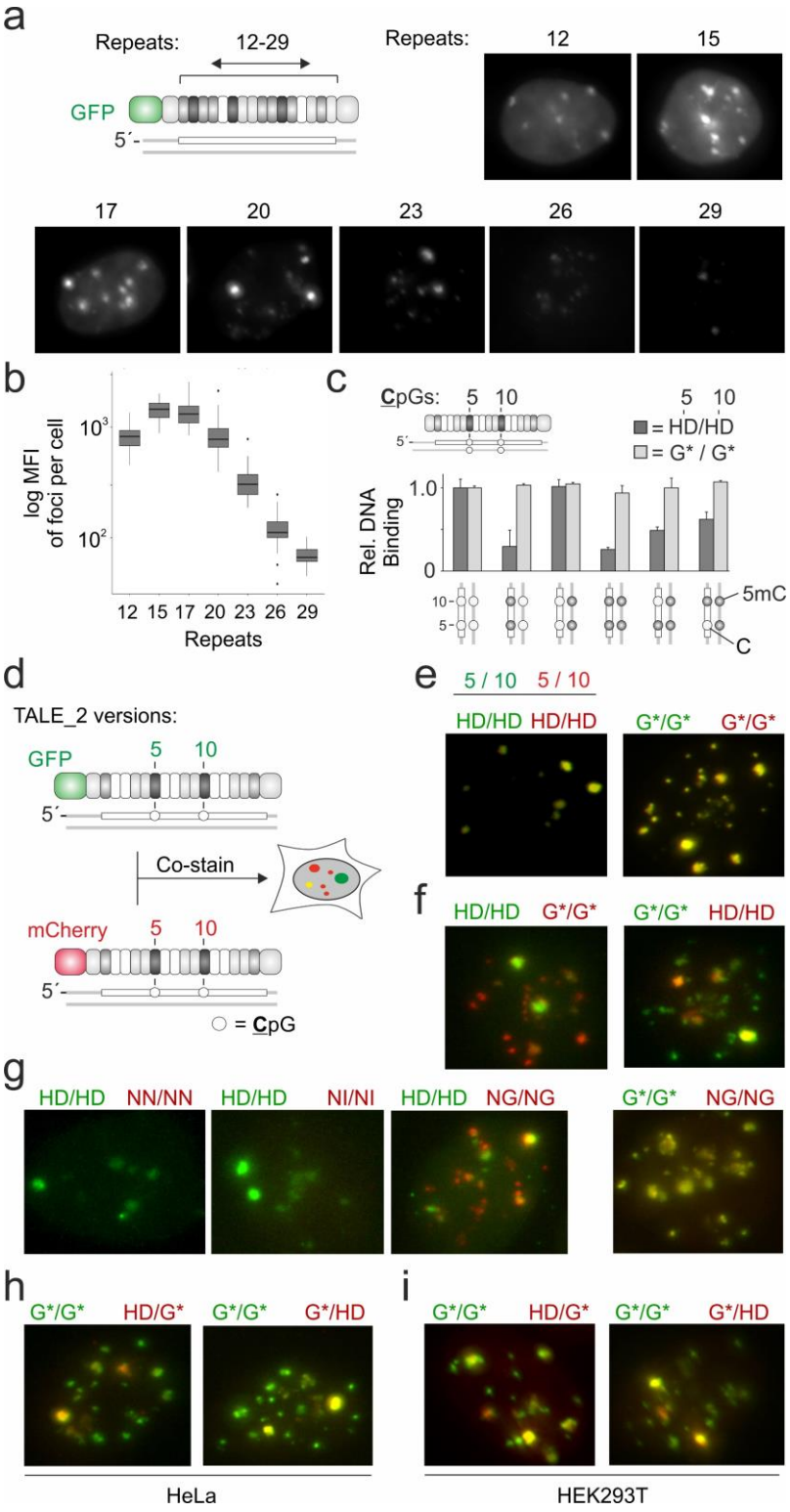


Figure 20. TALE length optimization, selectivity and nucleotide and strand resolution. a) Scheme depicting the eGFP-TALE construct targeting the SatIII locus and TALE-staining of HeLa cells. Only one cell per picture is shown. **b)** Log transformed mean fluorescence intensity of eGFP foci per cell from staining in Fig. 20a depending

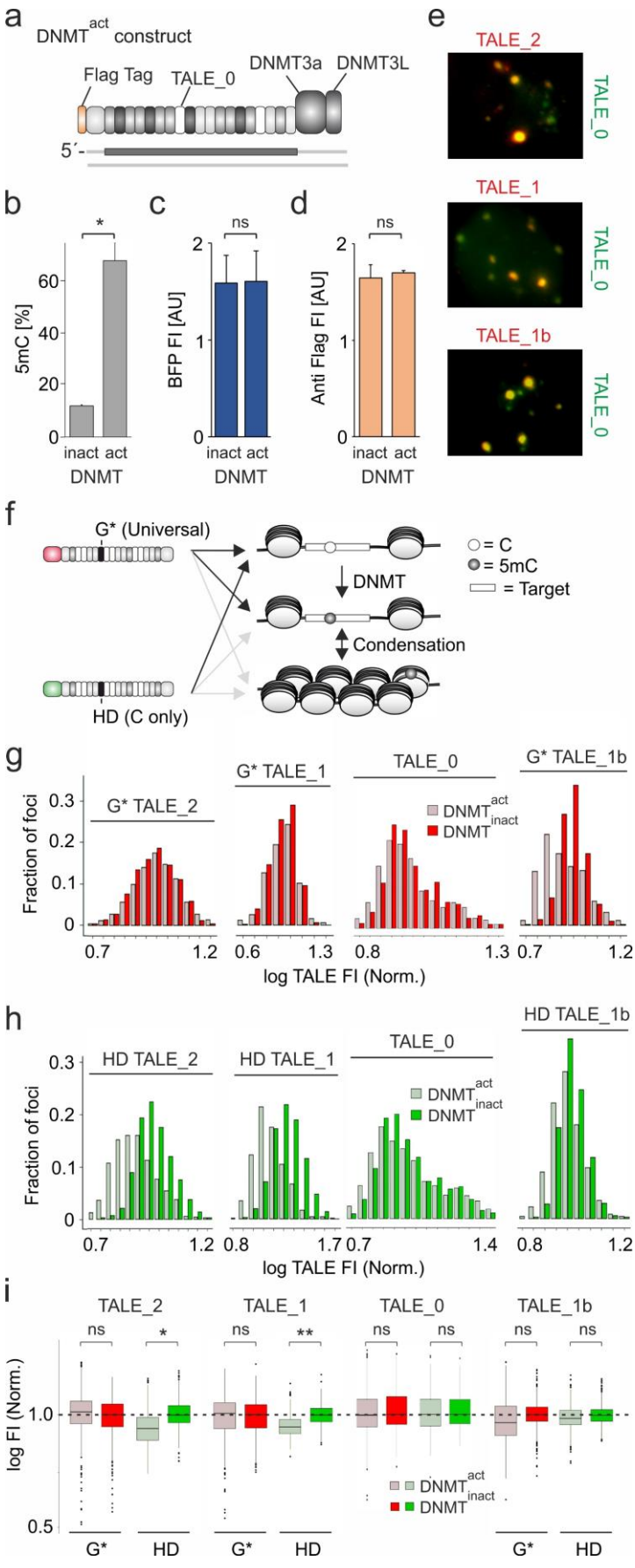
on the number of TALE repeats (N=512 cells). **c)** DNaseI footprinting assay showing nucleotide and strand resolution of HD TALE_2 and the universality of G* TALE_2. **d)** Schematic representation of the co-staining procedure using pairs of fluorescent TALEs. **e)** TALE co-staining in HeLa cells with either two HD TALEs (left) or two G* TALEs (right) in equimolar concentrations. Only one cell per image is shown. Magnification 60X **f)** Same as in e, but with the indicated pair of TALEs. **g)** Same as in e but with the indicated pair of TALEs. NN TALE and NI TALE did not afford any focus, while NG TALE presented many and perfectly colocalize with G* TALE. **h)** Same as e with the indicated pair of TALEs. Foci patterns from single RVD exchanges suggest nucleotide resolution. **i)** Same as in h, but in HEK293T.

Overall, this experiment showed the high selectivity of TALEs towards specific nucleobase changes in our staining procedure. Next, we evaluated the potential for single nucleotide resolution of our imaging method by co-staining HeLa cells with G* TALE_2 and a version of this TALE in which one of the G* was replaced by HD. This scaffold allows to ignore one of the CpG positions since G* is a universal binder and interrogate the other one for C/5mC levels with the sensitive RVD HD. The results revealed the same staining pattern as for co-staining with only HD plus only G* TALE_2 versions (**Fig. 20h** compared to **Fig. 20f,** right). The absence of HD fluorescence for many of the G* foci confirmed the high selectivity of HD for C, while the colocalization at other foci suggests that the nucleobase composition heterogeneity can be detected and analyzed with our approach. Similar results were obtained by co-staining of HEK 293T cells showing the versatility of the method to be applicable in different cell types (**Fig. 20f**).

The aim of this work is to detect methylation changes at single CpG level. To modify the 5mC level of the SatIII locus locally we developed a targeted methylation approach. Our targeted approach involves only local changes at the desired CpG simulating a more physiological effect, by contrast to DNMT inhibitors like 5'aza, DNMT gene knockouts or overexpression of DNMTs that induce global chromatin changes. We constructed "DNMT^{act}" consisting of an active DNMT3a3L fused to a TALE targeting the SatIII sequence "TGATTCCATTCCATTCCATT" (name TALE_0 for zero CpGs in target, **Figure 21a**). The target sequence of this TALE-DNMT fusion differs from the target sequence of TALE_2 and other analytical TALEs used later in this study to avoid competition. In addition, it does not contain any CpG to avoid potential binding inhibition by 5mC. As a negative control, we fused this same TALE to a DNMT3a3L construct with the inactivating mutation E756A

("DNMT^{inact}"). Bisulfite conversion and pyrosquencing analysis of genomic DNA of HEK 293T cells expressing DNMT^{act} presented a 6-fold higher methylation level at a relevant CpG from SatIII DNA in comparison to cells transfected with DNMT^{inact} (**Figure 21b**). Positively transfected cells were enriched by sorting cells with similar expression of BFP. This fluorescent protein is used as a reporter gene present in the same vector as the TALE-DNMT fusions. Comparison of the level of expression of BFP and TALE-DNMT fusions showed that similar levels of BFP presented similar levels of the corresponding TALE-DNMT fusion (**Figure 21c-d**). Therefore, by sorting BFP^{+} with the same expression level, we will ensure that these cells are also expressing similar amounts of the TALE-DNMT construct. Since both constructs, DNMT^{act} and DNMT^{inact}, have identical affinities (*163*) and revealed similar expression levels, we can rule out the possibility of differential competition with the fluorescent TALEs employed in subsequent stains. Importantly, TALE_0 of DNMT^{act}/DNMT^{inact} extensively colocalized with foci of TALEs used for later analyses (**Figure 21e**), providing many targets with expected differential methylation levels that can be studied with our imaging method.

Binding of TALEs to their target sequence can be affected by chromatin accessibility. This is especially important when analyzing 5mC, as DNA methylation is generally associated to chromatin compaction leading to potential reduction of target sequence accessibility. In this case, it would not be possible to discern if the reduction of binding by the 5mC sensitive scaffold (HD TALE) is due to increase in methylation level or rather intrinsic reduced chromatin accessibility. To account for this potential source of error we use a co-staining approach including the 5mC unsensitive G* TALE as control that will be affected only by decrease in target accessibility, but not by C/5mC differences. HD TALEs only bind to unmethylated SatIII target sequences, while G* TALEs equally bind to both C and 5mC containing sequences unless DNA accessibility is reduced by chromatin compaction as a consequence of methylation (**Figure 21f**). Therefore, increase in methylation level will be revealed as a decrease in fluorescence intensity of HD TALE, while overall reduced target accessibility would be revealed by negative response of both TALEs.


a
DNMT act construct
Flag Tag
TALE_0
DNMT3a
DNMT3L
5´-
e
TALE_2
TALE_0
TALE_1
TALE_0
TALE_1b
TALE_0
b
*
5mC [%]
inact act
DNMT
c
ns
BFP FI [AU]
inact act
DNMT
d
ns
Anti Flag FI [AU]
inact act
DNMT
f
G* (Universal)
HD (C only)
DNMT
Condensation
= C
= 5mC
= Target
g
G* TALE_2
G* TALE_1
TALE_0
G* TALE_1b
DNMT act inact
DNMT
Fraction of foci
log TALE FI (Norm.)
h
HD TALE_2
HD TALE_1
TALE_0
HD TALE_1b
DNMT act inact
DNMT
Fraction of foci
log TALE FI (Norm.)
i
TALE_2
TALE_1
TALE_0
TALE_1b
log FI (Norm.)
DNMT act inact
DNMT
G*
HD

◄ **Figure 21. Detection of differential levels of 5mC in SatIII locus by pairs of fluorescent TALEs in HEK 293T cells. a)** Schematic representation of TALE-DNMT fusion constructs. **b)** Methylation level at the target CpG in HEK293 cells transfected with either $DNMT^{act}$ or $DNMT^{inact}$ analyzed by bisulfite conversion and pyrosequencing (unpaired t-test; N = 3 experiments, * = $p<0.05$). **c)** Relative expression level of BFP by Flow Cytometry. (N= 4 experiments, ns = non-significant, AU = Arbitrary Units) **d)** Relative expression of the TALE-DNMT constructs in BFP+ cells analyzed by immunostaining of Flag Tag. (N = 4) **e)** Co-staining with indicated pairs of TALEs. TALE_0 represents the target sequence bound by the TALE-DNMT fusions, while the other ones are the analytical TALEs used in this study. **f)** Expected binding of G* and HD TALEs depending on methylation status and chromatin compaction. **g)** Histograms of G* TALE fluorescence intensity (FI) of foci from $DNMT^{act}$/$DNMT^{inact}$ cells. The FI of each focus is log transformed and then normalized by the average of the log transformed mean FI of all foci of the $DNMT^{inact}$ sample **h)** Same as Fig.21g, but with HD TALEs fluorescence intensity. **i)** Box plots showing the data from Fig 21g-h. Unpaired t-test with N = 5, 6, 3, 6 experiments respectively and >1400 foci; * = $p<0.05$; ** = $p<0.01$. Further statistics in SI.

Sorted HEK 293T cells expressing either $DNMT^{act}$ or $DNMT^{inact}$ were co-stained with equimolar concentrations of HD TALE_2 eGFP and G* TALE_2 mCherry. Cell nuclei were stained by DAPI and fluorescent signals of the three channels were recorded by fluorescence microscopy. For comparability, we normalized the fluorescence intensity of each focus by the average fluorescence intensity of the $DNMT^{inact}$ foci. G* TALE_2 presented similar fluorescence levels in both, unmethylated ($DNMT^{inact}$) and methylated cells ($DNMT^{act}$) indicating that there are no changes in target accessibility. By contrast, HD TALE_2 fluorescence level is significantly reduced in the methylated sample, revealing a selective response to 5mC. The sensitivity of HD TALE towards cytosine methylation resulted in a decrease in fluorescence. The same result was obtained with a pair of TALEs targeting the alternative sequence TGGAATCAACCCGAGTA. These TALEs, termed as HD TALE_1 and G* TALE_1 because the target sequence only contains one CpG, proved the applicability of the method in a different sequence context and moreover, single nucleotide resolution. Importantly, none of the versions of pair TALE_0, which targets a CpG-free and thus, non-methylatable sequence, showed no difference for cells expressing $DNMT^{act}$ and $DNMT^{inact}$ constructs. As expected, HD TALE_0 do not show any response because a target sequence lacking CpGs cannot be methylated. Finally, another pair of TALEs targeting a sequence with a single CpG (TALE_1b) showed a trend of reduced fluorescence intensity in $DNMT^{act}$ samples for both HD and G* TALE_1b. A negative

response from both TALEs indicated that the methylation is associated with reduced target accessibility for this particular sequence.

Next, we aimed to study the correlation between changes in 5mC levels and the stress-induced recruitment of heat-shock factor 1 (HSF1) at SatIII DNA. HSF1 is a transcription factor whose recruitment at SatIII locus is the initial step in the nSBs formation under stress conditions like heat-shock, triggering the expression of a long non-coding SatIII RNA (**Figure 22a**). This leads to the formation of nSBs and sequestration of transcription and splice factors as a pathway for global transcriptional and translational down-regulation. HSF1 binds to the consensus sequence nGAAn, which is often preceded by CpG in SatIII DNA, raising the possibility of 5mC control. Indeed, SatIII is hypomethylated and transcriptionally upregulated in several cancers and it can be induced by treatment with 5-azacytidine in HeLa cells. However, functional loss of DNMT is not sufficient to induce SatIII transcription, suggesting a multilayered regulation.

Immunostaining of HSF1 allows to visualize nSBs in cells grown under heat-shock conditions. We observed a stronger stress response in the osteosarcoma cell line U2Os in comparison with HeLa and HEK 293T cells (**Figure S11**). In addition, U2OS cells presented a higher number of nSB foci that extensively colocalize with TALEs designed in this study (**Figure 22b**). This demonstrates the recruitment of TALEs and HSF1 at SatIII locus and the possibility of studying the interplay between 5mC and HSF1 with our method. We incubated U2OS cells expressing either $DNMT^{act}$ or $DNMT^{inact}$ under heat-shock conditions and co-stained them with HD and G* TALE_2 pair combined with endogenous HSF1 immunostaining. We observed a reduced binding of HD TALE as a response to differential target methylation, but no difference in binding of the G* TALE indicating an unaltered overall target accessibility. Interestingly, we detected a slight increase of HSF1 recruitment in $DNMT^{act}$ samples (**Figure 22c**). Histogram analysis revealed that this was due to a subpopulation of cells with high HSF1 in the $DNMT^{act}$ sample (**Figure 22d**). To explore further the influence of 5mC on HSF1 recruitment, we plotted the HSF1 fluorescence intensity (FI) versus the ratio of G* TALE FI over HD TALE FI (as a measure of methylation level, **Figure 22e**). We identified a population of cells with high recruitment of HSF1 at SatIII locus in $DNMT^{act}$ cells which are also presenting a higher G* to HD TALE FI ratio, that is a higher methylation level. This may suggest that 5mC can play a

positive role in heat-shock induced HSF1 recruitment in U2OS cells than can only be detectable by analyzing at single focus and single cell level. This proves the advantages provided by our TALE approach for *in situ* correlation studies of 5mC and other imageable chromatin-interacting proteins.

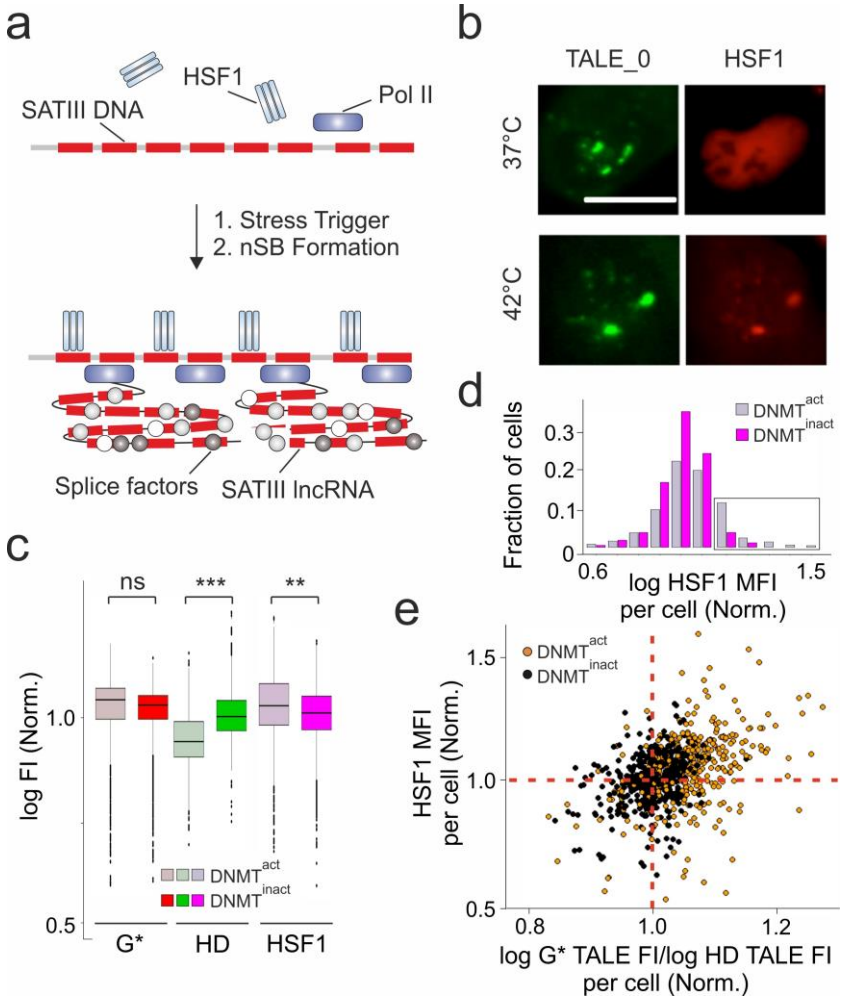


Figure 22. Study of the role of 5mC in heat-shock induced HSF1 recruitment at SatIII locus by combination of TALE co-staining and endogenous HSF1 immunostaining. a) Heat-shock induced formation of nuclear Stress Bodies by expression of lncRNAs from SatIII locus triggered by HSF1 **b)** Live cell imaging of mClover3-TALE_0 and mCherry-HSF1 under heat-shock and physiological temperature in U2OS cells. **c)** Box plots representing the foci FI from $DNMT^{act}$ or $DNMT^{inact}$ cells co-stained with HD and G* TALE_2, and anti-HSF1 antibody. Data normalized as in Fig. 21g-i. Unpaired t-test with N=7 experiments totaling 990 cells; ** = $p<0.01$; *** = $p<0.001$; ns = not significant. See Supplementary Information for further analysis (per cell and foci) and statistics. **d)** Histogram showing HSF1 FI from Fig. 22c analyzed per cell. See Material and Methods for data analysis. **e)** Scatter plot representing HSF1 MFI versus G* TALE FI/ HD TALE FI as a measure of methylation of foci of each cell from Fig. 22d.

To further improve the dynamic range of our system, we aimed to develop new RVDs with 5mC binding capabilities and inhibited by C as opposite to RVD HD. We constructed a size-reduced TALE repeat library containing an amino acid deletion at the RVD position 13 (**Figure 23a-b**). Position 11 was substituted by N and position 12 was randomized with a set of amino acids with polar side chain (**Figure 23a-b**). This strategy was proven to be successful for the discovery of RVDs with positive selectivity for different epigenetically modified bases (*149, 154*).

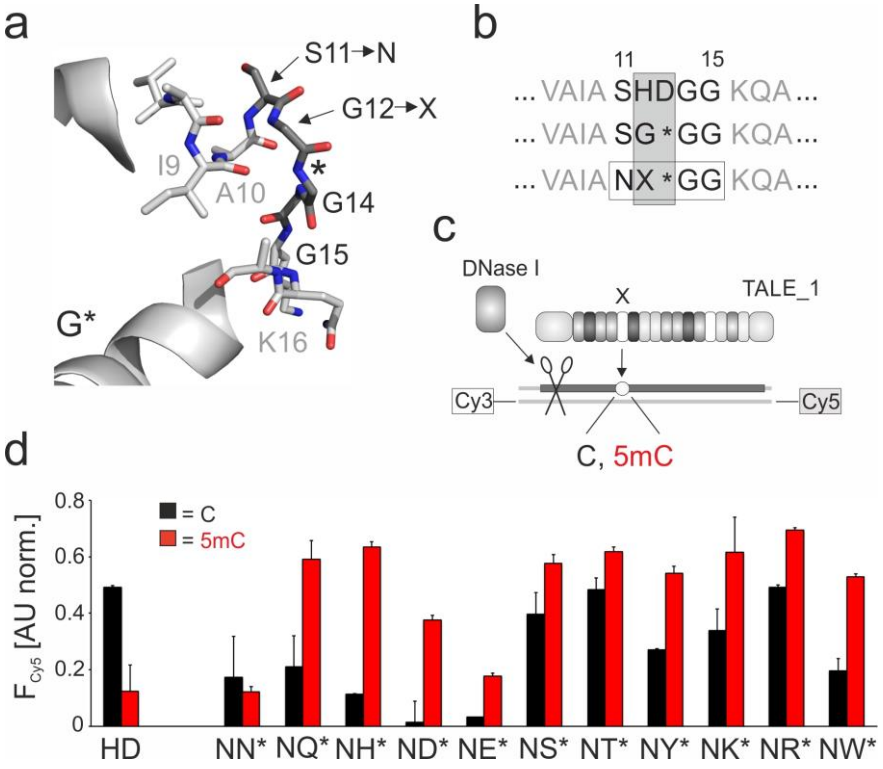


Figure 23. Generation and screening of library NX*. a) Structure of interhelical loop of G* TALE repeat model and positions targeted for mutation or deletion. X = randomized amino acid; * = deletion. **b)** Repeat sequences for RVDs HD, G* and library NX*. RVD position is shown in a grey box. **c)** DNaseI footprinting assay. A doubly labeled Cy3/Cy5 DNA duplex with the TALE target sequence and containing either C or 5mC at the position of interest (circle) is incubated with DNaseI and a TALE carrying the mutant repeat at the analytical position (X). **d)** Results of screening with DNase I competition assay conducted in duplicates with 0.5 µM of each TALE (RVDs indicated below), 0.1 µM of DNA duplex and 1 unit of DNaseI. Cy5 fluorescence level shown was detected 25 minutes after DNaseI addition. Background fluorescence corrected by subtracting the fluorescence of a control without TALE. Data normalized to the fluorescence of a control without DNaseI.

The novel engineered RVDs were assembled at the analytical position of a TALE scaffold containing an N-terminal GFP and a C-terminal Hisx6 tag for expression and purification in *E. coli*. We designed TALEs targeting a CpG containing sequence (TCTTCCGTTTCCACATC-3',

termed TALE_1c) from zebra fish HEY2 gene. Engineered RVD was placed opposite to the C of the single CpG and screened by DNaseI footprinting assay (**Figure 23c**). In this assay, TALEs are incubated with dual labelled Cy3/Cy5 oligonucleotide duplexes with either C or 5mC opposite the engineered RVD. Absence of binding by TALEs leaves oligonucleotides unprotected leading to DNA cleavage by DNaseI, resulting in decreased Förster resonance energy transfer (FRET) from Cy3 to Cy5. HD showed selective protection of the C-containing duplex from DNaseI clevage, but not of the 5mC-duplex, demonstrating once more the sensitivity of HD to methylation (**Figure 23d**). Engineered RVDs presented either no selectivity or different degree of selectivity for 5mC. Previous studies by our group, showed that a single amino acid deletion at position 13 generally helps to accommodate the methyl group, while amino acid at position 12 defines the selectivity by modulating the size and surface structure of the binding pocket (*149*). In the screening we present here, hydroxyl-bearing and basic residues showed low or no selectivity, while aromatic and acidic residues showed high selectivity (**Figure 23d**). Based on these results and additional screenings, we selected repeat ND* and the aromatic repeats NH*, NY* and NW* for further evaluation.

Next, we tested the performance of the selected novel TALE repeats in cellular context by luciferase assay. In this transcription activation assay, we assembled TALE_1c in a mammalian expression vector containing a C-terminal fusion of the transcription activator VP64 (**Figure 24a**). This vector was co-transfected with a reporter plasmid containing a methylated binding sequence of TALE_1c directly upstream of a minimal CMV promoter controlling the expression of firefly luciferase. The methylated binding site contained a single 5mC at the target CpG recognized by the screened RVDs. Quantification of luciferase activity 24 hours post-transfection in HEK 293T cells showed minimal or no signal when using a reporter plasmid without binding site for TALE_1c, in absence of TALE_1c or when using a TALE with the same RVD composition as TALE_1c, but in scrambled order (**Figure 24b**). Data was normalized to the activity induced by the TALE_1c version carrying the natural 5mC permissive RVD NG at the analytical position. TALE with engineered universal repeat G* induced similar expression levels as NG, while ND* TALE_1c and NW* TALE_1c showed lower levels. Interestingly, TALE versions carrying either NH* or NY* revealed higher activation levels, indicating that these novel engineered TALE repeats are potential candidates to be selective 5mC binders also in the

complex environment of the cell nucleus. These results are promising for its potential application in cell staining.

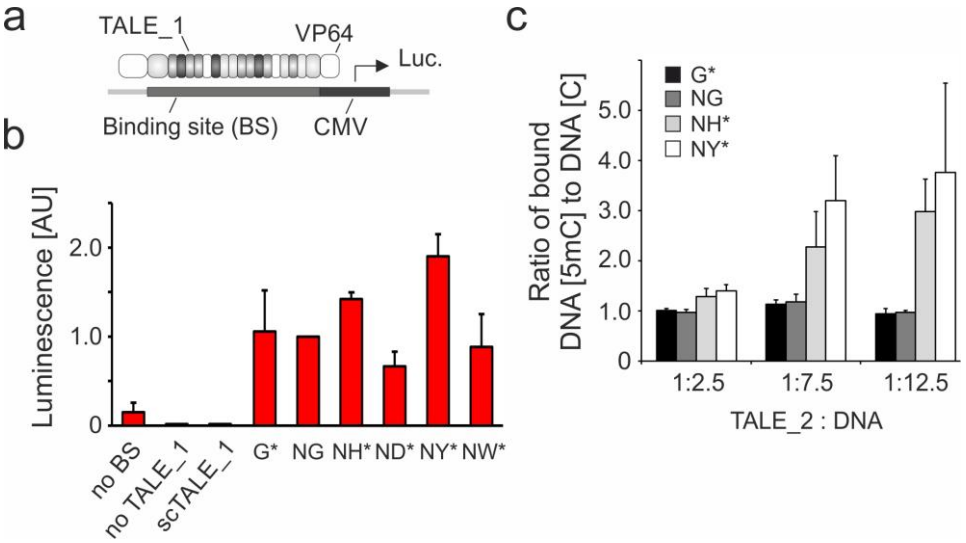
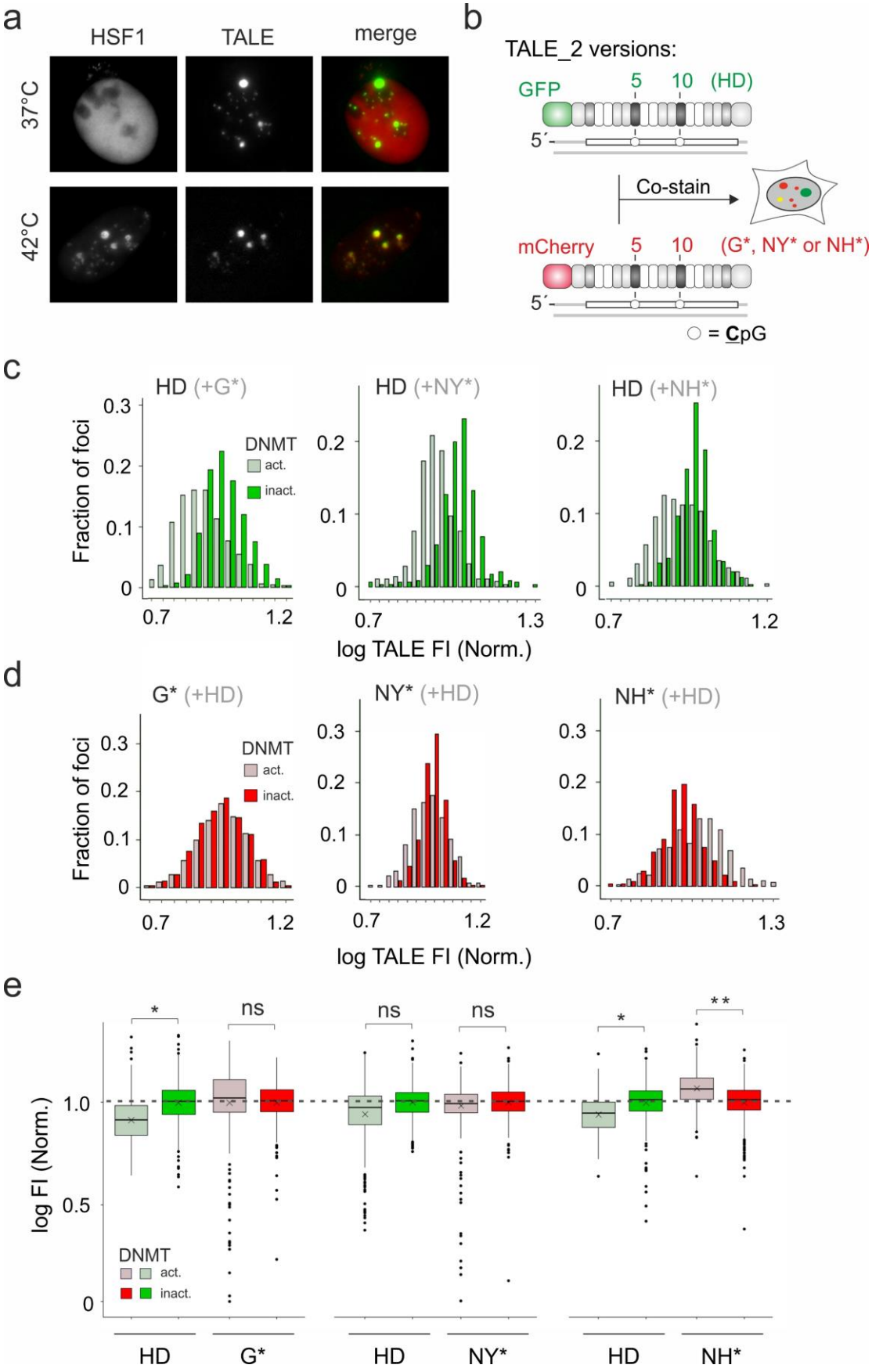


Figure 24. Characterization of potential 5mC binders in live cells and *in vitro*. a) Principle of Luciferase Transcription Activation Assay with TALEs fused to VP64 transcription activator. Luciferase gene is controlled by a minimal CMV promoter downstream the binding site of TALE_1c. **b)** Luminescence data from experiment depicted in Fig.24a. RVDs are shown below. Error bars show standard deviation from three independent biological replicates. **c)** 5mC selectivity of TALE_2 bearing two standard or engineered RVDs opposite to C in the two CpGs of the SatIII target sequence measured by EMSA. Fraction of DNA-bound TALE_2 versions in EMSA was quantified for different stoichiometries of TALE_2 and target DNA duplex containing either 5mC or C, and ratio between the two is depicted. Error bars show standard error from six independent experiments.

We further evaluated RVDs NH* and NY* in a sequence context relevant for cell staining and the application of our imaging-based method; the SatIII locus. We assembled NH* and NY* versions of the previously established TALE_2 targeting a sequence with two CpGs (5'-TGGAACGGAACGGAATG-3'). The different TALE_2 versions with appropriate RVDs were expressed and purified as mCherry fusions and evaluated by EMSA. TALEs were individually incubated with unlabeled DNA duplexes containing either C or 5mC at the two target CpGs and mCherry fluorescence was recorded. TALE versions carrying either G* or NG bound similarly to both methylated and unmethylated DNA (**Figure 24c**). By contrast, NH* and NY* TALEs showed a marked preference for methylated DNA, revealing the promising potential of these two RVDs as 5mC selective TALE repeats in a relevant sequence context.

Novel RVDs NY* and NH* demonstrated 5mC binding preference both *in vitro* and in live cells in two different sequence contexts: HEY2 and SatIII. This opened up their potential applicability in our imaging method to improve 5mC selectivity in the complex landscape of genomic chromatin. To evaluate this, we aimed to apply our established cell staining procedure to analyze methylation level at SatIII DNA sequences in HEK 293T cells. As previously shown in Figure 22b, specificity of TALEs for the SatIII locus was demonstrated by colocalization with immunostained HSF1, a transcription factor that is recruited to SatIII DNA upon heat-shock (**Figure 25a**). To alter 5mC level of SatIII in live cells, we employed our previously explained targeted methylation system based on TALEs fused to either active or inactive DNMT3a3L ($DNMT^{act}$ and $DNMT^{inact}$, respectively). These TALE-DNMT fusions target a CpG-free sequence from SatIII loci and the expression vector also contains BFP as reporter to facilitate cell sorting of positive cells. $DNMT^{act}$ and $DNMT^{inact}$ -treated cells were co-stained with equimolar amounts of HD eGFP TALE_2 and either G*, NY* or NH* mCherry TALE_2 counterpart (**Figure 25b**). Fluorescence from both channels was recorded and the fluorescence intensity of each focus was normalized by the average fluorescence intensity of the $DNMT^{inact}$ sample for each TALE. Co-staining with HD and G* TALE_2s showed, as previously observed, a decrease in fluorescence of the HD TALE in the $DNMT^{act}$ sample, displaying the 5mC sensitivity of RVD HD (**Figure 25c and e**). TALE_2 with universal binder G* showed no differences in methylated and unmethylated samples, indicating unaltered target accessibility (**Figure 25d and e**). Similar trend for HD TALE_2 was observed in HD versus NY* co-staining and in addition, no changes in fluorescence intensity of NY* TALE were detected. Therefore, NY* TALE behaved as a universal TALE similar to G* TALE_2. Co-staining with HD and NH* TALEs showed a significant negative response of HD TALE in the methylated sample ($DNMT^{act}$), demonstrating once more the selectivity of HD TALE towards 5mC increase at the target CpGs. Importantly, NH* TALE_2 showed higher fluorescence intensity for methylated cells, indicating a highly significant positive response for 5mC. This establishes RVD NH* as a 5mC binder whose selectivity has been demonstrated *in vitro*, in live cells and in cellular staining. Furthermore, imaging-based analysis of cellular 5mC with HD versus NH* co-staining will improve the dynamic range and allow to detect a double response by TALEs, in which one of the TALEs will respond negatively (HD) and the other one (NH*) positively.


a
HSF1
TALE
merge
37°C
42°C
b
TALE_2 versions:
GFP
5
10
(HD)
5´-
Co-stain
mCherry
5
10
(G*, NY* or NH*)
5´-
○ = CpG
c
HD (+G*)
HD (+NY*)
HD (+NH*)
DNMT
act.
inact.
Fraction of foci
log TALE FI (Norm.)
d
G* (+HD)
NY* (+HD)
NH* (+HD)
DNMT
act.
inact.
Fraction of foci
log TALE FI (Norm.)
e
log FI (Norm.)
DNMT
act.
inact.
*
ns
ns
ns
*
**
HD
G*
HD
NY*
HD
NH*

◄ **Figure 25. Evaluation of engineered repeats NY* and NH* for 5mC detection at user-defined CpGs by cell imaging. a)** Co-localization of mClover3-TALE_2 and mCherry-HSF1 by live imaging of transfected cells incubated at the indicated temperatures. **b)** Experimental setup for TALE co-staining and evaluation of engineered repeats for 5mC detection by cell imaging. Analytical RVDs are placed at positions 5 and 10. **c)** Histograms of eGFP foci fluorescence intensity from HD TALE_2 in co-staining with, from left to right, TALE_2 G*, NY* and NH*. For each TALE, log FI of each focus is normalized to the mean of the FI of all foci from the $DNMT^{inact}$ cells. **d)** Histograms of mCherry fluorescence intensity of, from left to right, G* TALE_2, NY* TALE_2 and NH* TALE_2 co-stained with HD TALE_2 (corresponding to Fig 25c). Data was normalized as stated in Fig. 25c. Histograms in both Fig. 25 c and d are cropped for clarity and outliers are not shown. All data points including outliers are shown in Fig. 25e. **e)** Box plots showing full data sets from Fig. 25c and d. Unpaired t-test with * = $p>0.05$; ns = not significant. N = 5, 4 and 4 independent biological experiments. For each experiment ~250 cells and >2000 foci were analyzed.

5. Conclusions and Outlook

We have developed an imaging-based method for sequence-specific analysis of 5mC with single nucleotide resolution. Our method exploits the potential of TALEs as programmable DNA binders with 5mC selectivity coupled to their fusion to fluorescent proteins to allow for cell staining and image-analysis. We designed pairs of TALEs targeting a CpG-containing sequence from the clustered and repetitive locus SatIII. Both TALEs of the pair are identical except for the RVD interacting with C at the CpG of interest. One of the TALEs contains the 5mC sensitive RVD HD and is fused to GFP, while the other carries the universal binder G* and is fused to mCherry. Cellular co-staining with both TALEs allows to visualize specific foci and analyze changes in methylation by differences in fluorescence intensity of HD TALE. By contrast, G* TALE is not affected by methylation and thus works as a control TALE whose fluorescence intensity only varies due to changes in target accessibility. This allows to distinguish actual changes in 5mC level from changes in chromatin structure and DNA accessibility whose origin could be different from DNA methylation. Following this principle, we have demonstrated the sensitivity of our method to detect changes in 5mC levels in SatIII target sequences containing one or two CpGs. In addition, we have combined our TALE co-staining with immunostaining to study the impact of DNA methylation on the recruitment of the heat-shock induced transcription factor HSF1 to SatIII sequences. Our results on single cell studies suggested that 5mC has a role on HSF1 recruitment in a small population of cells that presented high levels of HSF1 colocalizing with high methylation on SatIII foci.

Furthermore, we engineered and screened a library of size-reduced TALE repeats with potential 5mC selectivity. This library, NX*, contained a set of repeats with S11N substitution, a randomized amino acid (X) at position 12 and a deletion at position 13 (*). We found that the engineered RVD NH* is a methylation permissive repeat that binds selectively to 5mC, but not C. TALEs with this RVD at the CpG interacting position showed high selectivity and higher affinity for 5mC than other natural and engineered RVDs *in vitro* and in cell staining. Application of our method using an analytical pair of TALEs composed by a HD TALE-eGFP and a NH* TALE-mCherry offered improved image-based analysis with higher dynamic range and dual response to methylation, thanks to the opposite selectivity of each TALE.

In summary, we have developed an imaging-based method for 5mC analysis of user-defined repetitive sequences with nucleotide and strand resolution. By using two fluorescent TALEs with different analytical RVDs, one sensitive to 5mC (HD) and the other one universal (G*), we can detect changes in methylation level unambiguously and dissect them from mere changes in overall chromatin accessibility by contrast to other imaging approaches (*82–91*). Furthermore, our strategy based on sequence-specific fluorescent probes provides positional and topological information of the locus of interest and it is applicable for single cell and single locus *in situ* studies. Our method is also compatible with conventional immunostaining allowing to study the interplay between 5mC and other imageable chromatin features like transcription factors or epigenetic readers. In addition, thanks to the use of purified TALEs, it is potentially applicable to non-transfectable cell lines and clinical tissue samples. Finally, we developed RVD NH*, an engineered size-reduced TALE repeat that selectively binds to 5mC allowing for improved imaging-based analysis with increased dynamic range and response to methylation from both TALEs when combined with a HD TALE counterpart.

Our approach opened a new avenue in the study of 5mC and its role in shaping chromatin structure and function by cell imaging. However, until now it is only applicable to the analysis of highly repetitive and clustered sequences due to detection limits. Further development of our method will focus on enhancing the output signal and improve the signal-to-noise ratio to allow visualization of 5mC in sequences with low number of repeats and non-repetitive sequences. This will enable, for example, the study of promoter methylation status with a small subset of TALEs and its correlation with transcriptional activity. Next, our efforts will focus on the application of our method in live cells to examine the role of DNA methylation in dynamic epigenetically controlled processes like cell differentiation and development. This will require to explore the introduction of purified TALEs by electroporation or protein transfection to control protein levels precisely. Finally, it is expected that our method can be adapted for the detection of oxidized 5mC species, namely 5hmC, 5fC and 5caC. Specific engineered RVDs for the detection of these other epigenetically modified bases have been previously reported by our group and others (*134, 149, 154, 155*). A fluorescent TALE carrying the respective specific RVD at the CpG-interacting position is expected to be a viable option for the detection of oxi-5mC bases.

6. Material and Methods

Material and Methods were previously published in (*2*) by Wiley Co. (CC 4.0).

6.1 Plasmid cloning

Plasmid pÁlM1577 for expression of TALEs fused to N-terminal mCherry (for vector maps, see Supplementary Information) was generated by amplifying mCherry from pAnW1272 using primers o2994 and o2995 (see oligo table for primer sequences). PCR fragment and vector pAnI521 (*159*) were digested with NdeI and BsrGI enzymes and then ligated with T4 ligase (New England Biolabs, M0202T) using 3:1 insert:vector molar ratio.

pJaW876 was cloned by Gibson assembly (*164*). For this, vector pcDNA3.1-GoldenGate-VP64 (Addgene, #47389) was linearized by PCR using primers o2079 and o2080 and the insert DNMT3a3L CD was amplified by PCR using pET28-Dnmt3a3L-sc27 (Addgene, #71827) and o2071 and o2072 primers.

Vector pÁlM1285 was created by amplifying CMV-EBFP2 from plasmid EBFP2-N1 (Addgene #54595) using primers o2733 and o2734. Vector pJaW876 was digested using MfeI and previously amplified CMV-EBFP2 fragment was inserted by Gibson Assembly resulting in pÁlM1285. Mutation E756A for catalytically inactive DNMT3a3L was introduced by site-directed mutagenesis using primers o2038 and o2039 to generate pÁlM1560 vector.

The final entry vectors for Golden Gate Assembly pcDNA3.1-GoldenGate-VP64 (Addgene, #47389) and pTALYM3 (Addgene, #47874) were obtained from Addgene.

Vectors pAni521 and pÁlM1577 were used for expression in E. coli and purification of TALEs fused to eGFP or mCherry, respectively. Vectors pÁlM1285 and pÁlM1560 are for mammalian expression of TALEs fused to DNMT3a3L either wildtype or with E756A inactivating mutation, respectively. These two vectors contain an EBFP2 gene for sorting of transfected cells. Vector pTALYM3 was used for expression of TALEs fused to N-terminal mClover3 in mammalian cells for live-cell imaging and colocalization studies. Vector pcDNA3.1-GoldenGate-VP64 allows for mammalian expression of TALEs fused to VP64 transcriptional activator for targeted gene expression induction. Generation of fLuc vectors with singly methylated TALE target sequence for transcription activation assays is described in section of luciferase assay.

Library of G* RVD modules was prepared as follows: pG*1 module was generated by QuikChange site-directed mutagenesis (Agilent) using o2166 and o2167 primers with pHD1 vector. Modules pG*2 to pG*9 were derived from pHD2 to pHD9 (*110*) module vectors by site-directed mutagenesis using o2168 and o2169 primers. Vector pG*10 was created by cassette mutagenesis by digesting pHD10 with NcoI-HF and BsoBI restriction enzymes and ligating the double strand fragment generated by annealing o2142 and o2143 oligos.

Library NX* repeat modules for repeat position 5 (pNX*5) was generated as previously described (*149, 154*) by restriction-ligation using plasmid pHD5 as template. Briefly, pHD5 vector and respective annealed oligos (see SI) were digested with NcoI and XhoI (New England Biolabs) and ligated with T4 DNA ligase.

Repeat modules for position 10 (pNY*10 and pNH*10) were generated by Quikchange site-directed mutagenesis (Agilent) with oligos o3541/o3542 and o3553/o3554 respectively, using pNN10 as template.

TALEs were assembled by Golden Gate Assembly as previously described (*110*) (see TALEs assembly table for detailed RVD composition) using pAnI521 or pÁlM1577 as entry plasmids in Golden Gate 2 reactions, resulting in plasmids coding for the respective TALE proteins in frame with a C-terminal His6x tag and a N-terminal eGFP or mCherry, respectively. For mammalian live-cell imaging, TALEs were assembled in vector pTALYM3 (Addgene, #47874). For site-specific DNA methylation, TALEs were assembled in p1285 or p1560 for TALE-DNMT^{active} or TALE-DNMT^{inact} fusions, respectively.

HSF1 gene was amplified from HEK293T cDNA with primers HSF1-Bgl1-fw and HSF1-Sal1-rv. For cloning into pmCherry–C3 (Shaner 2004)(*165*), vector and insert were digested with Sal1 and Bgl1, and re-ligated in presence of the inserts with T4 Ligase at 16°C, overnight in order to obtain pHSF1-mCherry.

6.2 Cell culture and transfection

HEK 293T, HeLa and U2OS (Sigma Aldrich, 92022711-1VL) cells were cultured at 37°C and 5% of CO_2 in DMEM medium (PanBiotech, P04-03609) supplemented with 10% FBS (PanBiotech, P30-3302), 1% of L-Glutamine 200mM (PanBiotech, P04-80100) and 1% of Pen/Strep (PanBiotech, P06-07050).

For site-directed methylation, cells were seeded on 10 cm diameter dishes (Sarstedt, 83.3902.300) and transfected with 10 µg of $DNMT^{act}$ or $DNMT^{inact}$ (TALE_0 and TALE_1C assembled in p1285 and p1560) plasmid using FuGene 6 Transfection Reagent (Promega, E2691) following the manufacturer's protocol. The reagent:DNA ratio used was 3:1 in every case.

For live-imaging and colocalization studies of HSF1 and SatIII-TALEs after heat-shock, 300.000 U2OS cells were seeded on a µ-Dish 35 mm (ibidi, 81156) and transfected with 1 µg of TALE_0 assembled in pTALYM3 vector and 500 ng of pHSF1-mCherry as mentioned above.

For transfection regarding Luciferase Assay, see Luciferase Assay section.

6.3 Flow cytometry and cell sorting

Cells were washed with DPBS, trypsinized with Trypsin 0.05% / EDTA 0.02% (PanBiotech, P10-038100) for 5 minutes at 37°C and blocked with Full DMEM medium. Samples were pelleted by centrifugation at 300 g for 10 minutes, the supernatant was discarded and the cell pellet resuspended in 500 – 1000 µl of DPBS and transferred to a 5 ml FACS tube through a cell strainer.

Cells were sorted with a Sony Cell Sorter model LE-SH800SFP in targeted mode using the 405nm laser (filter FL1 450/50, Optical Filter Pattern 2) for detection of EBFP2 from DNMT-TALE plasmids. Gates were set to assure similar expression levels of fluorescent protein in both: samples transfected with $DNMT^{act}$ or $DNMT^{inact}$ plasmids (TALE_0 and TALE_1C assembled in p1285 and p1560). EBFP2+ cells were collected in tubes with Full DMEM + 2% of Pen/Strep (PanBiotech, P06-07050), counted and seeded in µ-Plate 96 Well Black ibiTreat, tissue culture treated plates (ibidi, 89626) previously coated with 0.02% of poly-L-Lysine (Sigma Aldrich,

P1274-100MG). Multicolor flow cytometry was compensated using single stained samples for each fluorophore.

6.4 TALE expression and purification

TALEs were expressed and purified as previously described (*151*). Briefly, TALE plasmids were transformed into electrocompetent BL21 DE3 Gold *E.coli* cells and grown overnight at 37°C on carbenicillin (Carb, 100 mg/ml) LB agar plates. Single colonies were picked and grown in 5 mL LB supplemented with 100 mg/ml of carbenicillin for 3 hours at 37°C and 220 r.p.m. This starter culture was transfer to a flask containing 250 mL of LB + Carb and grown under the same conditions until the OD^{600} reached 0.8 arbitrary units (au), followed by 0.5 mM IPTG induction and incubation overnight at 22°C and 220 r.p.m. Cultures were centrifuged at 3000 g and 4°C for 20 minutes. Supernatant was discarded and pellets were kept at -20°C during 2 hours. Pellet was resuspended in 10 mL of Deep Lysis Buffer (10 mM Tris-HCl, 300 mM NaCl, 2,5 mM MgCl2, 5 % DMSO, 0.2 % sodium lauroyl sarcosinate (AppliChem), 0.1 % Triton X-100, pH = 9) containing 1 mM PMSF, 1 mM of DTT and 50 µg/mL lysozyme (Sigma Aldrich). The suspension was lysed by repeated sonication on ice (2 runs of 3 minutes; 20 % amplitude; 4s on, 2s off), letting samples cool on ice for 1 minute between runs. Samples were centrifuged at 14000 g and 4°C for 20 minutes to remove cell debris. Supernatant was then incubated with 1 ml HisPur™ Ni-NTA Resin (ThermoFisher Scientific, 88221) overnight spinning at 4°C. Beads were then washed twice with PBS, three times with Lysis buffer + 20 mM Imidazole + 1mM DTT and twice with Lysis buffer + 50 mM Imidazole + 1mM DTT. Elution of the TALEs was performed by incubating the beads with 1mL of Lysis buffer + 500 mM Imidazole + 1mM DTT overnight. Samples were centrifuged at 12000 g for 5 minutes. Supernatant was taken and load into a dialysis cassette (Thermo Scientific, 66381). Samples were dialyzed in TALE Storage Buffer (200 mM NaCl, 20 mM Tris, 10% glycerol, pH = 7.5) + 1 mM DTT stirring at 4°C and exchanging the buffer every 2 hours for 5 times. Last dialysis step was performed overnight. Samples were centrifuge at 12.000 g and 4°C for 5 minutes and supernatant was aliquot, snap-froze with liquid nitrogen and stored at -80°C. Protein concentration was measured by BCA using Microplate BCA Protein Assay Kit – Reducing Agent Compatible (ThermoFisher Scientific, #23252) following manufacturer's instructions.

6.5 TALE staining

20.000 HEK 293T or HeLa cells or 17.000 U2OS cells per well were seeded on a µ-Plate 96 Well Black ibiTreat tissue culture treated plates (ibidi, 89626) and incubated overnight prior fixation. Cell were washed once with DPBS (PanBiotech, P04-361000) and then fixed with ice-cold methanol at -20ºC for 10 minutes. After washing for 5 minutes with DPBS, cells were treated with 2N HCl for 5 min at room temperature, followed by two washes with DPBS and incubation overnight with blocking buffer (DPBS-T + 1% BSA). Samples were then stained with 200 µl 0.8 nM of each (eGFP and mCherry) purified TALEs in TALE Binding Buffer (20 mM Tris-HCl pH= 8, 50 mM NaCl, 5 mM $MgCl_2$, 50 ng/µl Salmon Sperm DNA, 0.1 mg/ml BSA and 10% glycerol) for TALE_1, TALE_1b and TALE_0 and DPBS + 50mM of NaCl for TALE_2 at room temperature for 30 minutes in the dark. After incubation, each well was washed four times with the respective buffer for 5 minutes at room temperature and shaking at 300 rpm. Plates were kept in DPBS + 5% FBS overnight at 4ºC. Nucleus staining was performed by incubating the samples with 1 uL per well of Vectashield with DAPI in 200 µl of DPBS (Vector Laboratories, H-1200) for 10 minutes at room temperature. Finally, cells were washed twice with DPBS for 5 minutes at room temperature and kept in DPBS for microscopy.

6.6 Co-staining with TALEs and HSF1 antibody

Cells were incubated under heat-shock conditions for 2 hours at 43ºC and immediately after, they were washed with DPBS and then fixed with formaldehyde 1% for 5 minutes at room temperature, followed by 10 minutes at 4ºC. After washing three times with DPBS-T (0.05% Tween-20), samples were incubated with ice cold methanol for 10 minutes at -20ºC and then washed again. Cells were blocked with blocking buffer (DBPS with 0.05% of Tween-20 and 1% of BSA) overnight at 4ºC and then treated with HCl 2N for 5 minutes at room temperature. Samples were thoroughly washed with DPBS-T three times and then incubated with 1:500 dilution of anti-HSF1 antibody (Cell Signaling Technology, 4356S) in blocking buffer for 1 hour at room temperature, followed by three washes and incubation with 1:250 dilution of Alexa Fluor 750 secondary antibody (Invitrogen, A21039) under the same conditions. After washing, cells were incubated with TALEs as described above. Each well was washed four times with the

respective buffer for 5 minutes at room temperature and shaking at 300 rpm. Finally, nucleus staining was performed as described above.

6.7 FLAG-tag immunostaining

HEK 293T cells were trypsinized for 5 minutes at 37°C, centrifuged down and fixed using the FIX & PERM Cell Permeabilization Kit (ThermoFisher Scientific, GAS003) following manufacturer's instruction. Then samples were first blocked with 1% of BSA in DPBS-T for 1 hour at room temperature and then incubated with primary anti-FLAG (Sigma-Aldrich, F1804-200UG) in blocking buffer for 1 hour at room temperature. After washing three times, staining with secondary Alexa Fluor 488 antibody (Invitrogen, A11001) was performed under the same conditions. Cells were washed and resuspended in FACS buffer (DPBS + 2% FBS + 2 mM EDTA) for flow cytometry.

6.8 Bisulfite conversion

For the bisulfite conversions the EpiTect® Bisulfite Kit (QIAGEN, 59104) was performed according to the protocol of the manufacturer. In brief, for each reaction 200 – 250 ng DNA was used. After the bisulfite conversion the solution was loaded onto EpiTect spin columns and centrifuged at maximum speed (13,400 x g). After discarding the flow through and washing, 500 µL of desulfonation buffer was applied to each column. After 15 minutes of incubation at room temperature, the columns were centrifuged again for one minute at max. speed. After two further washing steps and one dry centrifugation the samples were eluted in 20 µL of elution buffer.

6.9 Pyromark PCR

Pyromark PCR was done with the Pyromark PCR kit (QIAGEN, 978703). The reaction mixture for the PCR was prepared according to the manufacturer's protocol. A final primer concentration of 0.2 µM of oligos SatIII-Pyro-fw and SatIII-Pyro-rv was applied and 10 ng of template DNA were added.

6.10 Pyrosequencing

The Pyromark PCR product was bound to sepharose beads on a 96-well plate utilizing the biotinylation of either forward or reverse primers. The bound PCR products were shaken on a plate shaker for five minutes at room temperature. In the meantime, 10 µM of sequencing primers were diluted in annealing buffer and added onto a separate PSQ 96-well plate. A vacuum filter station was then used for the washing of the PCR product bound to the sepharose beads. The filters were first washed with water before being immersed into the DNA solution on the first 96-well plate. The filters were then immersed into 70% ethanol, 0.2 M NaOH and washing buffer containing 10 mM tris-acetate (pH 7.6). Afterwards, the vacuum was turned off and the filters were placed into the PSQ 96-well plate containing the diluted sequencing primers, which led to the detachment of the sepharose beads from the filters. After letting the filters stand in the solution for several minutes, the plate, now containing both, the primers as well as the DNA templates, was placed on a heating block for 2 minutes at 85°C. After the heating, the plate was kept at room temperature. The pyrosequencing reaction was performed in a PSQ HS 96ATwo Pyrosequencer and analysed using the PSQ HS 96A software.

6.11 Electromobility shift assay (EMSA)

6.11.1 EMSAS for nucleotide and strand resolution performance of TALEs

Electromobility Shift Assays were performed as previously reported (*157*) using Cy5-labelled oligo o3546 and o3552. Briefly, pairs of oligos were hybridized in Annealing Buffer (20 mM Tris, 50 mM NaCl, 5 mM MgCl2 and 5% v/v glycerol, pH = 8) by incubating for 5 minutes at 95°C, followed by 30 minutes at room temperature. Then, 7.5 pmol of annealed oligos were incubated with 1.5 pmol of TALEs in TALE Binding Buffer (see above) for 1 hour at room temperature and 30 minutes at 4°C. In case of competitive EMSA, equimolar amounts (1.5 pmol) of both HD-EGFP and G*-mCherry TALEs were incubated with unlabeled annealed oligos. The native polyacrylamide gel (0.5 x TAE buffer, 8 % Rotiphorese gel 40 (Carl-Roth), 0.1 % APS and 0.01 % TEMED) was pre-run at 4 °C for 30 minutes in a Mini Protean vertical electrophoresis cell (Bio-Rad) at 70 V. The gel was loaded with 7,5 µl sample and then run for 90 minutes at 4 °C and 70 V. GFP fluorescence of the gels was recorded with a Typhoon FLA-

9500 laser scanner (GE Healthcare) and analyzed using the software ImageQuant TL 8.1 (GE Healthcare).

6.11.2 EMSAs for screening of potential 5mC-binding engineered TALE repeats

EMSAs were performed as mentioned above with some variations. A methylated (o3545) or an unmethylated (o3552) oligo with the target sequence of SatIII TALE as forward strand was hybridized to a complementary unmethylated reverse oligo (o3547) in Annealing Buffer (20 mM Tris, 50 mM NaCl, 5 mM MgCl2 and 5% v/v glycerol, pH = 8) by incubating 5 minutes at 95 °C, followed by a cool-down to room temperature for 3 hours. Annealed oligos at 15 to 400 nm concentration were incubated with 200 nM of the respective m-Cherry-TALE in TALE Binding Buffer (20 mM Tris-HCl pH= 8, 50 mM NaCl, 5 mM MgCl2, 50 ng/µl salmon sperm DNA, 0.1 mg/ml BSA and 10 % glycerol) in 10 µl final volume. This mixture was incubated 1 hour at room temperature and 30 minutes at 4 °C in the dark. Samples were run on pre-run native polyacrylamide gels (0.5 x TAE buffer, 8 % Rotiphorese gel 40 (Carl-Roth), 0.1 % APS and 0.01 % TEMED) for 90 minutes at 4 °C in a Mini Protean vertical electrophoresis cell (Bio-Rad) with a voltage of 70 V. mCherry fluorescence was recorded with a Typhoon FLA-9500 laser scanner (GE Healthcare) using a 532 nm laser and LPG filter. Images were analyzed using ImageQuant TL 8.1 software (GE Healthcare).

6.12 Library screening by DNaseI Footprinting Assay

Assays were performed in 384-well plate format (Greiner Bio-one) as previously described. Briefly, HEY2 gene target sequences either methylated (o465) or unmethylated (o476) were hybridized with the reverse complementary oligo o1892 (5′-Cy5- and 3′-Cy3-labeled) at a concentration of 200 nM in 3 µL Hybridization buffer (40 mM Tris-HCl (pH = 8.0), 100 mM NaCl, 10 mM MgCl2, 0.2 mg/mL BSA, 10 % glycerol) by incubation at 95 °C for 5 min and then at room temperature for 30 min. TALE proteins were added in 3 µl TALE Storage Buffer (200 mM NaCl, 20 mM Tris-HCl (pH = 7.5), 1 mM DTT, 10 % glycerol) to result in 0.5 µM final concentration and were incubated 30 min at room temperature. 6 µL of a mixture of 1 U DNase I (New England Biolabs) in DNase I buffer (20 mM Tris-HCl (pH = 7.5), 5 mM MgCl2 and 0.2 mM CaCl2) were added per well and the plate containing the mix was placed immediately into

a TECAN M1000 plate reader, pre-heated at 37 °C. Excitation of Cy3 was performed at 552 nm and Cy5 Emission was acquired at 665 nm every 5 minutes over one hour. Background Cy5 fluorescence was subtracted from control wells lacking TALE, and the ratio of Cy5 fluorescence of the TALE samples to that of a control without DNase I were plotted as relative Cy5 fluorescence.

6.13 Luciferase Assay

Luciferase reporter plasmids were generated as previously reported. Oligonucleotides o2520/o2501 containing the TALE target sequence-containing methylated at the single CpG were hybridized at 10 μM each in 150 mM NaCl by heating at 95 °C for 5 min and cooling down at an interval of 3 °C / min to 10 °C. Hybridized oligos and vector pAnW755 (see SI) were digested with SalI and SpeI (New England Biolabs) and purified using a PCR purification kit (Thermo Fisher Scientific). Digested inserts and vector were ligated with T4 ligase (New England Biolabs,) at 3:1 insert:vector ratio for 4 hours at 16 °C. Products were directly used for transfection of HEK 293T cells. 1.4 · 104 cells HEK 293T cells were seeded in a 96 well plate (Merck) one day before transfection. Co-transfection was performed with 100 ng of luciferase reporter plasmid and 100 ng of the respective TALE-VP64 plasmid using Lipofectamine 2000 Transfection Reagent (Thermo Fisher Scientific) with a reagent:DNA ratio of 3:1, following the manufacturer's protocol. For each condition, three replicates were prepared. Cells were lysed 24 h post transfection with lysis buffer containing 100 mM NaH2PO4 and 0.2 % Triton X 100. After incubation on ice for 20 min, 20 μL of lysate were mixed with 90 μL of premixed Bright-GloTM Luciferase reagent (Promega) in a second 96 well plate. The luminescence of each well was analyzed by a Tecan Infinite M1000 plate reader (wavelength 380-600 nM). Luminescence data for different TALEs were normalized to the reaction of TALE NG.

6.14 Microscopy

Microscopy was performed with an Olympus IX81 microscope coupled with a Hamamatsu model C10600-10B-H camera. Pictures were taken as z-stack images of 6 µm (step size = 0.3 µm) using a 60x oil objective covered with immersion oil (ibidi) compatible with µ-Plate 96 Well Black ibiTreat, tissue culture treated plates (ibidi). Excitation settings for each channel were as follows: DAPI was acquired using the DAPI excitation filter (387/11), EGFP with excitation filter GFPFret (470/22) and mCherry with Cy3(560/25). The three fluorescence channels were detected using triple band dichroic DaFICy3 cube and the DaFICy3 quad band emission filter. When using Alexa Fluor 750, a fourth channel was added and images were acquired using Cy7 HC Filterset cube (F36-577) and Cy5-Cy7 emission filter (ET700/75). Exposure times were 30, 100 and 50 ms for DAPI, EGFP and mCherry respectively. For HSF1 and TALEs co-staining 30, 200, 100, 100 ms exposure times were used for DAPI, EGFP, mCherry and Alexa Fluor 750, respectively. For live-cell imaging of co-transfections with mClover3-TALE and HSF1-mCherry, plates were either incubated at 37 ^{o}C (control) or at 42 ^{o}C (heat-shock) for 3 hours before microscopy. Then, cells were immediately imaged as described above. Microscope chamber was pre-warmed at the appropriate temperature with a heat unit.

6.15 Image processing and analysis

The intensity, area and subcellular localization of foci was analyzed from maximal intensity Z-projections of image stacks (1344 x 1024 pixels, 12 bits) using the Fiji distribution of ImageJ(*166, 167*) as follows: The mean background intensity of an out-of-interest region was subtracted from each channel of the stack, and the nuclear regions selected from the DAPI channel (10 µm^2 minimum area, circularity between 0.5–1.0). We then used the "GaussFit OnSpot" plugin(*168, 169*) to analyze the intensity and size of the spots recorded in the GFP/FITC using elliptical shape and Levenberg Marquard fit mode with a 10 pixel rectangle half size. In order to select foci-like objects from this set, spots outside the nuclear regions or with a prominence smaller than 30 (signal-to-noise ratio) or spots larger than 25 pixels were excluded. Mask (ROIs) generated by foci detection from GFP/FITC channel were directly

applied to mCherry and Cy7 (when applicable) channels to measure the mean fluorescence intensity. The images were processed in batch using an ImageJ macro script. For each nucleus, the number, size and intensity of the associated foci was recorded.

6.16 Data analysis and statistics

The data was analyzed and plotted using R(*170, 171*). DAPI areas larger than 25000 square pixels were filtered out to exclude unsplit nuclei. For each TALE, mean fluorescence intensities of each focus were log transformed and then normalized to the average fluorescence intensity of all foci of the $DNMT^{inact}$ sample of each independent experiment. Single cell data was performed by calculating the average mean fluorescence intensity of all foci within a single nucleus and then, log transformed and normalized as explained above. Graphs were plotted using ggplot2 library (*172*) and statistical analyses were performed using GraphPad to carry out unpaired t-test considering number of independent experiments as sample size ($N \geq 3$ independent experiments in every case). Statistical t-test analyses by cell and by foci were performed with ggpubr (*173*). Significance of unpaired t-test studies and sample sizes for each experiment are shown in the tables below:

Sample Size			
	N (# of experiments)	# of cells	# of foci
TALE_2	5	751	2560
TALE_1	6	842	1484
TALE_1b	6	869	2764
TALE_0	3	248	1444
HSF1 experiment	7	990	2969

	Significance for unpaired t-test								
	Biological Replicates			Cells			Foci		
	HD	G*	HSF1	HD	G*	HSF1	HD	G*	HSF1
TALE_2	*	ns	-	****	ns	-	****	ns	-
TALE_1	**	ns	-	****	ns	-	****	ns	-
TALE_1b	ns	ns	-	****	****	-	****	****	-
TALE_0	ns	ns	-	ns	ns	-	ns	ns	-
HSF1 experiment	***	ns	**	****	ns	***	****	ns	****

Sample sizes and number of experiments of staining experiments

Sample Size			
Co-staining	N (# of experiments)	# of cells	# of foci
HD vs G*	5	751	2560
HD vs NY*	4	725	3738
HD vs NH*	4	739	4086

Statistical data. Significance and p-values for unpaired t-test

HD + G* (N=5)	
	p-value
HD	* 0.0110
G*	ns 0.8907

HD + NY* (N=4)	
	p-value
HD	ns 0.2331
NY*	ns 0.3561

HD + NH* (N=4)	
	p-value
HD	* 0.0486
NH*	** 0.0036

7. Supplementary Information

7.1 Supplementary Tables

7.1.1 Oligos tables

7.1.1.1 Oligos for vector construction

Oligo name	Sequence (5' → 3')
o2994	TATACATATGGTGAGCAAGGGCGAGGAG
o2995	AGCGTAGTCCGGAACGTCG
o2079	ATTAACTACCCGTACGACGTTC
o2080	GGCGCGCCCAACTTTGCGTTT
o2071	AAACGCAAAGTTGGGCGCGCCAACCATGACCAGGAATTTGACC
o2072	GAACGTCGTACGGGTAGTTAATAAGAGGAAGTGAGTTTTGAG
o2733	GAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACGCGTTACATAACTTACGGTAAATGG
o2734	AGCAGCGCAAAACGCCTAACCCTAAGCAGATTCTTCATGCGCTTACTTGTACAGCTCGTCC
o2038	CCCTTCTTCTGGCTCTTTGCCAATGTGGTGGCCATGGGCG
o2039	CCATGGCCACCACATTGGCAAAGAGCCAGAAGAAGGGGCG
HSF1-Bgl1-Fw	TTTAAGTAGATCTGATCTGCCCGTGGGCCCCGG
HSF1-Bgl1-Rv	TTTAAGTGTCGACCTAGGAGACAGTGGGGTCC

7.1.1.2 Oligos for generation of G* repeat modules 1-10 by QuikChange site-directed mutagenesis or restriction-ligation

Oligo name	Sequence (5' → 3')
o2166	GGTCTCGCTATCGCCAGCGGCGGCGGCAAG
o2167	GCCGCCGCCGCTGGCGATAGCGAGACCTC
o2168	GCTATCGCCAGCGGCGGCGGCAAGCAAGCG
o2169	TTGCTTGCCGCCGCCGCTGGCGATAGCCAC
o2142	CATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCGGAGGCGGGAGACCC
o2143	TCGAGGGTCTCCCGCCTCCGCTGGCGATAGCCACCACTTGGTCCGGAGTCAGGC

7.1.1.3 Primers for pyrosequencing of bisulfite converted SatIII locus

Oligo name	Sequence (5' → 3')
SatIII-Pyro-Fw	GGAATGGATTCAACTTGAATG
SatIII-Pyro-Rv	TTCCATTCCATTCCTATACT

7.1.1.4 Oligos for library construction of repeat module 5 by restriction-ligation

Oligo ID	Description	Sequence (5' → 3')
o1691	VAIA**NN***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACAATGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1692	VAIA**NQ***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACCAGGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1693	VAIA**NH***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACCATGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1694	VAIA**ND***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACGATGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1695	VAIA**NE***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACGAAGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1696	VAIA**NS***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACAGCGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1697	VAIA**NT***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACACCGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1698	VAIA**NY***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACTATGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1699	VAIA**NK***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACAAAGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1700	VAIA**NR***GG	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACCGTGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1701	VAIA**NW***G G	TTTTCCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAACTGGGGC GGCAAGCAAGCGCTCGAAACGGTGCAGGAG
o1746	Reverse primer	TTTTCTCGAGGGTCTCCTGCACCGTTTCGAGCGCTTG

7.1.1.5 Construction of modules pNY*10 and pNH*10 by Quikchange

Oligo ID	Description	Sequence (5' → 3')
o3541 fw	VAIA**NY***GG	GCTATCGCCAACTATGGCGG
o3542 rv	VAIA**NY***GG	CCGCCATAGTTGGCGATAGC
o3553 fw	VAIA**NH***GG	CCGCCATGGTTGGCGATAGC
o3554 rv	VAIA**NH***GG	GCTATCGCCAACCATGGCGG

7.1.1.6 DNaseI Footprinting Assay target sequence oligos

Oligo ID	Description	Sequence (5' → 3')
o476	Unmethylated	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTT CCACCGTGAGCTCTTC**C**GTTTCCACATCC
o465	Methylated	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTT CCACCGTGAGCTCTTC**N**GTTTCCACATCC **N**= 5mC
o1892	Reverse oligo	XGGATGTGGAAA**C**GGAAGAY **X**= Cy5, **Y**= Cy3

7.1.1.7 Luciferase Assay target sequence oligos

Oligo ID	Description	Sequence (5' → 3')
o1899	Unmethylated	TTTTGTCGACTCTTC**C**GTTTCCACATCTACTAGTTTTT
o1900	Unmethylated	AAAAACTAGTAGATGTGGAAA**C**GGAAGAGTCGACAAAA
o2520	Mehtylated	TTTTGTCGACTCTTC**N**GTTTCCACATCTACTAGTTTTT **N**= 5mC
o2501	Mehtylated	AAAAACTAGTAGATGTGGAAA**N**GGAAGAGTCGACAAAA **N**= 5mC

7.1.1.8 EMSA target sequence oligos

Oligo ID	Description	Sequence (5' → 3')
o3552	Unmethylated	CCAATGGAA**C**GGAA**C**GGAATGGAATG
o3545	Methylated	CCAATGGAA**N**GGAA**N**GGAATGGAATG **N**= 5mC
o3547	Reverse	CATTCCATTCCGTTCCGTTCCATTGG
o3546	Reverse	NCATTCCATTCCGTTCCGTTCCATTGG N=Cy5

7.1.2 TALEs Assemblies

TALE_2 Length Optimization

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
29	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN	NI	NI
25	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG				
23	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN	NI						
20	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG									
17	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN												
15	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG														
12	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI																

TALE_2 Selectivity

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
HD/HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN
G*/G*	NN	NN	NI	NI	G*	NN	NN	NI	NI	G*	NN	NN	NI	NI	NG	NN	NN
NN/NN	NN	NN	NI	NI	NN	NN	NN	NI	NI	NN	NN	NN	NI	NI	NG	NN	NN
NI/NI	NN	NN	NI	NI	NI	NN	NN	NI	NI	NI	NN	NN	NI	NI	NG	NN	NN
NG/NG	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN
HD/G*	NN	NN	NI	NI	HD	NN	NN	NI	NI	G*	NN	NN	NI	NI	NG	NN	NN
G*/HD	NN	NN	NI	NI	G*	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN

TALE_1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
HD	NN	NN	NI	NI	NG	HD	NI	NI	HD	HD	HD	NN	NI	NN	NG	NI
G*	NN	NN	NI	NI	NG	HD	NI	NI	HD	HD	G*	NN	NI	NN	NG	NI

TALE_1b

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
HD	NN	NN	NI	NI	NG	HD	NI	NI	HD	NI	HD	NN	NI	NN	NG	NN	NN
G*	NN	NN	NI	NI	NG	HD	NI	NI	HD	NI	G*	NN	NI	NN	NG	NN	NN

TALE_0

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
HD	NN	NI	NG	NG	HD	HD	NI	NG	NG	HD	HD	NI	NG	NG	HD	HD	NI	NG	NG

TALE_1c (targeting HEY2 gene)

	1	2	3	4	5	6	7	8	9	10	11	12
HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN

7. Supplementary Information

TALE_1c (for DNaseI Footprinting Assay, assembled in pAni521)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	LR
HD	HD	NG	NG	HD	HD	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NN*	HD	NG	NG	HD	NN*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NQ*	HD	NG	NG	HD	NQ*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NH*	HD	NG	NG	HD	NH*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
ND*	HD	NG	NG	HD	ND*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NE*	HD	NG	NG	HD	NE*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NS*	HD	NG	NG	HD	NS*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NT*	HD	NG	NG	HD	NT*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NY*	HD	NG	NG	HD	NY*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NK*	HD	NG	NG	HD	NK*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NR*	HD	NG	NG	HD	NR*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NW*	HD	NG	NG	HD	NW*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG

TALE_1c (for Luciferase Assay, assembled in pcDNA3.1-GoldenGate-VP64)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	LR
G*	HD	NG	NG	HD	G*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NG	HD	NG	NG	HD	NG	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NH*	HD	NG	NG	HD	NH*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
ND*	HD	NG	NG	HD	ND*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NY*	HD	NG	NG	HD	NY*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG
NW*	HD	NG	NG	HD	NW*	NN	NG	NG	NG	HD	HD	NI	HD	NI	NG	HD	NG

TALE_2 (for imaging analysis, assembled in pAni521 or pÁlM1577)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	LR
HD/HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN
G*/G*	NN	NN	NI	NI	G*	NN	NN	NI	NI	G*	NN	NN	NI	NI	NG	NN	NN
NY*/NY*	NN	NN	NI	NI	NY*	NN	NN	NI	NI	NY*	NN	NN	NI	NI	NG	NN	NN
NH*/NH*	NN	NN	NI	NI	NH*	NN	NN	NI	NI	NH*	NN	NN	NI	NI	NG	NN	NN
HD/NG	NN	NN	NI	NI	HD	NN	NN	NI	NI	NG	NN	NN	NI	NI	NG	NN	NN

TALE_0 (For site-directed methylation, assembled in pÁlM1285 and pÁlM1560)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	LR
HD	NN	NI	NG	NG	HD	HD	NI	NG	NG	HD	HD	NI	NG	NG	HD	HD	NI	NG	NG

7.2 Supplementary Figures

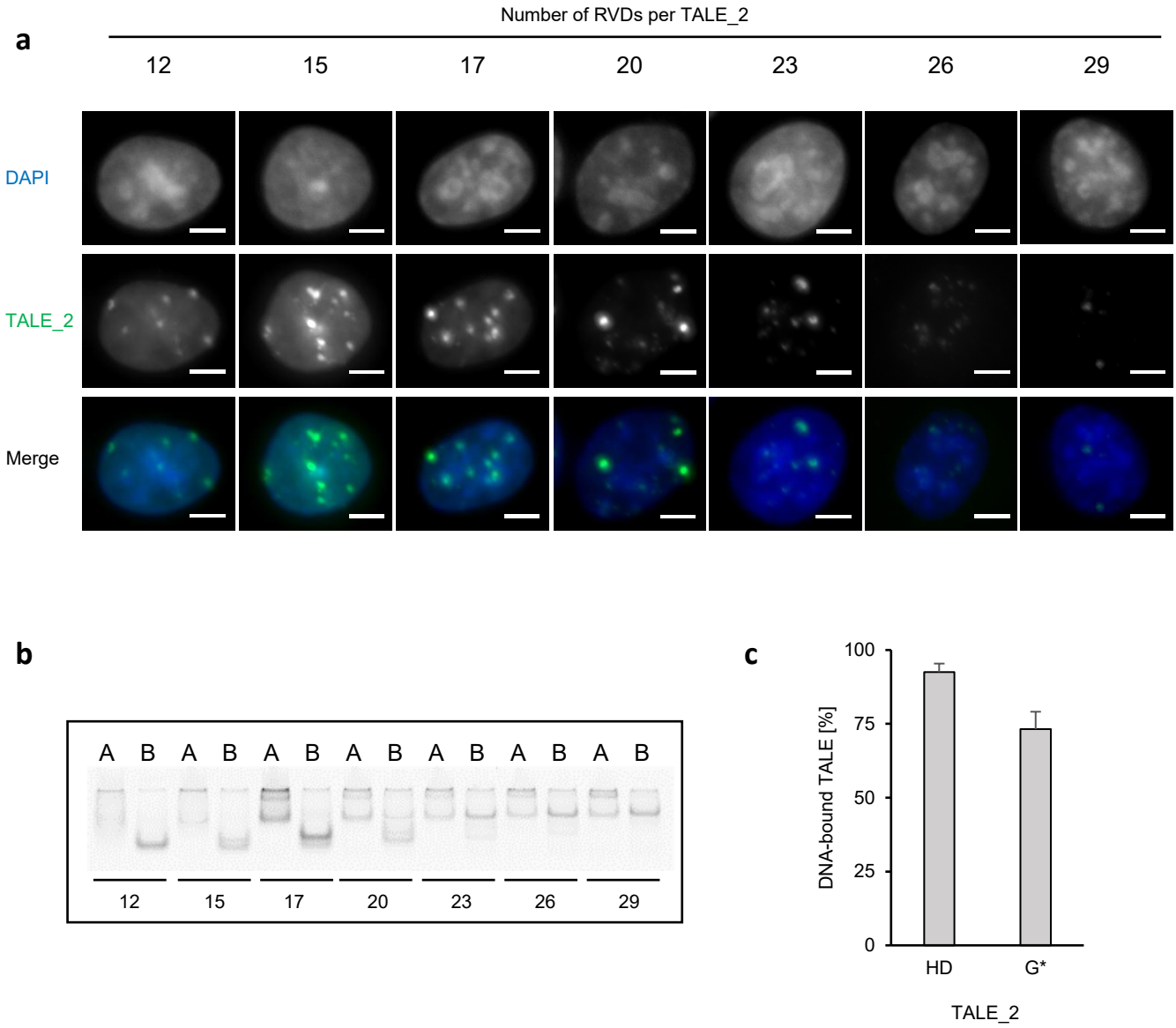
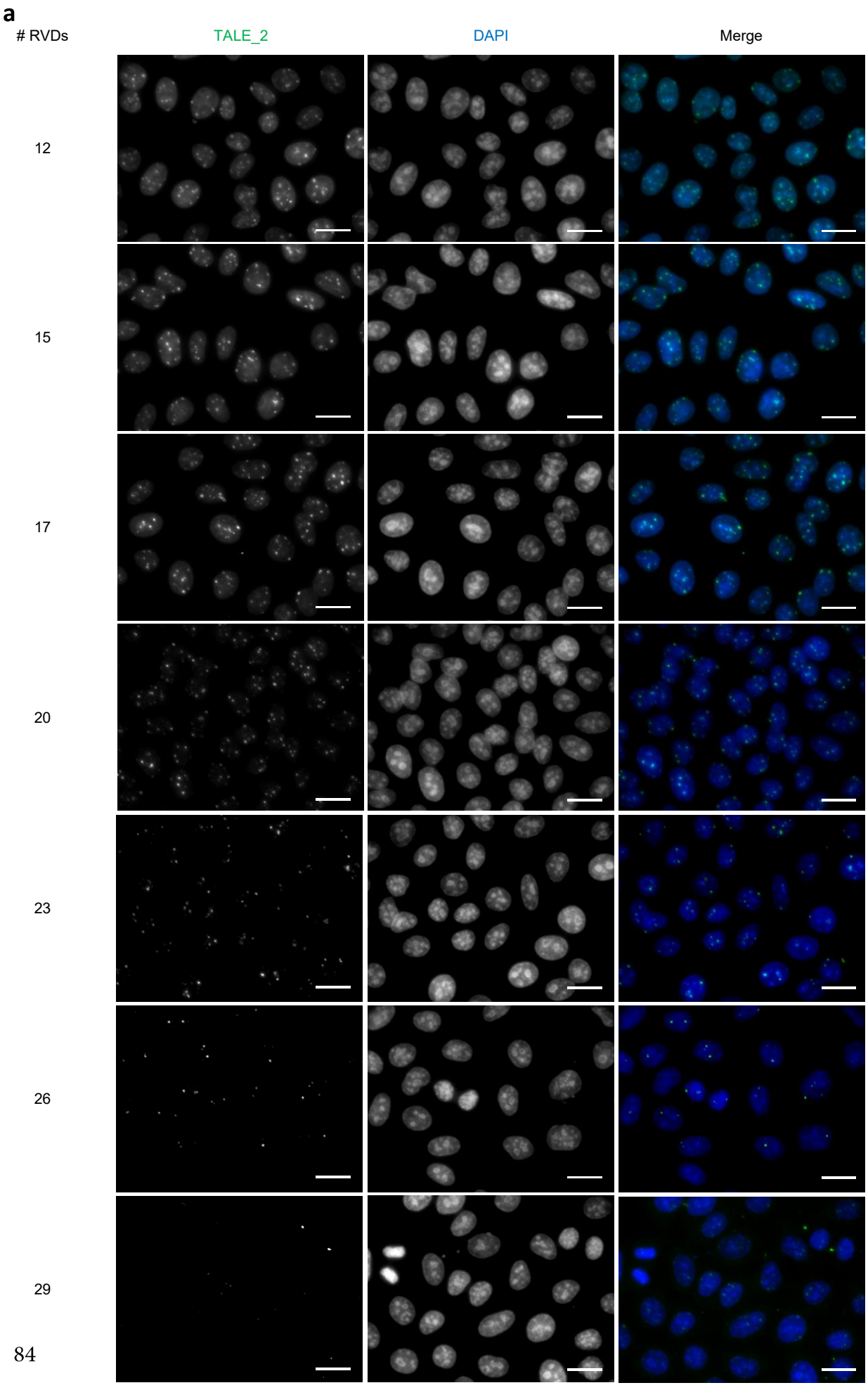


Figure S1. TALE Length Optimization. a) TALE stainings for length optimization using different truncations (29 – 12 RVDs, see TALEs Assembly table) of TALE_2 (2 CpGs) fused to EGFP in HeLa cells. Scale bar is 5µm. **b)** Electromobility Shift Assay of the versions of TALE_2. A indicates only TALE, B indicates that the TALE of the given length was used in combination with DNA. **c)** Competitive EMSA with equimolar concentrations of TALE_2-EGFP bearing HD and TALE_2-mCherry bearing G* at positions 5 and 10.

7. Supplementary Information



◄ **Figure S2. TALE Length optimization. a)** TALE staining of different truncations of TALE_2 (12 – 29 RVDs) fused to EGFP in HeLa cells. Scale bar is 20 µm.

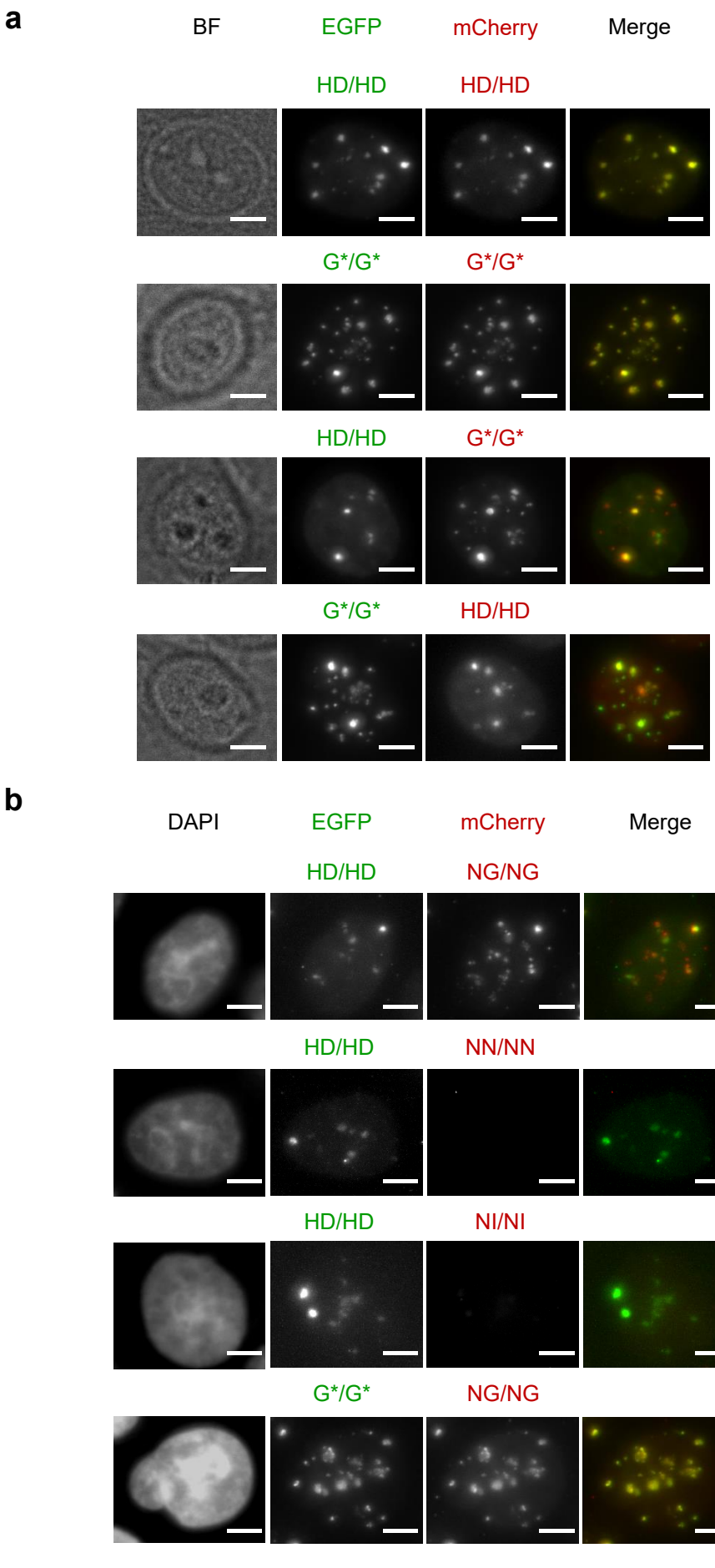


Figure S3. TALEs selectivity. a) HeLa co-stains with TALE_2 bearing indicated repeats at positions 5 and 10. **b)** Co-stains as in Fig. S3 with indicated repeats. Scale bar is 5 µm.

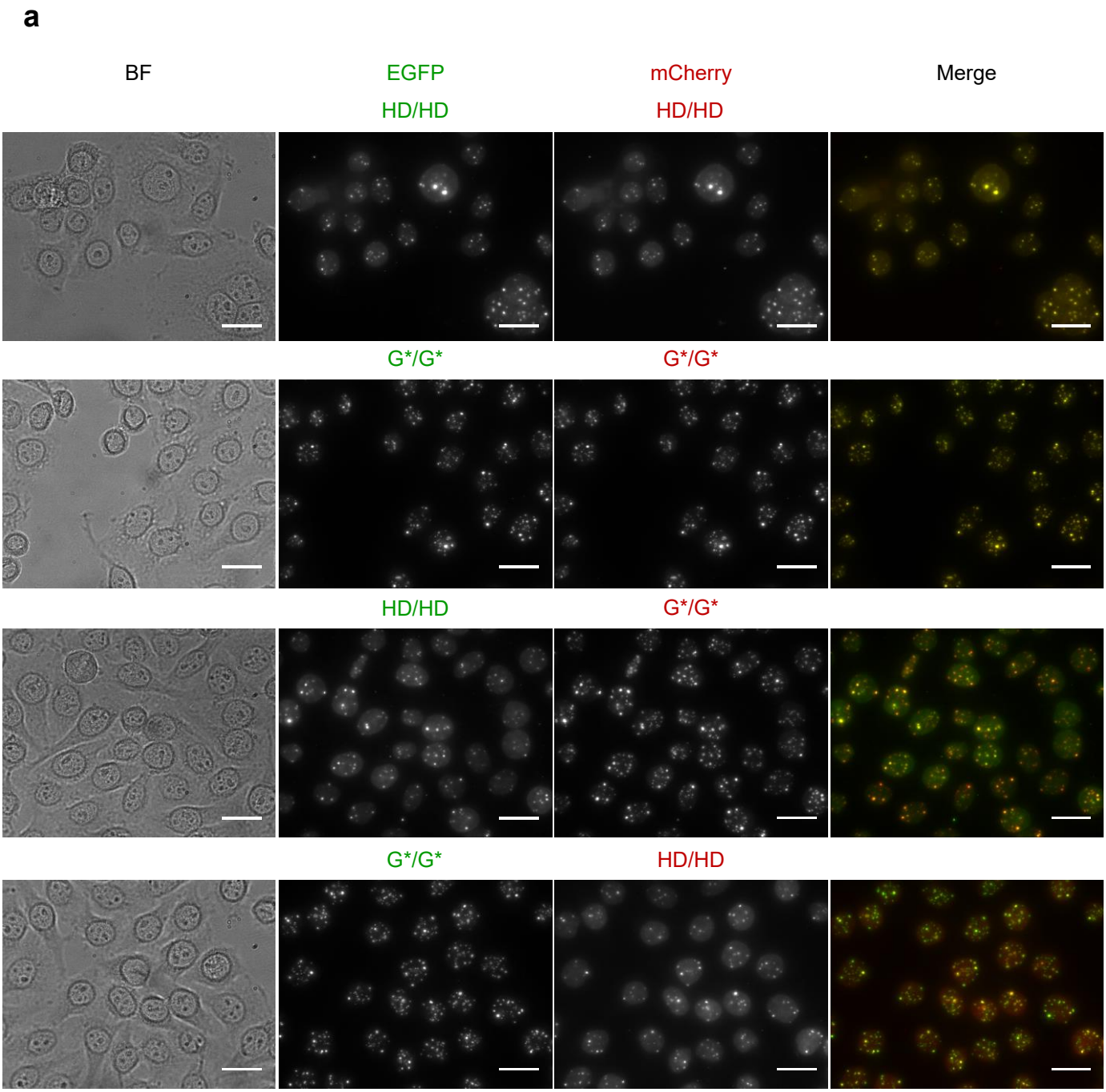


Figure S4. TALEs selectivity. a) HeLa co-stainings with TALE_2 bearing indicated repeats at positions 5 and 10. Scale bar is 20 μm.

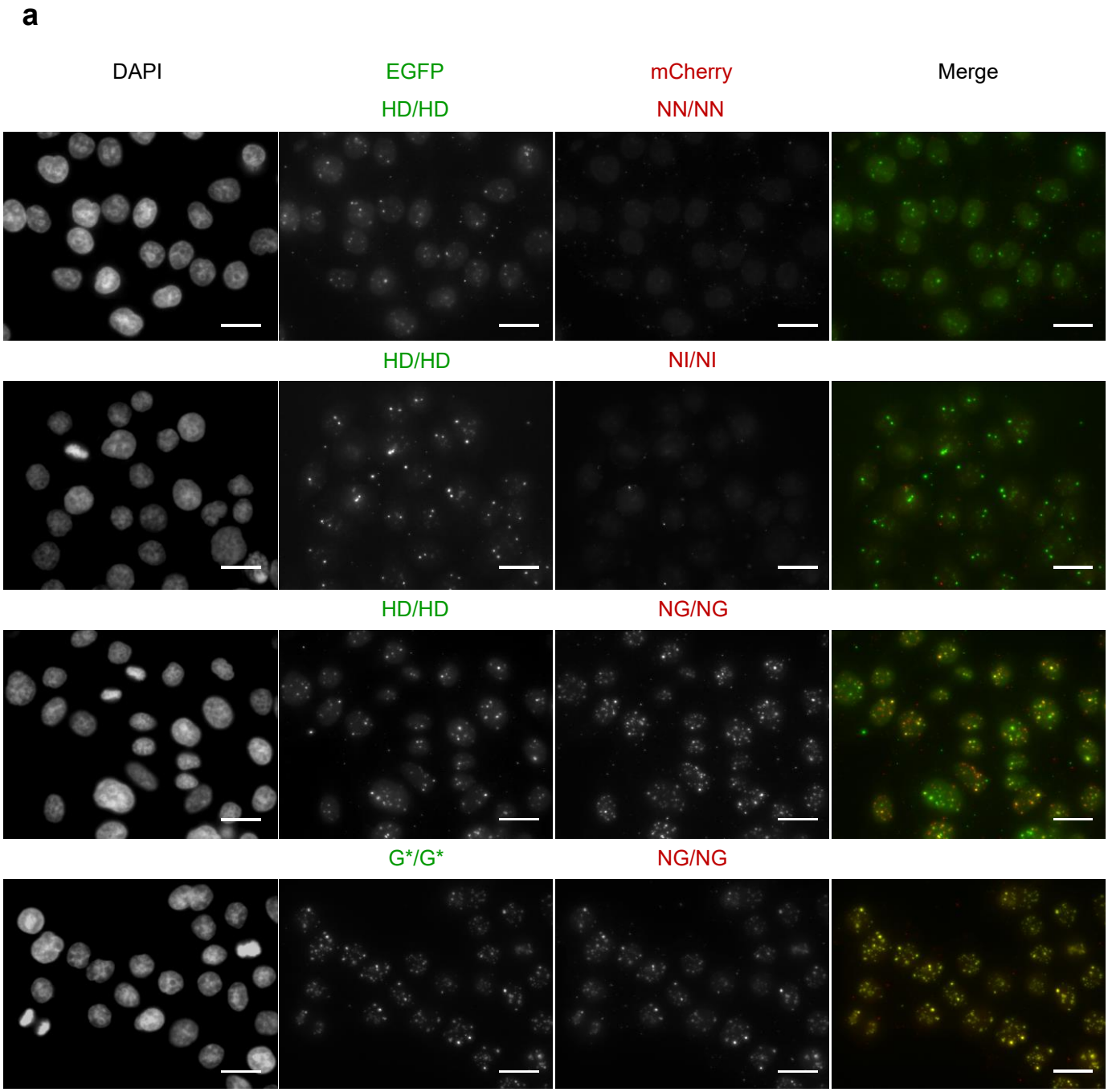
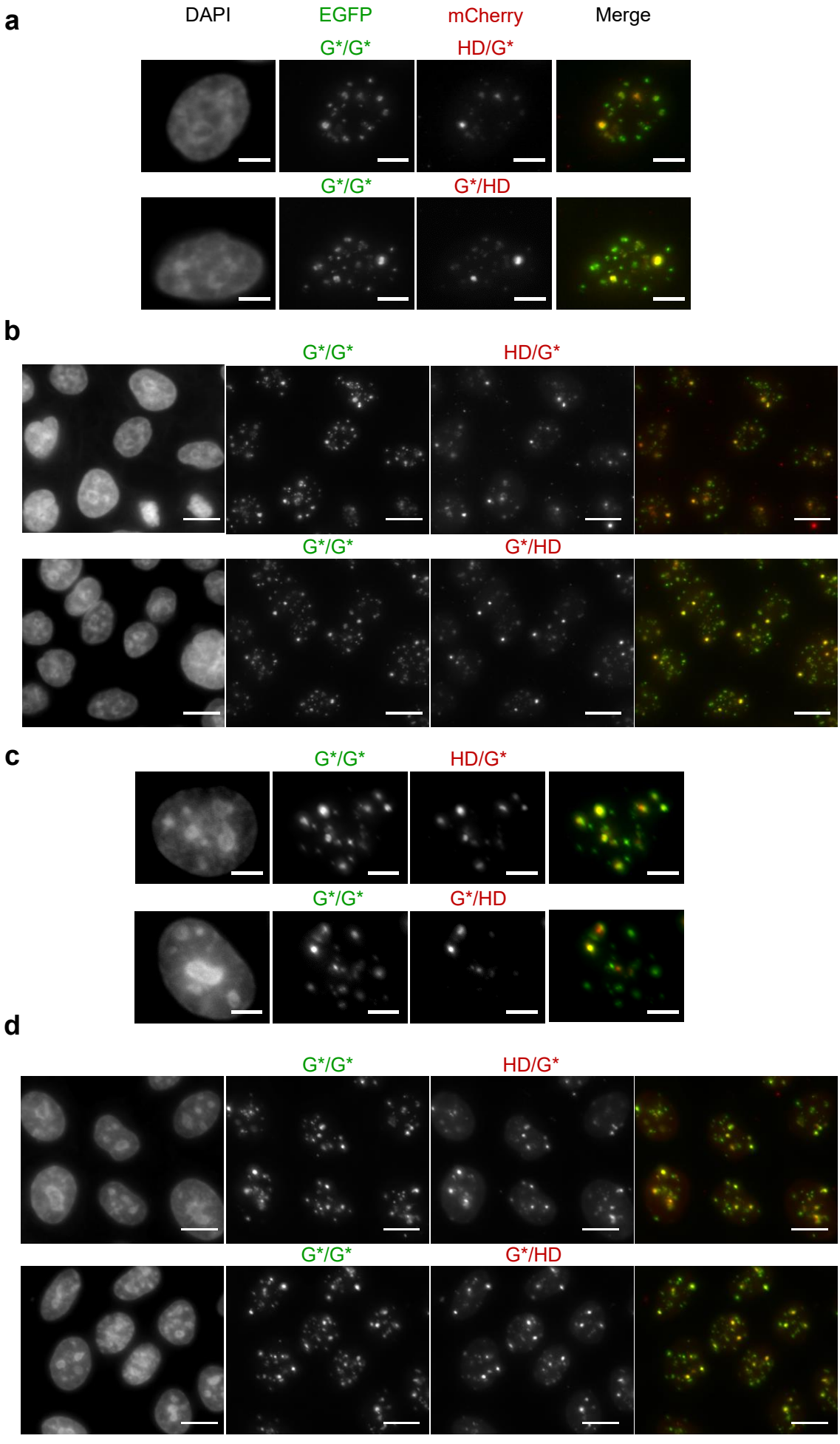
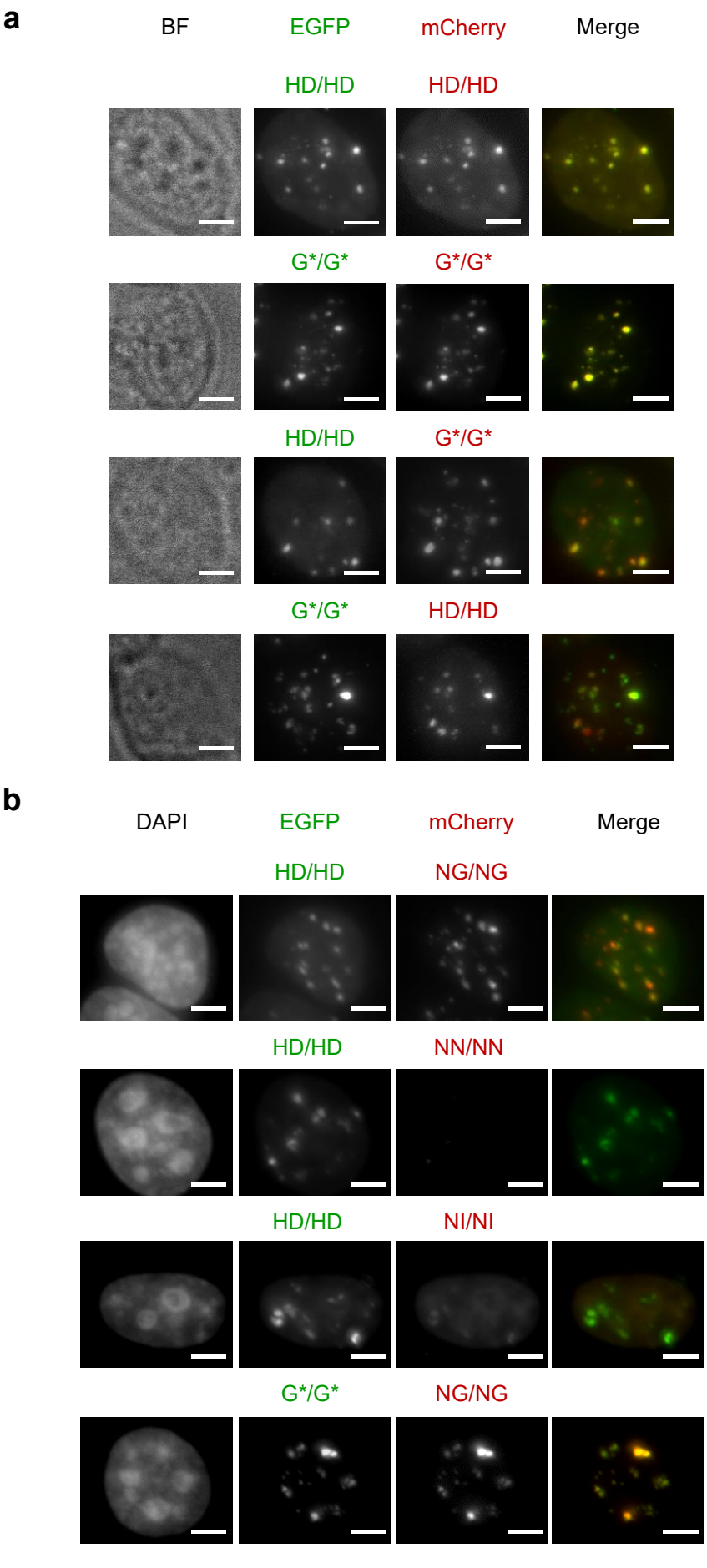


Figure S5. TALEs selectivity. a) HeLa co-stainings with TALE_2 bearing indicated repeats at positions 5 and 10. Scale bar is 20 µm.


a
DAPI
EGFP
mCherry
Merge
G*/G*
HD/G*
G*/G*
G*/HD
b
G*/G*
HD/G*
G*/G*
G*/HD
c
G*/G*
HD/G*
G*/G*
G*/HD
d
G*/G*
HD/G*
G*/G*
G*/HD

◄ **Figure S6. TALEs selectivity in HeLa and HEK 293T cells. a)** TALE stains with TALE_2 bearing indicated RVDs at positions 5 and 10 in HeLa cells. Scale bar is 5 μm **b)** TALE stain as in Fig. S6a showing multiple cells. Scale bar is 20 μm. **c)** TALE stains of HEK 293T cells with TALE_2 bearing indicated RVDs at positions 5 and 10. Scale bar is 5 μm. **d)** TALE stains as in Fig. S6c showing multiple cells. Scale bar is 20 μm.


a
BF
EGFP
mCherry
Merge
HD/HD
HD/HD
G*/G*
G*/G*
HD/HD
G*/G*
G*/G*
HD/HD
b
DAPI
EGFP
mCherry
Merge
HD/HD
NG/NG
HD/HD
NN/NN
HD/HD
NI/NI
G*/G*
NG/NG

◄ **Figure S7. TALE selectivity in HEK 293T cells. a)** TALE stains with TALE_2 bearing indicated RVDs at positions 5 and 10 in HEK 293T cells. Scale bar is 5 μm. **b)** Stains as in Fig. S7a with TALE_2 carrying the indicated RVDs. Scale bar is 5 μm.

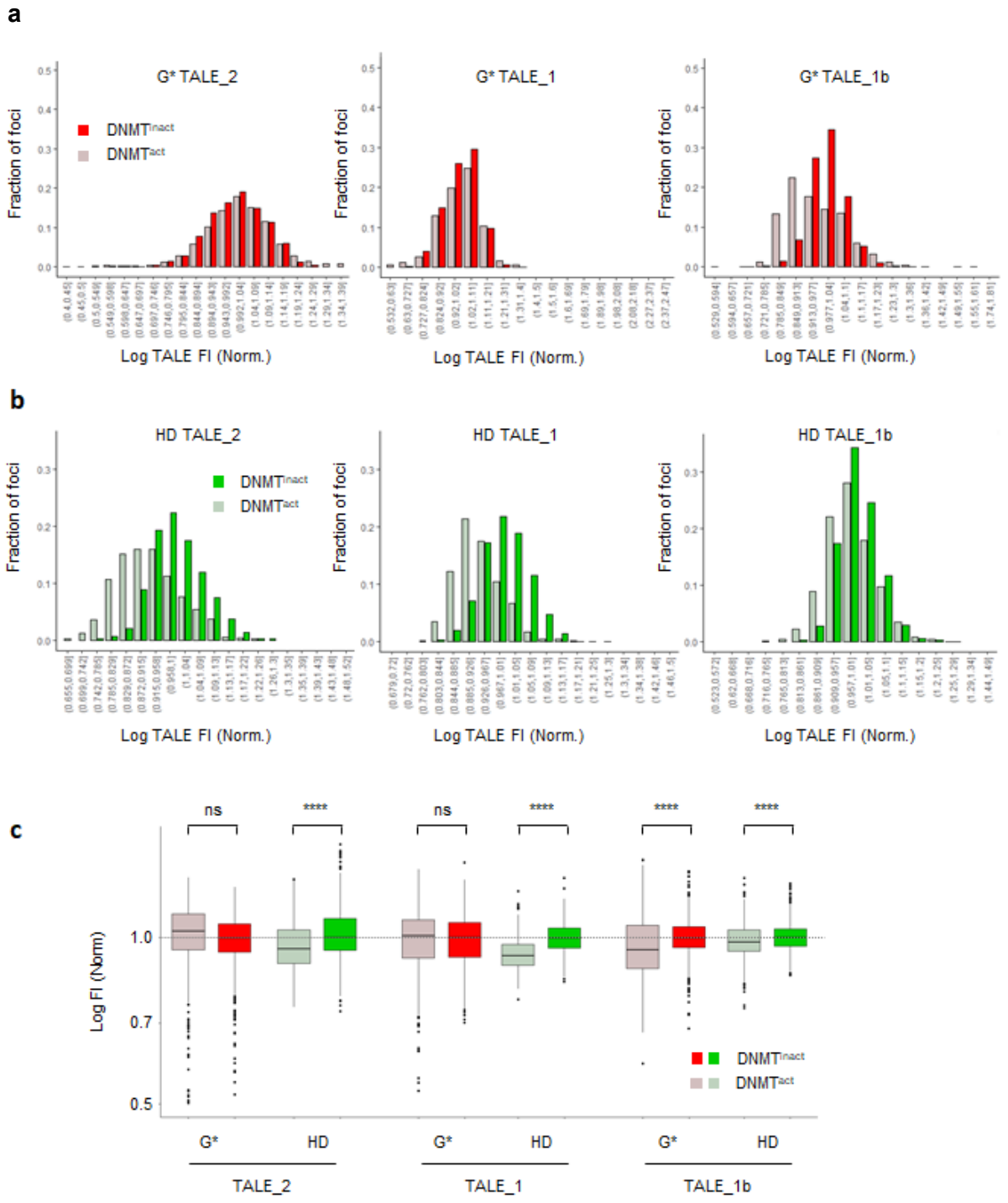


Figure S8. TALE staining and imaging of HEK 293T cells enables 5mC analysis at user-defined CpGs. Foci Analysis. a) Full histogram of G* TALE fluorescent intensities of foci (log transformed and normalized by the average fluorescent intensities of foci in $DNMT^{inact}$ sample for each experiment) co-stained with G* and HD TALEs in cells transfected with either $DNMT^{act}$ or $DNMT^{inact}$. **b)** Same as in Fig. S8a, but for HD TALEs fluorescent intensities. **c)** Box plots of data presented in Fig. S8a-b with t-test using as sample size the number of foci (N >1000 foci). See Data Analysis and Statistics section.

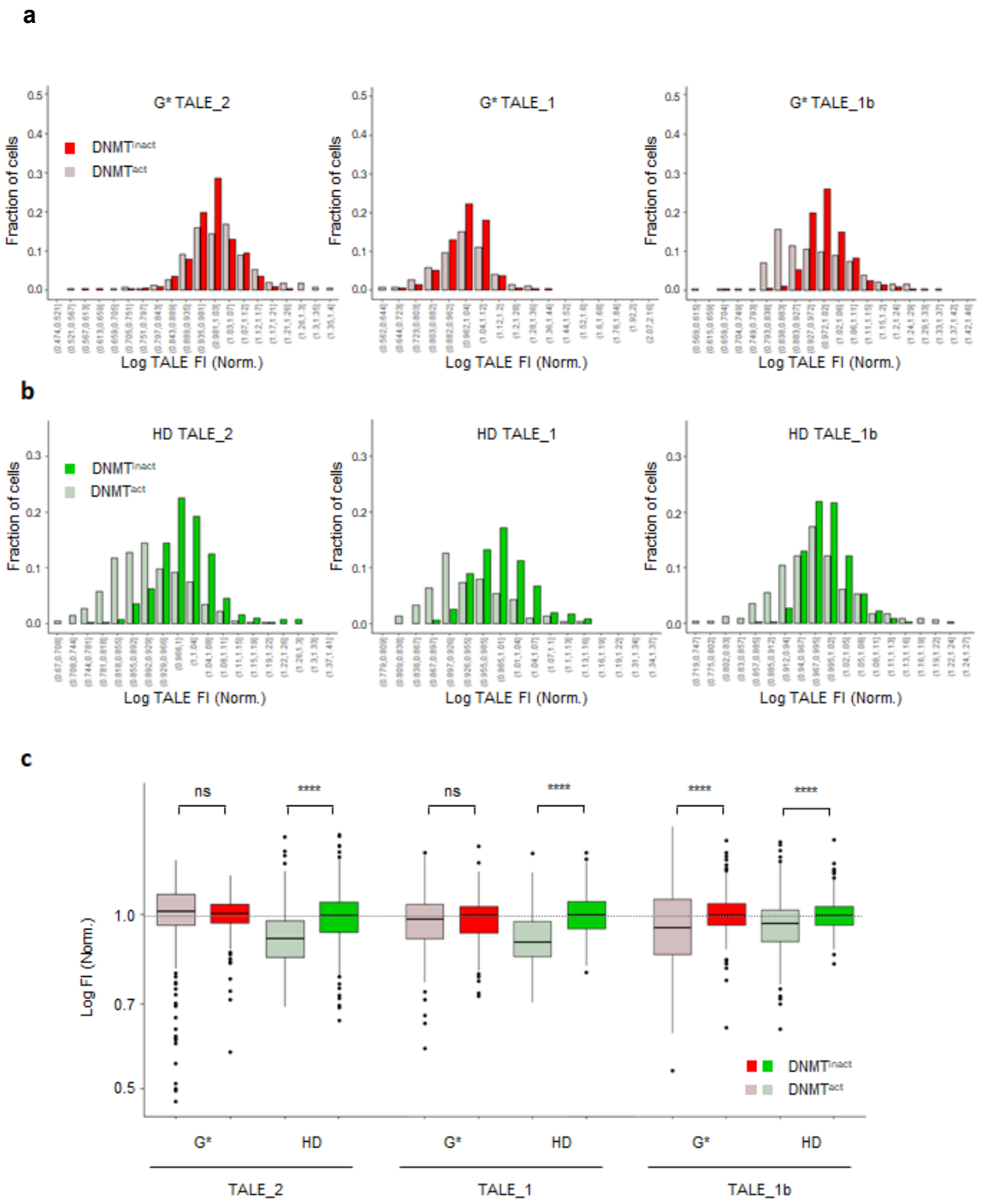
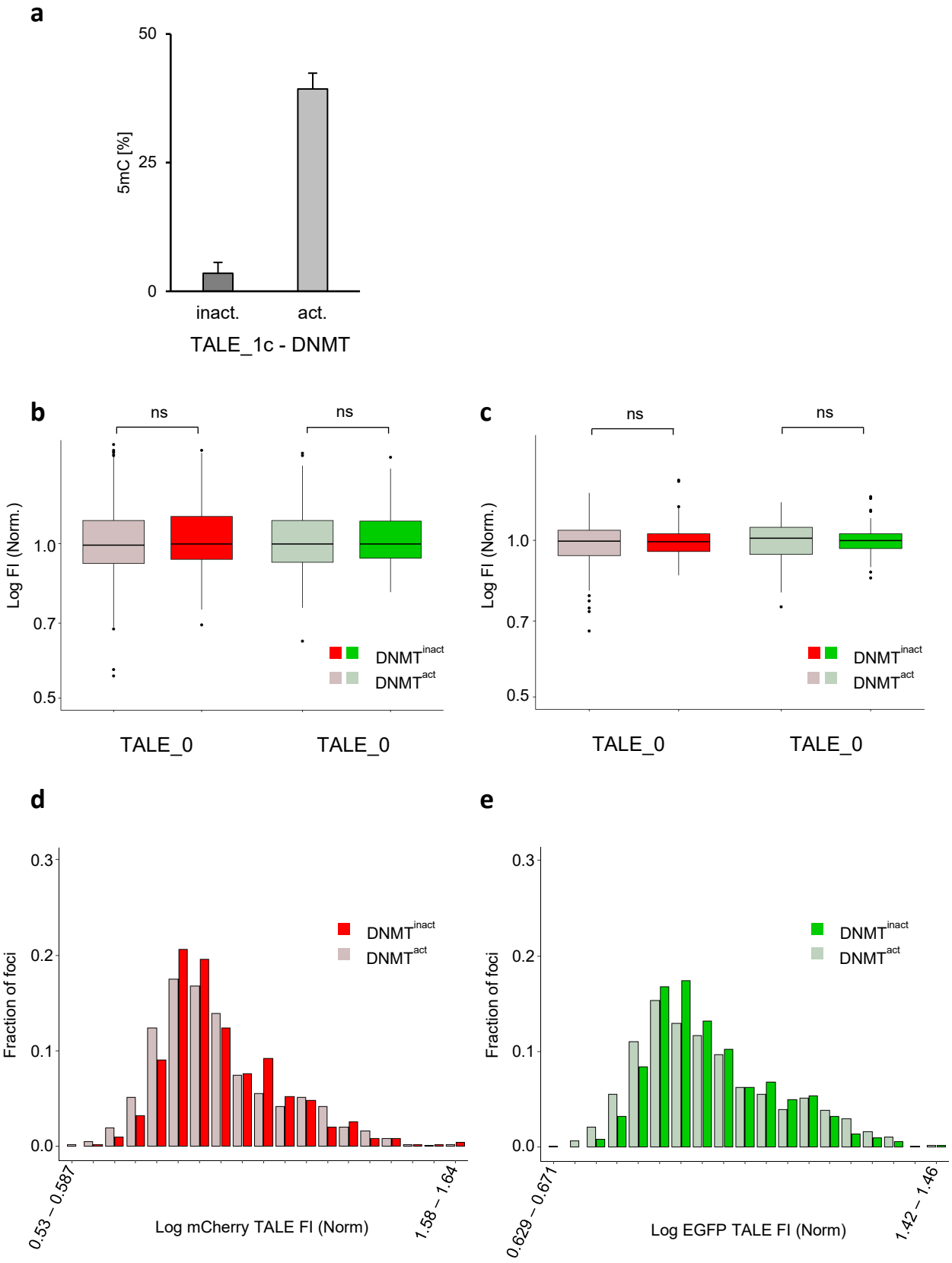


Figure S9. TALE staining and imaging of HEK 293T cells enables 5mC analysis at user-defined CpGs. Cell Analysis. a) Full histogram of G* TALE average fluorescent intensities of cells (log transformed and normalized by the average fluorescent intensities of foci within the same nucleus in DNMTinact sample for each experiment) co-stained with G* and HD TALEs in cells transfected with either DNMTact or DNMTinact. **b)** Same as in Fig. S8a, but for HD TALEs fluorescent intensities. **c)** Box plots of data presented in Fig. S8a-b with t-test using as sample size the number of cells (N > 200 cells). See Data Analysis and Statistics section.


a
50
25
0
5mC [%]
inact.
act.
TALE_1c - DNMT
b
ns
ns
Log FI (Norm.)
1.0
0.7
0.5
DNMT^inact
DNMT^act
TALE_0
TALE_0
c
ns
ns
Log FI (Norm.)
1.0
0.7
0.5
DNMT^inact
DNMT^act
TALE_0
TALE_0
d
Fraction of foci
0.3
0.2
0.1
0.0
DNMT^inact
DNMT^act
0.53 – 0.587
Log mCherry TALE FI (Norm)
1.58 – 1.64
e
Fraction of foci
0.3
0.2
0.1
0.0
DNMT^inact
DNMT^act
0.629 – 0.671
Log EGFP TALE FI (Norm)
1.42 – 1.46

◄ **Figure S10. TALEs targeting CpG-free target sequence are not affected by 5mC. a)** Pyrosequencing analysis of SATIII target CpG methylation for TALE_1c-DNMT^{act} or TALE_1c-DNMT^{inact} in HEK 293T cells. **b)** Log transformed fluorescent intensities of foci for mCherry (red) and EGFP TALE_0 (green) co-stains in DNMTact or DNMTinact (TALE_1c) transfected HEK 293T cells. Statistical analysis is a t-test using foci as sample size (N > 1000 foci). See Data Analysis and Statistics section. **c)** Boxplots as in Fig. S10b, but showing average fluorescent intensity of foci within each nucleus (log transformed and normalized), t-test was performed using number of cells as sample size (N > 200). See Data Analysis and Statistics section. **d)** Full histogram of mCherry TALE_0 fluorescent intensity of foci. **e)** Same as in Fig. S10e, but with EGFP TALE_0.

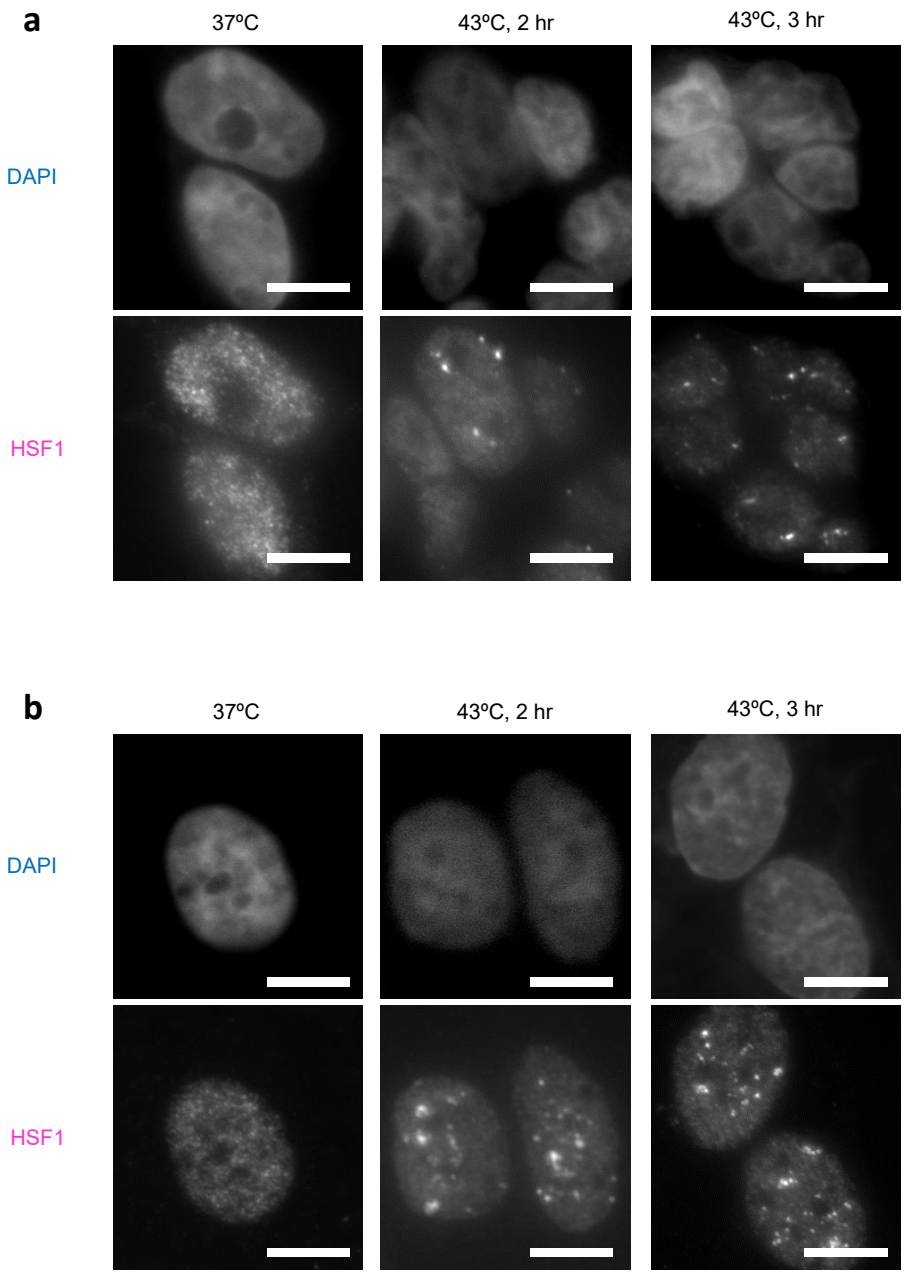
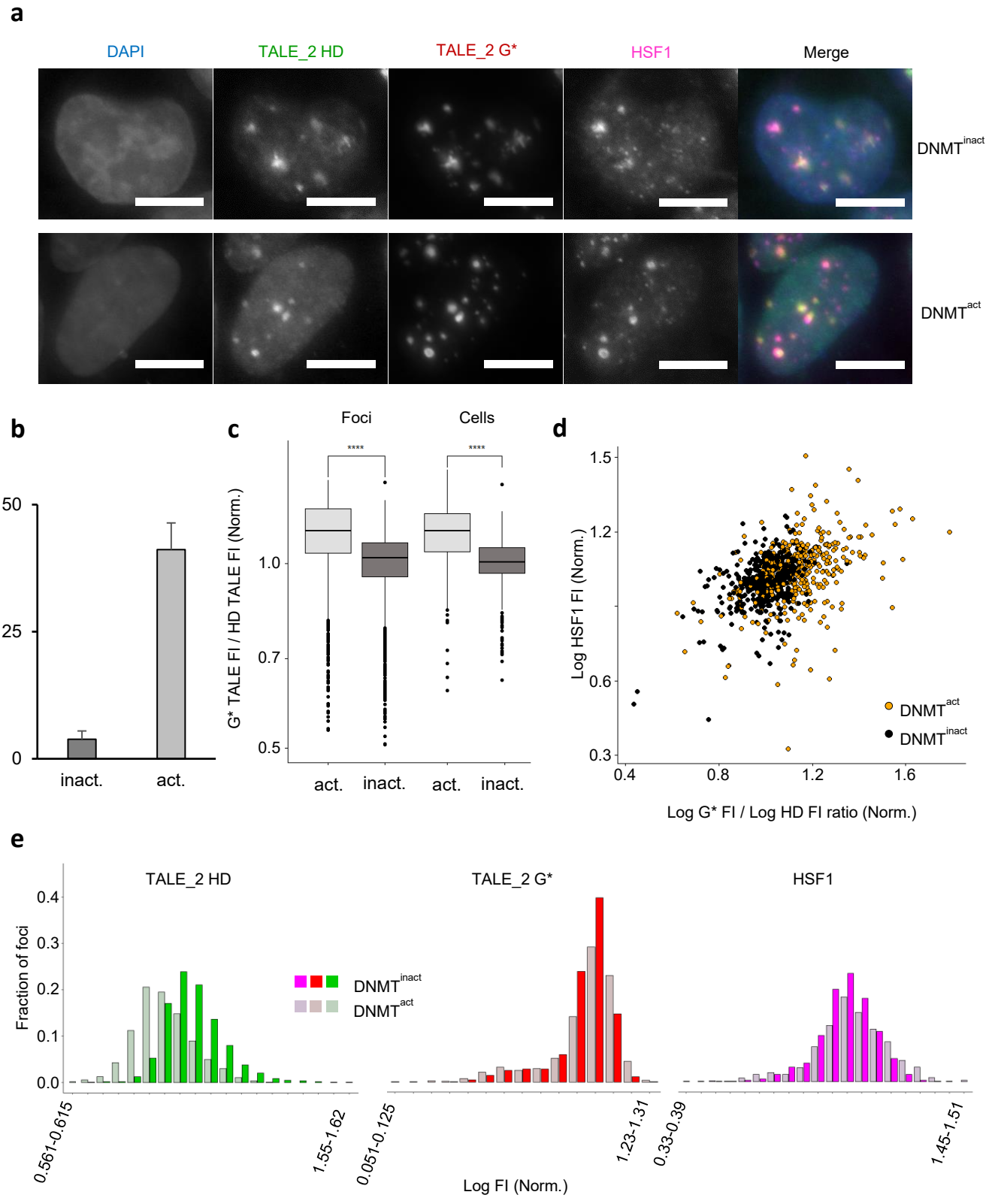


Figure S11. Nuclear stress body formation after Heat-Shock in HEK 293T and U2OS cells. a) Endogenous HSF1 immunostaining of HEK 293T cells grown at 37°C or incubated at 43°C for indicated times. Scale bar is 10 μm. **b)** HSF1 immunostaining as in Fig. S11a, but in U2OS cells. Scale bar is 10 μm.



◂ **Figure S12. TALEs reveal a role of 5mC in the regulation of heat-shock-induced recruitment of HSF1 to SATIII U2OS cells. a)** Co-stain of HD and G* TALE_2 combined with endogenous HSF1 immunostaining in U2OS cells. Cells were incubated for 2 hours at 43°C. Scale bar is 10 µm. **b)** Pyrosequencing analysis of SATIII target CpG methylation for DNMT^{act} or DNMT^{inact} U2OS cells. **c)** Boxplot showing the ratio of G* TALE

fluorescent intensity over HD TALE fluorescent intensity per foci and per cell(log transformed and normalized) in U2OS cells. T-test was perform using either the number of foci or the cells as sample size (N > 1000 foci, N > 200 cells). See Data Analysis and Statistics section. **d)** Scatter plot showing the ratio of G* fluorescent intensity over HD fluorescent intensity per cells versus fluorescent intensity of HSF1 (log transformed and normalized). **e)** Full histograms showing the fraction of foci with indicated fluorescent intensities for HD and G* TALEs and HSF1 immunostaining.

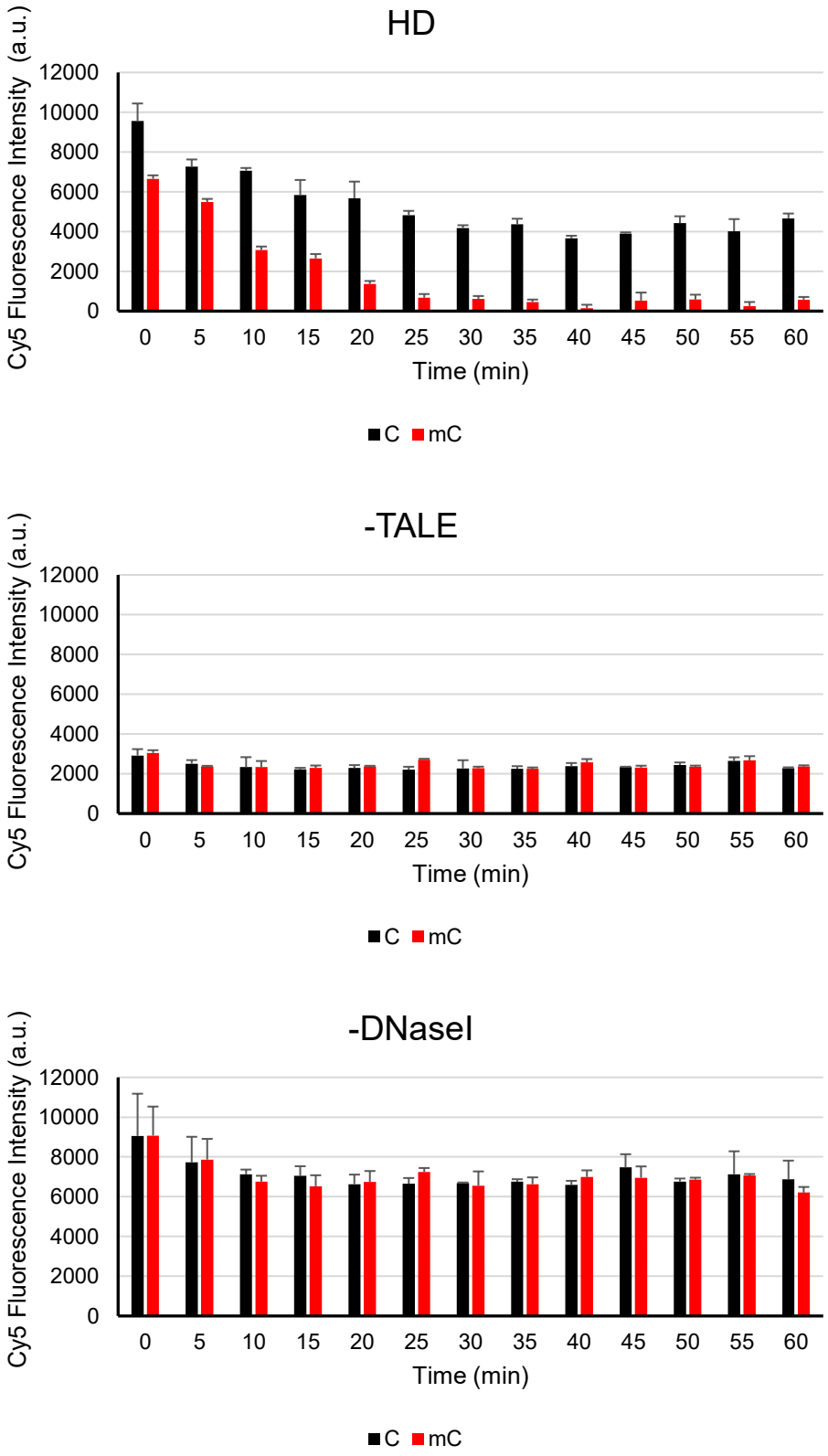


Figure S13. Fret Kinetics for TALE_1 HD and controls. Controls shown are performed under the same experimental conditions as TALE_1 HD either without TALE (middle) or without addition of DNaseI (bottom). Error bars show standard deviation.

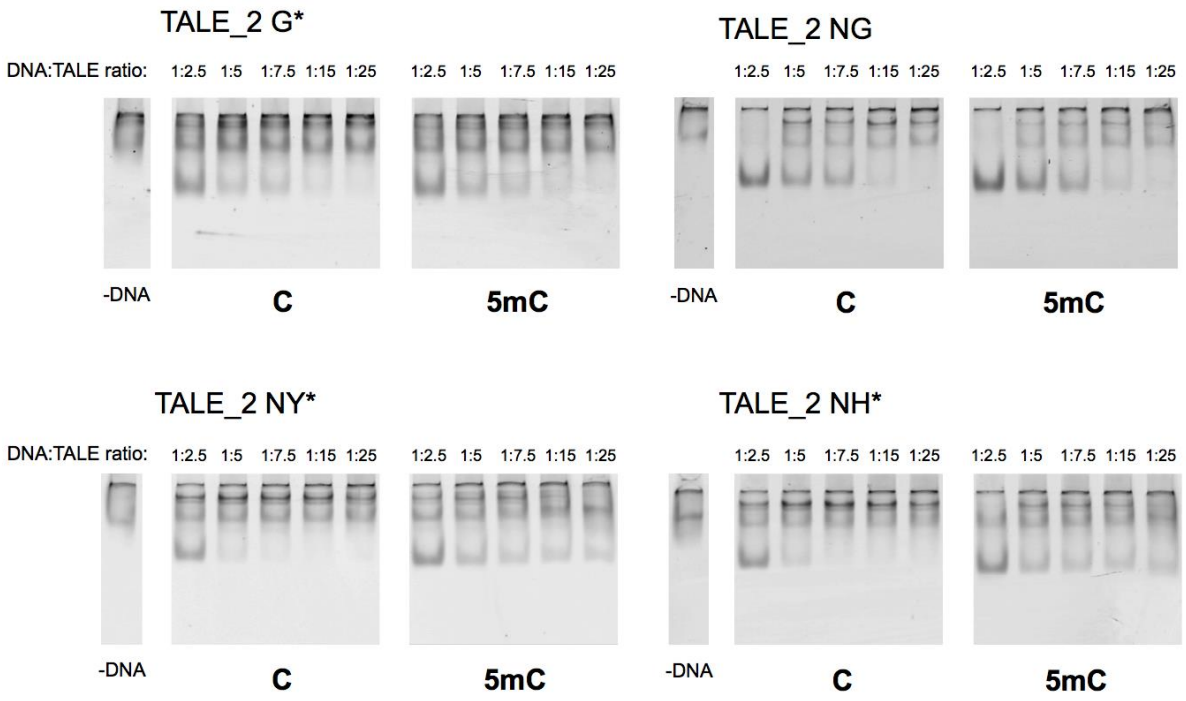


Figure S14. EMSA assays. EMSA with indicated, SATIII-targeting TALE versions bearing an N-terminal GFP and unlabeled DNA with C or 5mC opposite the target CpGs. Note that fluorescence of GFP is recorded.

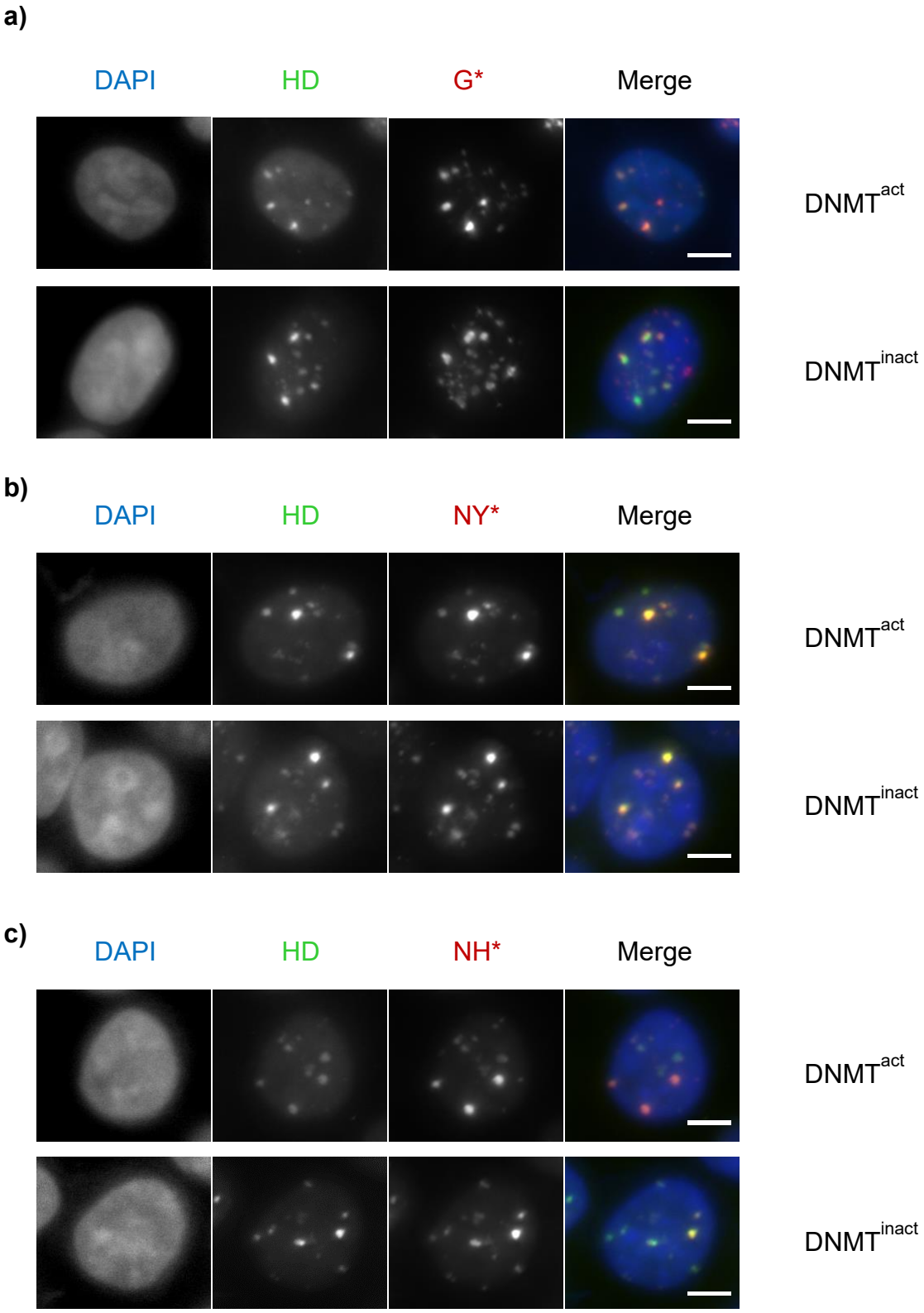


Figure S15. TALE staining. a) HEK 293T cells co-stained with TALE_2 bearing indicated repeats at positions 5 and 10 (HD in eGFP and G* in mCherry). **b)** Same as in a, but using NY* at positions 5 and 10 in the mCherry-TALE. **c)** Same as in a, but with RVD NH* for positions 5 and 10 of mCherry-TALE. Scale bar is 5 µm.

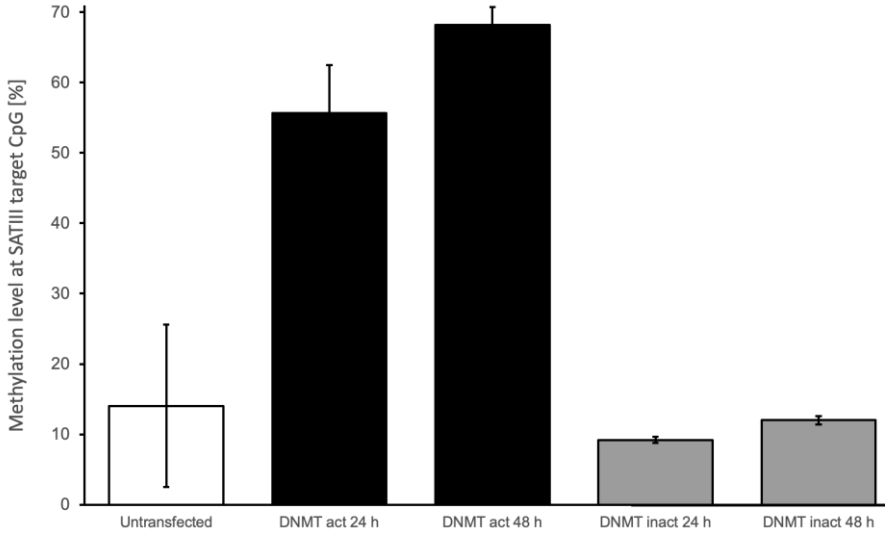
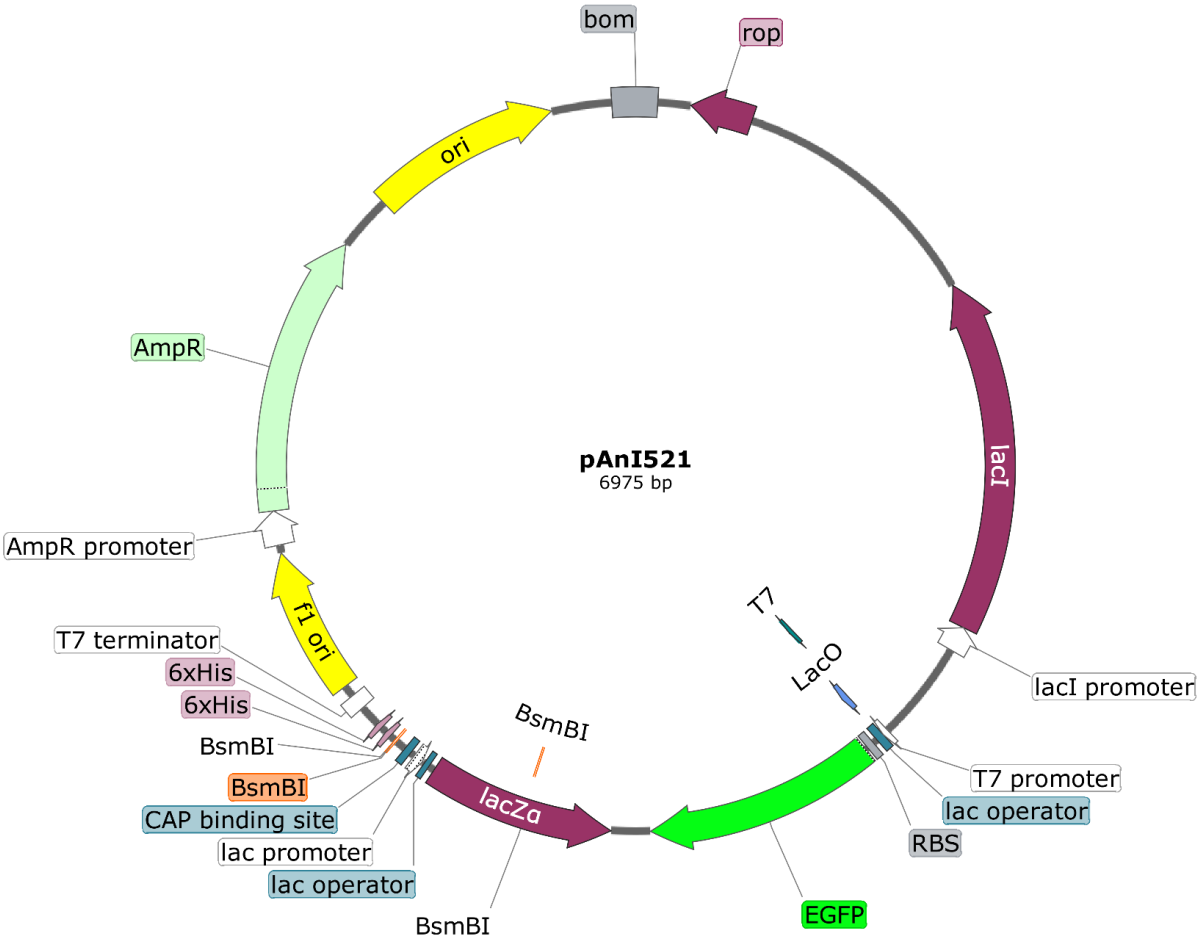


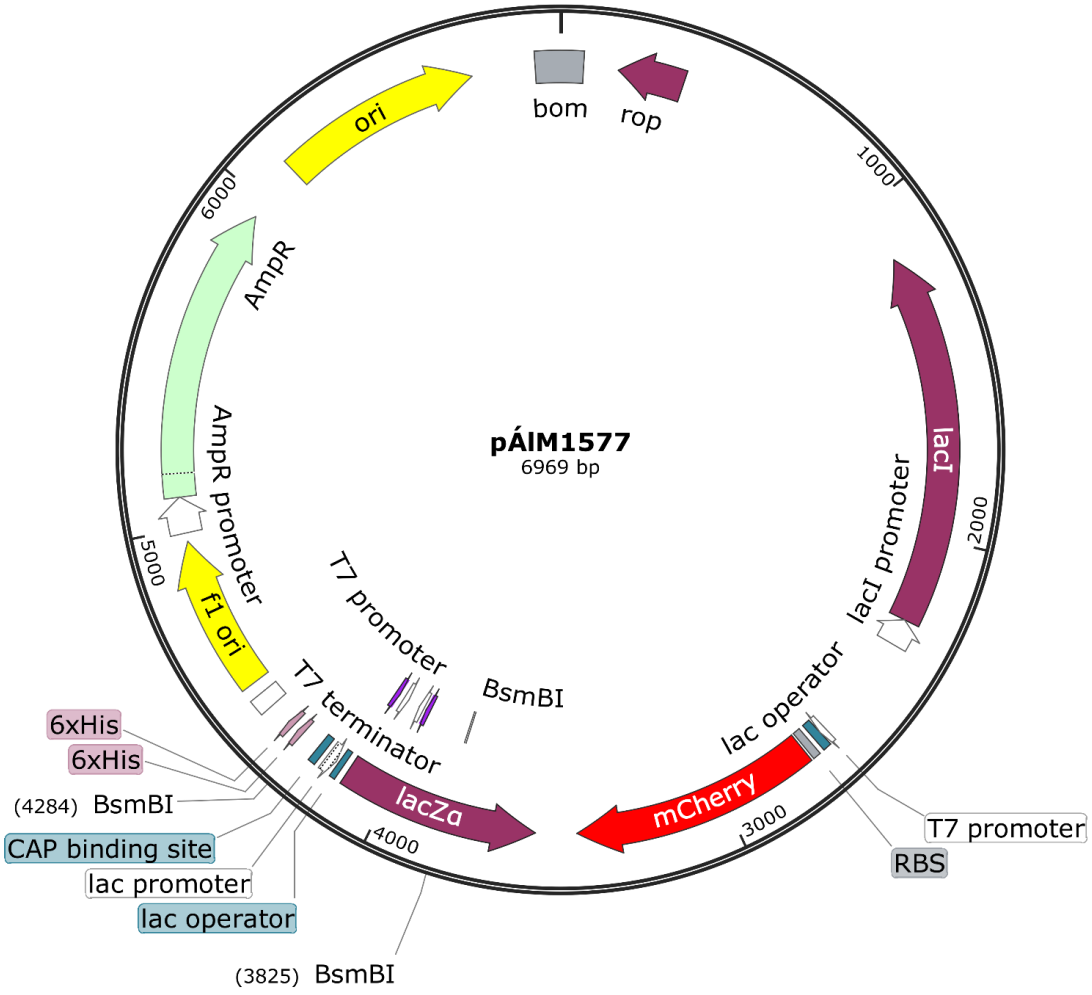
Figure S16. Methylation of SATIII target CpG. HEK293T cells were not transfected or transfected with the TALE_0-DNMT3a3L fusion protein "DNMT^{act}" or the catalytically inactive version "DNMT^{inact}". Shown are 5mC levels from bisulfite PCR and Pyrosequencing analysis of the SATIII target CpG at the indicated timepoints after transfection (Error bars from biological duplicates).

7.3 Vector Maps



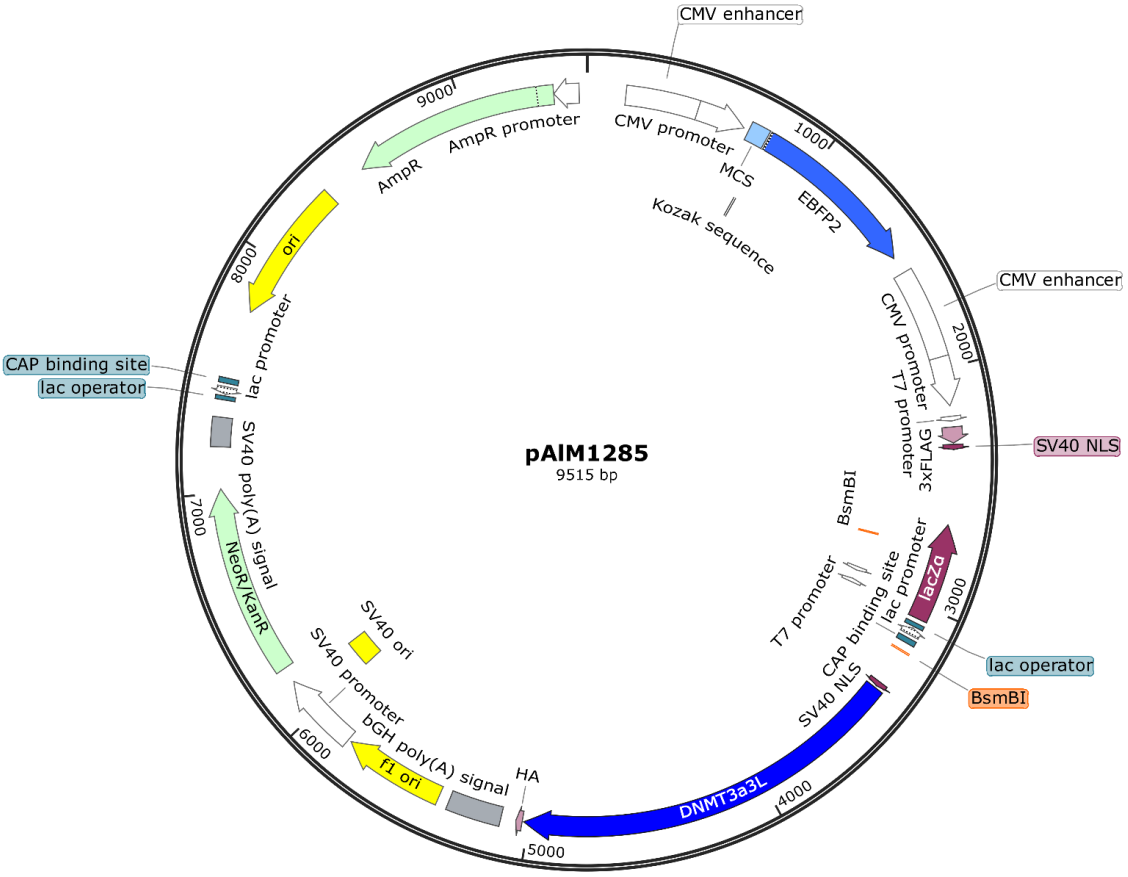
pAnI521

GG2 Entry Vector for bacterial expression of TALEs fused to N-terminal EGFP and C-terminal 6xHis for protein purification.



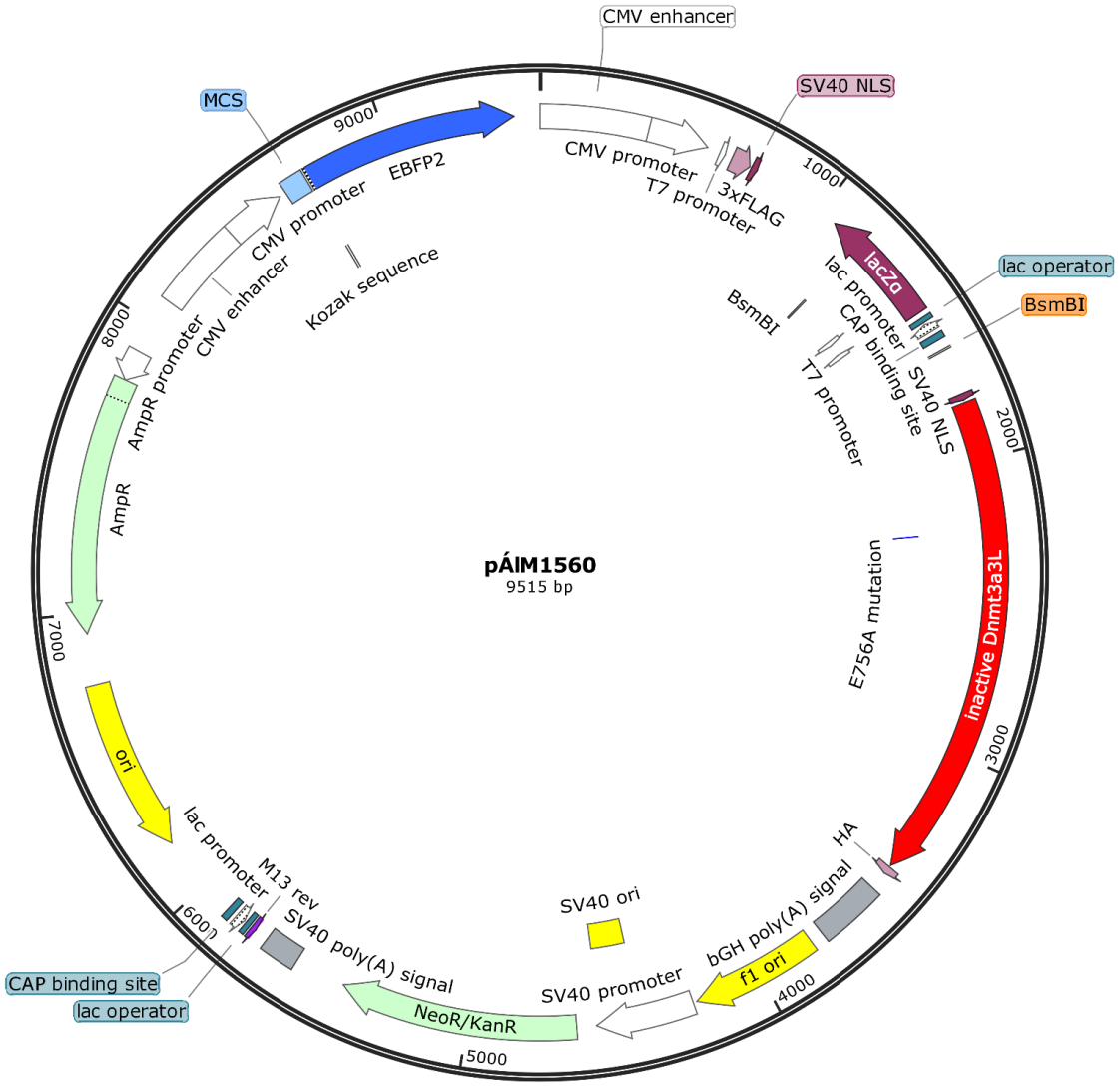
pÁlM1577

Vector for bacterial expression of TALEs fused to N-terminal mCherry and C-terminal 6xHis for protein purification.



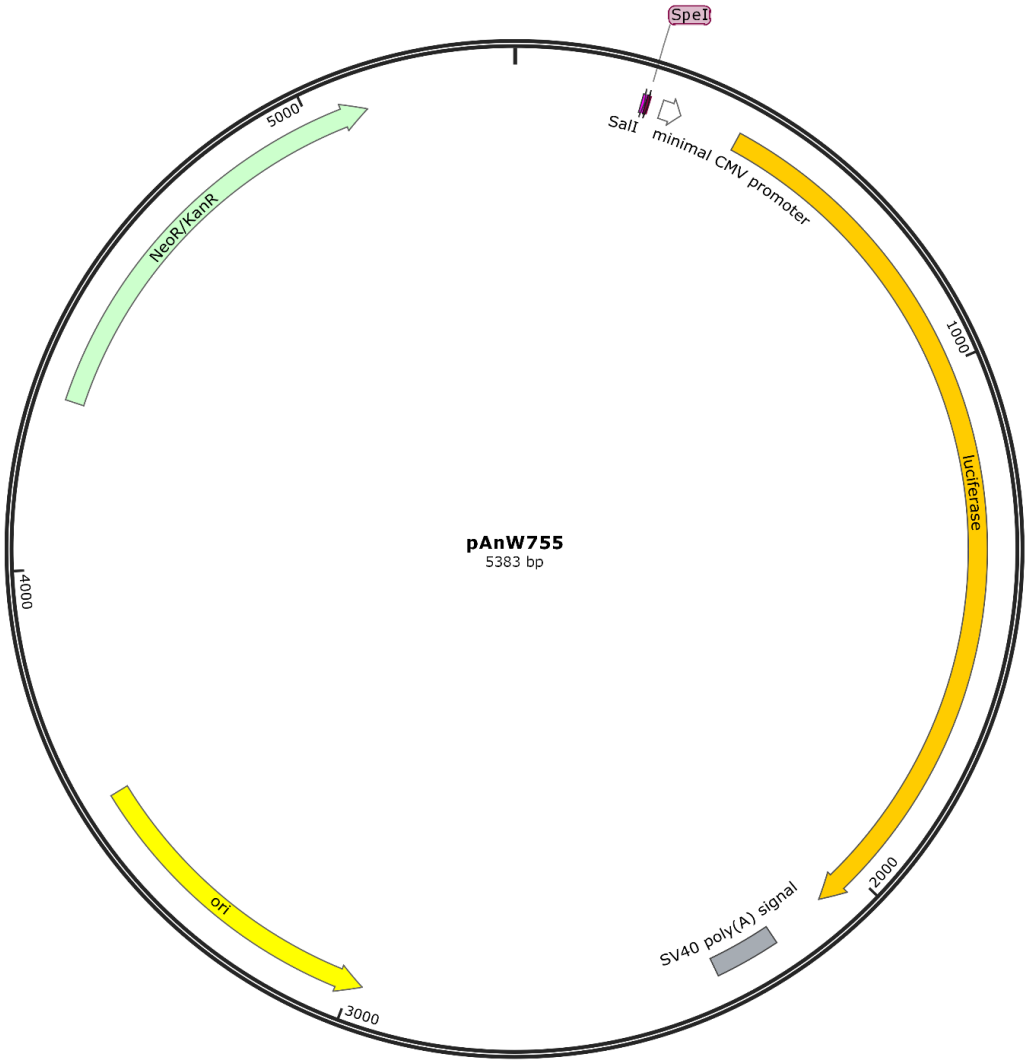
pÁlM1285

Vector for mammalian expression of TALEs fused to C-terminal DNMT3a3L and HA-tag and N-terminal 3XFLAG-tag for for live-cell site-specific DNA methylation. It also contains an EBFP2 fluorophore under the same promoter (CMV) as transfection control.



pÁIM1560

Vector for mammalian expression of TALEs fused to C-terminal inactive DNMT3a3L (carrying E756A mutation) and HA-tag and N-terminal 3XFLAG-tag. It also contains an EBFP2 fluorophore under the same promoter (CMV) as transfection control. This construct is used as negative control for site-specific DNA-methylation.



pAnW755

Plasmid map of luciferase reporter plasmid pAnW755. Shown are the restriction sites SalI (pink: SalI) and SpeI (purple: SpeI) for insertion of the TALE target sequence, a minimal CMV promoter (white: minimal CMV promoter) for luciferase expression (orange: luciferase) with a C-terminal SV40 poly(A) signal (grey: SV40 poly(A) signal), the origin for bacterial replication (yellow: ori) and the gene for neomycin/kanamycin resistance (bright green: NeoR/KanR).

8. References

1. B. Alberts, *Molecular biology of the cell, The problems book* (Garland Science, New York, 2015).
2. Á. Muñoz-López *et al.*, Designer Receptors for Nucleotide-Resolution Analysis of Genomic 5-Methylcytosine by Cellular Imaging. *Angewandte Chemie (International ed. in English)*. **59**, 8927–8931 (2020), doi:10.1002/anie.202001935.
3. D. W. Deamer, J. W. Szostak, *The origins of life, A subject collection from Cold Spring Harbor perspectives in biology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 2010).
4. J. D. Watson, F. H. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953. *Nature*. **248**, 765 (1974), doi:10.1038/248765a0.
5. E. CHARGAFF, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*. **6**, 201–209 (1950), doi:10.1007/BF02173653.
6. L. Pauling, R. B. Corey, A Proposed Structure For The Nucleic Acids. *Proceedings of the National Academy of Sciences of the United States of America*. **39**, 84–97 (1953), doi:10.1073/pnas.39.2.84.
7. G. Kubik, D. Summerer, TALEored Epigenetics: A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *Chembiochem : a European journal of chemical biology*. **17**, 975–980 (2016), doi:10.1002/cbic.201600072.
8. J. M. Berg, J. L. Tymoczko, G. J. Gatto jr., L. Stryer, *Stryer Biochemie* (Springer Berlin Heidelberg; Imprint: Springer Spektrum, Berlin, Heidelberg, ed. 8, 2018).
9. S. BRENNER, F. JACOB, M. MESELSON, An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*. **190**, 576–581 (1961), doi:10.1038/190576a0.
10. A. KORNBERG, Biologic synthesis of deoxyribonucleic acid. *Science (New York, N.Y.)*. **131**, 1503–1508 (1960), doi:10.1126/science.131.3412.1503.

11. I. R. LEHMAN, M. J. BESSMAN, E. S. SIMMS, A. KORNBERG, Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *The Journal of biological chemistry.* **233**, 163–170 (1958).
12. M. Méchali, Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews. Molecular cell biology.* **11**, 728–738 (2010), doi:10.1038/nrm2976.
13. A. Schmoldt, H. F. Benthe, G. Haberland, Digitoxin metabolism by rat liver microsomes. *Biochemical pharmacology.* **24**, 1639–1641 (1975).
14. T. Okazaki *et al.*, Structure and metabolism of the RNA primer in the discontinuous replication of prokaryotic DNA. *Cold Spring Harbor symposia on quantitative biology.* **43 Pt 1**, 203–219 (1979), doi:10.1101/sqb.1979.043.01.026.
15. T. Okazaki, Days weaving the lagging strand synthesis of DNA - A personal recollection of the discovery of Okazaki fragments and studies on discontinuous replication mechanism. *Proceedings of the Japan Academy. Series B, Physical and biological sciences.* **93**, 322–338 (2017), doi:10.2183/pjab.93.020.
16. R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, A. Sugino, Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proceedings of the National Academy of Sciences of the United States of America.* **59**, 598–605 (1968), doi:10.1073/pnas.59.2.598.
17. G. C. Fareed, C. C. Richardson, Enzymatic breakage and joining of deoxyribonucleic acid. II. The structural gene for polynucleotide ligase in bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America.* **58**, 665–672 (1967), doi:10.1073/pnas.58.2.665.
18. M. MESELSON, F. W. Stahl, THE REPLICATION OF DNA IN ESCHERICHIA COLI. *Proceedings of the National Academy of Sciences of the United States of America.* **44**, 671–682 (1958), doi:10.1073/pnas.44.7.671.
19. M. Grunberg-Manago, P. J. Ortiz, S. Ochoa, Enzymic synthesis of polynucleotides I. polynucleotide phosphorylase of Azotobacter vinelandii. *Biochimica et Biophysica Acta.* **20**, 269–285 (1956), doi:10.1016/0006-3002(56)90286-4.
20. P. Cramer, Organization and regulation of gene transcription. *Nature.* **573**, 45–54 (2019), doi:10.1038/s41586-019-1517-4.

21. V. Haberle, A. Stark, Eukaryotic core promoters and the functional basis of transcription initiation. *Nature reviews. Molecular cell biology.* **19**, 621–637 (2018), doi:10.1038/s41580-018-0028-8.
22. F. H. Crick, L. BARNETT, S. BRENNER, R. J. WATTS-TOBIN, General nature of the genetic code for proteins. *Nature.* **192**, 1227–1232 (1961), doi:10.1038/1921227a0.
23. R. S. GARDNER *et al.*, Synthetic polynucleotides and the amino acid code. VII. *Proceedings of the National Academy of Sciences of the United States of America.* **48**, 2087–2094 (1962), doi:10.1073/pnas.48.12.2087.
24. M. W. NIRENBERG, J. H. MATTHAEI, The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America.* **47**, 1588–1602 (1961), doi:10.1073/pnas.47.10.1588.
25. C. T. Caskey, P. Leder, The RNA code: nature's Rosetta Stone. *Proceedings of the National Academy of Sciences of the United States of America.* **111**, 5758–5759 (2014), doi:10.1073/pnas.1404819111.
26. A. J. WAHBA *et al.*, Synthetic polynucleotides and the amino acid code. VIII. *Proceedings of the National Academy of Sciences of the United States of America.* **49**, 116–122 (1963), doi:10.1073/pnas.49.1.116.
27. J. Ling, N. Reynolds, M. Ibba, Aminoacyl-tRNA synthesis and translational quality control. *Annual review of microbiology.* **63**, 61–78 (2009), doi:10.1146/annurev.micro.091208.073210.
28. R. M. Voorhees, V. Ramakrishnan, Structural basis of the translational elongation cycle. *Annual review of biochemistry.* **82**, 203–236 (2013), doi:10.1146/annurev-biochem-113009-092313.
29. G. Li, M. Levitus, C. Bustamante, J. Widom, Rapid spontaneous accessibility of nucleosomal DNA. *Nature structural & molecular biology.* **12**, 46–53 (2005), doi:10.1038/nsmb869.
30. A. J. Andrews, K. Luger, Nucleosome structure(s) and stability: variations on a theme. *Annual review of biophysics.* **40**, 99–117 (2011), doi:10.1146/annurev-biophys-042910-155329.

31. S. A. Grigoryev, C. L. Woodcock, Chromatin organization - the 30 nm fiber. *Experimental cell research.* **318**, 1448–1455 (2012), doi:10.1016/j.yexcr.2012.02.014.
32. R. C. Allshire, H. D. Madhani, Ten principles of heterochromatin formation and function. *Nature reviews. Molecular cell biology.* **19**, 229–244 (2018), doi:10.1038/nrm.2017.119.
33. L. Armstrong, *Epigenetics* (Garland Science, New York, London, op. 2014).
34. R. Sales-Gil, P. Vagnarelli, How HP1 Post-Translational Modifications Regulate Heterochromatin Formation and Maintenance. *Cells.* **9** (2020), doi:10.3390/cells9061460.
35. J. W. Shay, W. E. Wright, Telomeres and telomerase: three decades of progress. *Nature reviews. Genetics.* **20**, 299–309 (2019), doi:10.1038/s41576-019-0099-1.
36. B. E. Black, L. E. T. Jansen, D. R. Foltz, D. W. Cleveland, Centromere identity, function, and epigenetic propagation across cell divisions. *Cold Spring Harbor symposia on quantitative biology.* **75**, 403–418 (2010), doi:10.1101/sqb.2010.75.038.
37. J. C. Venter *et al.,* The sequence of the human genome. *Science.* **291**, 1304–1351 (2001), doi:10.1126/science.1058040.
38. E. S. Lander *et al.,* Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001), doi:10.1038/35057062.
39. J. M. Claverie, Gene number. What if there are only 30,000 human genes? *Science.* **291**, 1255–1257 (2001), doi:10.1126/science.1058969.
40. D.-K. Niu, L. Jiang, Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and biophysical research communications.* **430**, 1340–1343 (2013), doi:10.1016/j.bbrc.2012.12.074.
41. E. S. Lander, Initial impact of the sequencing of the human genome. *Nature.* **470**, 187–197 (2011), doi:10.1038/nature09792.
42. T. R. Gregory, Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics.* **6**, 699–708 (2005), doi:10.1038/nrg1674.
43. J. Padeken, P. Zeller, S. M. Gasser, Repeat DNA in genome organization and stability. *Current opinion in genetics & development.* **31**, 12–19 (2015), doi:10.1016/j.gde.2015.03.009.
44. K. H. Miga, Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome research : an international journal on the molecular,*

supramolecular and evolutionary aspects of chromosome biology. **23**, 421–426 (2015), doi:10.1007/s10577-015-9488-2.

45. K. H. Miga, Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes.* **10** (2019), doi:10.3390/genes10050352.
46. N. B. Atkin, V. Brito-Babapulle, Heterochromatin polymorphism and human cancer. *Cancer Genetics and Cytogenetics.* **3**, 261–272 (1981), doi:10.1016/0165-4608(81)90093-5.
47. N. B. Atkin, V. Brito-Babapulle, Chromosome 1 heterochromatin variants and cancer: A reassessment. *Cancer Genetics and Cytogenetics.* **18**, 325–331 (1985), doi:10.1016/0165-4608(85)90154-2.
48. R. Berger, A. Bernheim, U. Kristoffersson, F. Mitelman, H. Olsson, C-band heteromorphism in breast cancer patients. *Cancer Genetics and Cytogenetics.* **18**, 37–42 (1985), doi:10.1016/0165-4608(85)90037-8.
49. E. M. Black, S. Giunta, Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes.* **9** (2018), doi:10.3390/genes9120615.
50. N. I. Enukashvily, R. Donev, I. S.-R. Waisertreiger, O. I. Podgornaya, Human chromosome 1 satellite 3 DNA is decondensed, demethylated and transcribed in senescent cells and in A431 epithelial carcinoma cells. *Cytogenetic and genome research.* **118**, 42–54 (2007), doi:10.1159/000106440.
51. D. Ferreira *et al.,* Satellite non-coding RNAs: the emerging players in cells, cellular pathways and cancer. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology.* **23**, 479–493 (2015), doi:10.1007/s10577-015-9482-8.
52. D. T. Ting *et al.,* Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science.* **331**, 593–596 (2011), doi:10.1126/science.1200801.
53. F. I. Sahin *et al.,* Chromosome heteromorphisms: an impact on infertility. *Journal of assisted reproduction and genetics.* **25**, 191–195 (2008), doi:10.1007/s10815-008-9216-3.
54. C. Tyler-Smith, W. R.A. Brown, Structure of the major block of alphoid satellite DNA on the human Y chromosome. *Journal of Molecular Biology.* **195**, 457–470 (1987), doi:10.1016/0022-2836(87)90175-6.
55. R. Bandyopadhyay, C. McQuillan, S. L. Page, K. H. Choo, L. G. Shaffer, Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome*

research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology. **9**, 223–233 (2001), doi:10.1023/a:1016648404388.

56. K. J. Billingsley *et al.*, Analysis of repetitive element expression in the blood and skin of patients with Parkinson's disease identifies differential expression of satellite elements. *Scientific reports*. **9**, 4369 (2019), doi:10.1038/s41598-019-40869-z.
57. Y. Saito *et al.*, Expression of mRNA for DNA methyltransferases and methyl-CpG-binding proteins and DNA methylation status on CpG islands and pericentromeric satellite regions during human hepatocarcinogenesis. *Hepatology (Baltimore, Md.)*. **33**, 561–568 (2001), doi:10.1053/jhep.2001.22507.
58. K. H. Miga *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. **585**, 79–84 (2020), doi:10.1038/s41586-020-2547-7.
59. M. Stadtfeld, K. Hochedlinger, Induced pluripotency: history, mechanisms, and applications. *Genes & development*. **24**, 2239–2263 (2010), doi:10.1101/gad.1963910.
60. J. B. GURDON, The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of embryology and experimental morphology*. **10**, 622–640 (1962).
61. S. L. Klemm, Z. Shipony, W. J. Greenleaf, Chromatin accessibility and the regulatory epigenome. *Nature reviews. Genetics*. **20**, 207–220 (2019), doi:10.1038/s41576-018-0089-8.
62. A. D'Urso, J. H. Brickner, Mechanisms of epigenetic memory. *Trends in genetics : TIG*. **30**, 230–236 (2014), doi:10.1016/j.tig.2014.04.004.
63. M. M. Müller, T. W. Muir, Histones: at the crossroads of peptide and protein chemistry. *Chemical reviews*. **115**, 2296–2349 (2015), doi:10.1021/cr5003529.
64. H. Kimura, Histone modifications for human epigenome analysis. *Journal of human genetics*. **58**, 439–445 (2013), doi:10.1038/jhg.2013.66.
65. H. Santos-Rosa, C. Caldas, Chromatin modifier enzymes, the histone code and cancer. *European journal of cancer (Oxford, England : 1990)*. **41**, 2381–2402 (2005), doi:10.1016/j.ejca.2005.08.010.
66. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell research*. **21**, 381–395 (2011), doi:10.1038/cr.2011.22.

67. Á. Muñoz-López, D. Summerer, Recognition of Oxidized 5-Methylcytosine Derivatives in DNA by Natural and Engineered Protein Scaffolds. *Chemical record (New York, N.Y.).* **18**, 105–116 (2018), doi:10.1002/tcr.201700088.
68. Y. Yin *et al.,* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science.* **356** (2017), doi:10.1126/science.aaj2239.
69. Y. Zeng, T. Chen, DNA Methylation Reprogramming during Mammalian Development. *Genes.* **10** (2019), doi:10.3390/genes10040257.
70. Z. D. Smith, A. Meissner, DNA methylation: roles in mammalian development. *Nature reviews. Genetics.* **14**, 204–220 (2013), doi:10.1038/nrg3354.
71. A. Bird, DNA methylation patterns and epigenetic memory. *Genes & development.* **16**, 6–21 (2002), doi:10.1101/gad.947102.
72. J. A. Law, S. E. Jacobsen, Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics.* **11**, 204–220 (2010), doi:10.1038/nrg2719.
73. S. Ito *et al.,* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science.* **333**, 1300–1303 (2011), doi:10.1126/science.1210597.
74. M. Tahiliani *et al.,* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science.* **324**, 930–935 (2009), doi:10.1126/science.1170116.
75. Y.-F. He *et al.,* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science.* **333**, 1303–1307 (2011), doi:10.1126/science.1210944.
76. J. Yang, N. Bashkenova, R. Zang, X. Huang, J. Wang, The roles of TET family proteins in development and stem cells. *Development (Cambridge, England).* **147** (2020), doi:10.1242/dev.183129.
77. A. Maiti, A. C. Drohat, Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *The Journal of biological chemistry.* **286**, 35334–35338 (2011), doi:10.1074/jbc.C111.284620.
78. D.-Q. Shi, I. Ali, J. Tang, W.-C. Yang, New Insights into 5hmC DNA Modification: Generation, Distribution and Function. *Frontiers in genetics.* **8**, 100 (2017), doi:10.3389/fgene.2017.00100.

79. M. Frommer *et al.,* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America.* **89**, 1827–1831 (1992), doi:10.1073/pnas.89.5.1827.

80. K. Tanaka, A. Okamoto, Degradation of DNA by bisulfite treatment. *Bioorganic & medicinal chemistry letters.* **17**, 1912–1915 (2007), doi:10.1016/j.bmcl.2007.01.040.

81. Z. Staševskij, P. Gibas, J. Gordevičius, E. Kriukienė, S. Klimašauskas, Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. *Molecular cell.* **65**, 554-564.e6 (2017), doi:10.1016/j.molcel.2016.12.012.

82. Y. Liu *et al.,* Dynamic epigenetic states of maize centromeres. *Frontiers in plant science.* **6**, 904 (2015), doi:10.3389/fpls.2015.00904.

83. I. Ohki *et al.,* Solution Structure of the Methyl-CpG Binding Domain of Human MBD1 in Complex with Methylated DNA. *Cell.* **105**, 487–497 (2001), doi:10.1016/S0092-8674(01)00324-5.

84. T. Yamazaki, K. Yamagata, T. Baba, Time-lapse and retrospective analysis of DNA methylation in mouse preimplantation embryos by live cell imaging. *Developmental biology.* **304**, 409–419 (2007), doi:10.1016/j.ydbio.2006.12.046.

85. H. Yoshioka, J. R. McCarrey, Y. Yamazaki, Dynamic nuclear organization of constitutive heterochromatin during fetal male germ cell development in mice. *Biology of reproduction.* **80**, 804–812 (2009), doi:10.1095/biolreprod.108.072603.

86. Y. Hori *et al.,* Synthetic-Molecule/Protein Hybrid Probe with Fluorogenic Switch for Live-Cell Imaging of DNA Methylation. *Journal of the American Chemical Society.* **140**, 1686–1690 (2018), doi:10.1021/jacs.7b09713.

87. J. L. Furman, P.-W. Mok, S. Shen, C. I. Stains, I. Ghosh, A turn-on split-luciferase sensor for the direct detection of poly(ADP-ribose) as a marker for DNA repair and cell death. *Chemical communications (Cambridge, England).* **47**, 397–399 (2011), doi:10.1039/C0CC02229B.

88. C. Lungu, S. Pinter, J. Broche, P. Rathert, A. Jeltsch, Modular fluorescence complementation sensors for live cell detection of epigenetic signals at endogenous genomic sites. *Nature communications.* **8**, 649 (2017), doi:10.1038/s41467-017-00457-z.

89. K. Yamagata, DNA methylation profiling using live-cell imaging. *Methods (San Diego, Calif.)*. **52**, 259–266 (2010), doi:10.1016/j.ymeth.2010.04.008.

90. K. Tanaka, K. Tainaka, T. Kamei, A. Okamoto, Direct labeling of 5-methylcytosine and its applications. *Journal of the American Chemical Society*. **129**, 5612–5620 (2007), doi:10.1021/ja068660c.

91. Y. Li *et al.*, Sequence-specific microscopic visualization of DNA methylation status at satellite repeats in individual cell nuclei and chromosomes. *Nucleic acids research*. **41**, e186 (2013), doi:10.1093/nar/gkt766.

92. U. Bonas, R. E. Stall, B. Staskawicz, Genetic and structural characterization of the avirulence gene avrBs3 from Xanthomonas campestris pv. vesicatoria. *Molecular & general genetics : MGG*. **218**, 127–136 (1989), doi:10.1007/bf00330575.

93. G. V. Minsavage, Gene-For-Gene Relationships Specifying Disease Resistance in Xanthomonas campestris pv. vesicatoria - Pepper Interactions. *MPMI*. **3**, 41 (1990), doi:10.1094/MPMI-3-041.

94. F. van Gijsegem, S. Genin, C. Boucher, Conservation of secretion pathways for pathogenicity determinants of plant and animal bacteria. *Trends in Microbiology*. **1**, 175–180 (1993), doi:10.1016/0966-842X(93)90087-8.

95. K. Wengelnik, C. Marie, M. Russel, U. Bonas, Expression and localization of HrpA1, a protein of Xanthomonas campestris pv. vesicatoria essential for pathogenicity and induction ofthe hypersensitive reaction. *Journal of bacteriology*. **178**, 1061–1069 (1996), doi:10.1128/jb.178.4.1061-1069.1996.

96. C. M. Hopkins, F. F. White, S. H. Choi, A. Guo, J. E. Leach, Identification of a family of avirulence genes from Xanthomonas oryzae pv. oryzae. *MPMI*. **5**, 451–459 (1992), doi:10.1094/mpmi-5-451.

97. M. J. Moscou, A. J. Bogdanove, A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)*. **326**, 1501 (2009), doi:10.1126/science.1178817.

98. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)*. **326**, 1509–1512 (2009), doi:10.1126/science.1178811.

99. R. Moore, A. Chandrahas, L. Bleris, Transcription activator-like effectors: a toolkit for synthetic biology. *ACS synthetic biology*. **3**, 708–716 (2014), doi:10.1021/sb400137b.

100. T. Gaj, C. A. Gersbach, C. F. Barbas, ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology.* **31**, 397–405 (2013), doi:10.1016/j.tibtech.2013.04.004.

101. M. Bochtler, Structural basis of the TAL effector-DNA interaction. *Biological chemistry.* **393**, 1055–1066 (2012), doi:10.1515/hsz-2012-0164.

102. D. Deng, C. Yan, J. Wu, X. Pan, N. Yan, Revisiting the TALE repeat. *Protein & cell.* **5**, 297–306 (2014), doi:10.1007/s13238-014-0035-2.

103. D. Deng *et al.*, Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science (New York, N.Y.).* **335**, 720–723 (2012), doi:10.1126/science.1215670.

104. H. Gao, X. Wu, J. Chai, Z. Han, Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research.* **22**, 1716–1720 (2012), doi:10.1038/cr.2012.156.

105. A. N.-S. Mak, P. Bradley, R. A. Cernadas, A. J. Bogdanove, B. L. Stoddard, The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science (New York, N.Y.).* **335**, 716–719 (2012), doi:10.1126/science.1216211.

106. D. Deng *et al.*, Recognition of methylated DNA by TAL effectors. *Cell research.* **22**, 1502–1504 (2012), doi:10.1038/cr.2012.127.

107. S. Stella *et al.*, Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism. *Acta crystallographica. Section D, Biological crystallography.* **69**, 1707–1716 (2013), doi:10.1107/S0907444913016429.

108. S. Tsuji, S. Futaki, M. Imanishi, Creating a TALE protein with unbiased 5'-T binding. *Biochemical and biophysical research communications.* **441**, 262–265 (2013), doi:10.1016/j.bbrc.2013.10.060.

109. L. Cuculis, Z. Abil, H. Zhao, C. M. Schroeder, Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nature communications.* **6**, 7277 (2015), doi:10.1038/ncomms8277.

110. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research.* **39**, e82 (2011), doi:10.1093/nar/gkr218.

111. C. Engler, R. Gruetzner, R. Kandzia, S. Marillonnet, Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS one.* **4**, e5553 (2009), doi:10.1371/journal.pone.0005553.

112. C. Engler, R. Kandzia, S. Marillonnet, A one pot, one step, precision cloning method with high throughput capability. *PloS one.* **3**, e3647 (2008), doi:10.1371/journal.pone.0003647.

113. P. Deng, S. Carter, K. Fink, Design, Construction, and Application of Transcription Activation-Like Effectors. *Methods in molecular biology (Clifton, N.J.).* **1937**, 47–58 (2019), doi:10.1007/978-1-4939-9065-8_3.

114. Y. Luo *et al.,* Generation of TALE nickase-mediated gene-targeted cows expressing human serum albumin in mammary glands. *Scientific reports.* **6**, 20657 (2016), doi:10.1038/srep20657.

115. Y. Feng, S. Zhang, X. Huang, A robust TALENs system for highly efficient mammalian genome editing. *Scientific reports.* **4**, 3632 (2014), doi:10.1038/srep03632.

116. A. C. Mercer, T. Gaj, R. P. Fuller, C. F. Barbas, Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic acids research.* **40**, 11163–11172 (2012), doi:10.1093/nar/gks875.

117. W. Qasim *et al.,* Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Science translational medicine.* **9** (2017), doi:10.1126/scitranslmed.aaj2013.

118. M. J. Pino-Barrio *et al.,* TALEN mediated gene editing in a mouse model of Fanconi anemia. *Scientific reports.* **10**, 6997 (2020), doi:10.1038/s41598-020-63971-z.

119. C. Mussolino *et al.,* TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. *Nucleic acids research.* **42**, 6762–6773 (2014), doi:10.1093/nar/gku305.

120. K. Bloom, A. Ely, C. Mussolino, T. Cathomen, P. Arbuthnot, Inactivation of hepatitis B virus replication in cultured cells and in vivo with engineered transcription activator-like effector nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy.* **21**, 1889–1897 (2013), doi:10.1038/mt.2013.170.

121. T. Fujita, H. Fujii, Applications of Engineered DNA-Binding Molecules Such as TAL Proteins and the CRISPR/Cas System in Biology Research. *International journal of molecular sciences.* **16**, 23143–23164 (2015), doi:10.3390/ijms161023143.

122. J. C. Miller *et al.,* A TALE nuclease architecture for efficient genome editing. *Nature biotechnology.* **29**, 143–148 (2011), doi:10.1038/nbt.1755.

123. F. Zhang *et al.,* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology.* **29**, 149–153 (2011), doi:10.1038/nbt.1775.

124. S. Bultmann *et al.,* Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic acids research.* **40**, 5368–5377 (2012), doi:10.1093/nar/gks199.

125. J. Hu *et al.,* Direct activation of human and mouse Oct4 genes using engineered TALE and Cas9 transcription factors. *Nucleic acids research.* **42**, 4375–4390 (2014), doi:10.1093/nar/gku109.

126. X. Gao *et al.,* Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. *Nucleic acids research.* **42**, e155 (2014), doi:10.1093/nar/gku836.

127. X. Wang *et al.,* Designed transcription activator-like effector proteins efficiently induced the expression of latent HIV-1 in latently infected cells. *AIDS research and human retroviruses.* **31**, 98–106 (2015), doi:10.1089/AID.2014.0121.

128. Y.-H. Zhan *et al.,* CELSR1 Is a Positive Regulator of Endothelial Cell Migration and Angiogenesis. *Biochemistry. Biokhimiia.* **81**, 591–599 (2016), doi:10.1134/S0006297916060055.

129. J. P. Tremblay, J. Rousseau, P. Chapdelaine, 604. TALE-VP64 Targeting the Frataxin Promoter Increase the Expression of That Gene in Friedreich Fibroblasts. *Molecular Therapy.* **23**, S240 (2015), doi:10.1016/S1525-0016(16)34213-7.

130. Le Cong, R. Zhou, Y.-C. Kuo, M. Cunniff, F. Zhang, Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications.* **3**, 968 (2012), doi:10.1038/ncomms1962.

131. Z. Zhang, E. Wu, Z. Qian, W.-S. Wu, A multicolor panel of TALE-KRAB based transcriptional repressor vectors enabling knockdown of multiple gene targets. *Scientific reports.* **4**, 7338 (2014), doi:10.1038/srep07338.

132. J. Masuda *et al.,* Transient Tcf3 Gene Repression by TALE-Transcription Factor Targeting. *Applied biochemistry and biotechnology.* **180**, 1559–1573 (2016), doi:10.1007/s12010-016-2187-4.

133. K. Bloom *et al.,* Inhibition of replication of hepatitis B virus using transcriptional repressors that target the viral DNA. *BMC infectious diseases.* **19**, 802 (2019), doi:10.1186/s12879-019-4436-y.

134. Y. Zhang *et al.,* Deciphering TAL effectors for 5-methylcytosine and 5-hydroxymethylcytosine recognition. *Nature communications.* **8**, 901 (2017), doi:10.1038/s41467-017-00860-6.

135. S. Tsuji, M. Imanishi, Modified nucleobase-specific gene regulation using engineered transcription activator-like effectors. *Advanced drug delivery reviews.* **147**, 59–65 (2019), doi:10.1016/j.addr.2019.08.011.

136. D. L. Bernstein, J. E. Le Lay, E. G. Ruano, K. H. Kaestner, TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. *The Journal of clinical investigation.* **125**, 1998–2006 (2015), doi:10.1172/JCI77321.

137. C.-L. Lo, S. R. Choudhury, J. Irudayaraj, F. C. Zhou, Epigenetic Editing of Ascl1 Gene in Neural Stem Cells by Optogenetics. *Scientific reports.* **7**, 42047 (2017), doi:10.1038/srep42047.

138. V. Kameswaran *et al.,* The Dysregulation of the DLK1-MEG3 Locus in Islets From Patients With Type 2 Diabetes Is Mimicked by Targeted Epimutation of Its Promoter With TALE-DNMT Constructs. *Diabetes.* **67**, 1807–1815 (2018), doi:10.2337/db17-0682.

139. M. L. Maeder *et al.,* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology.* **31**, 1137–1142 (2013), doi:10.1038/nbt.2726.

140. C. Jiang, J. Guo, H. Cheng, Y.-H. Feng, Induced Expression of Endogenous CXCR4 in iPSCs by Targeted CpG Demethylation Enhances Cell Migration Toward the Ligand CXCL12. *Inflammation.* **42**, 20–34 (2019), doi:10.1007/s10753-018-0869-5.

141. H. Ma, P. Reyes-Gutierrez, T. Pederson, Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. *Proceedings of the National Academy of Sciences of the United States of America.* **110**, 21048–21053 (2013), doi:10.1073/pnas.1319097110.

142. Y. Miyanari, C. Ziegler-Birling, M.-E. Torres-Padilla, Live visualization of chromatin dynamics with fluorescent TALEs. *Nature structural & molecular biology.* **20**, 1321–1324 (2013), doi:10.1038/nsmb.2680.

143. K. Thanisch *et al.*, Targeting and tracing of specific DNA sequences with dTALEs in living cells. *Nucleic acids research.* **42**, e38 (2014), doi:10.1093/nar/gkt1348.

144. H. Hu, X. Yang, C. Tang, Visualization of Genomic Loci in Living Cells with BiFC-TALE. *Current protocols in cell biology.* **82**, e78 (2019), doi:10.1002/cpcb.78.

145. H. Hu *et al.*, Live visualization of genomic loci with BiFC-TALE. *Scientific reports.* **7**, 40192 (2017), doi:10.1038/srep40192.

146. R. Ren *et al.*, Visualization of aging-associated chromatin alterations with an engineered TALE system. *Cell research.* **27**, 483–504 (2017), doi:10.1038/cr.2017.18.

147. R. Taneja, B. K. Kennedy, T(ell)TALE signs of aging. *Cell research.* **27**, 453–454 (2017), doi:10.1038/cr.2017.33.

148. J. Valton *et al.*, Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *The Journal of biological chemistry.* **287**, 38427–38432 (2012), doi:10.1074/jbc.C112.408864.

149. S. Maurer, M. Giess, O. Koch, D. Summerer, Interrogating Key Positions of Size-Reduced TALE Repeats Reveals a Programmable Sensor of 5-Carboxylcytosine. *ACS chemical biology.* **11**, 3294–3299 (2016), doi:10.1021/acschembio.6b00627.

150. G. Kubik, M. J. Schmidt, J. E. Penner, D. Summerer, Programmable and highly resolved in vitro detection of 5-methylcytosine by TALEs. *Angewandte Chemie (International ed. in English).* **53**, 6002–6006 (2014), doi:10.1002/anie.201400436.

151. G. Kubik, D. Summerer, Achieving single-nucleotide resolution of 5-methylcytosine detection with TALEs. *Chembiochem : a European journal of chemical biology.* **16**, 228–231 (2015), doi:10.1002/cbic.201402408.

152. G. Kubik, S. Batke, D. Summerer, Programmable sensors of 5-hydroxymethylcytosine. *Journal of the American Chemical Society.* **137**, 2–5 (2015), doi:10.1021/ja506022t.

153. P. Rathi, S. Maurer, G. Kubik, D. Summerer, Isolation of Human Genomic DNA Sequences with Expanded Nucleobase Selectivity. *Journal of the American Chemical Society.* **138**, 9910–9918 (2016), doi:10.1021/jacs.6b04807.

154. S. Maurer *et al.*, Overcoming conservation in TALE-DNA interactions: a minimal repeat scaffold enables selective recognition of an oxidized 5-methylcytosine. *Chemical science.* **9**, 7247–7252 (2018), doi:10.1039/c8sc01958d.

155. L. Liu *et al.*, Structural Insights into the Specific Recognition of 5-methylcytosine and 5-hydroxymethylcytosine by TAL Effectors. *Journal of Molecular Biology.* **432**, 1035–1047 (2020), doi:10.1016/j.jmb.2019.11.023.

156. S. Tsuji, S. Futaki, M. Imanishi, Sequence-specific recognition of methylated DNA by an engineered transcription activator-like effector protein. *Chemical communications (Cambridge, England).* **52**, 14238–14241 (2016), doi:10.1039/c6cc06824c.

157. P. Rathi, A. Witte, D. Summerer, Engineering DNA Backbone Interactions Results in TALE Scaffolds with Enhanced 5-Methylcytosine Selectivity. *Scientific reports.* **7**, 15067 (2017), doi:10.1038/s41598-017-15361-1.

158. M. Gieß, A. Witte, J. Jasper, O. Koch, D. Summerer, Complete, Programmable Decoding of Oxidized 5-Methylcytosine Nucleobases in DNA by Chemoselective Blockage of Universal Transcription-Activator-Like Effector Repeats. *Journal of the American Chemical Society.* **140**, 5904–5908 (2018), doi:10.1021/jacs.8b02909.

159. M. Gieß, Á. Muñoz-López, B. Buchmuller, G. Kubik, D. Summerer, Programmable Protein-DNA Cross-Linking for the Direct Capture and Quantification of 5-Formylcytosine. *Journal of the American Chemical Society.* **141**, 9453–9457 (2019), doi:10.1021/jacs.9b01432.

160. S. Fujimoto, S. S. Sugano, K. Kuwata, K. Osakabe, S. Matsunaga, Visualization of specific repetitive genomic sequences with fluorescent TALEs in Arabidopsis thaliana. *Journal of experimental botany.* **67**, 6101–6110 (2016), doi:10.1093/jxb/erw371.

161. F. C. Rinaldi, L. A. Doyle, B. L. Stoddard, A. J. Bogdanove, The effect of increasing numbers of repeats on TAL effector DNA binding specificity. *Nucleic acids research.* **45**, 6960–6970 (2017), doi:10.1093/nar/gkx342.

162. S. Tsuji, K. Shinoda, S. Futaki, M. Imanishi, Sequence-specific 5mC detection in live cells based on the TALE-split luciferase complementation system. *The Analyst.* **143**, 3793–3797 (2018), doi:10.1039/c8an00562a.

163. S. Reither, F. Li, H. Gowher, A. Jeltsch, Catalytic Mechanism of DNA-(cytosine-C5)-methyltransferases Revisited: Covalent Intermediate Formation is not Essential for Methyl Group Transfer by the Murine Dnmt3a Enzyme. *Journal of Molecular Biology.* **329**, 675–684 (2003), doi:10.1016/S0022-2836(03)00509-6.

164. D. G. Gibson *et al.,* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods.* **6**, 343–345 (2009), doi:10.1038/nmeth.1318.

165. N. C. Shaner *et al.,* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature biotechnology.* **22**, 1567–1572 (2004), doi:10.1038/nbt1037.

166. J. Schindelin *et al.,* Fiji: an open-source platform for biological-image analysis. *Nature methods.* **9**, 676–682 (2012), doi:10.1038/nmeth.2019.

167. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nature methods.* **9**, 671–675 (2012), doi:10.1038/nmeth.2089.

168. P. Haub, T. Meckel, GaussFit OnSpot (2014).

169. A. D. Edelstein *et al.,* Advanced methods of microscope control using μManager software. *Journal of biological methods.* **1** (2014), doi:10.14440/jbm.2014.36.

170. M. Dowle, A. Srinivasan, *data.table: Extension of `data.frame`. R package version 1.12.6.* (available at https://CRAN.R-project.org/package=data.table).

171. R Core Team, *R: A language and environment for statistical computing.* (2019) (available at https://www.R-project.org/).

172. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).

173. A. Kassambara, *ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.4.* (2019) (available at https://CRAN.R-project.org/package=ggpubr).

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)

5-methylcytosine (5mC) is a fundamental epigenetic modification in mammalian genomes involved in development, cell differentiation and genomic imprinting. In addition, aberrant DNA methylation patterns are responsible for the pathogenesis of many diseases including neurodegenerative disorders, cardiovascular affections and cancer. In this thesis, we describe the development of a novel method for image-based analysis of 5mC using pairs of fluorescent Transcription Activator Like-Effectors (TALEs). These DNA binding proteins can recognize specific sequences of canonical or epigenetically modified DNA via modular repeats that interact with nucleobases in a one-to-one correspondence. We employed fluorescent TALE pairs that differ only in the repeat responsible for recognizing cytosine (C) at CpG dinucleotides and the fluorophore fused to them (either eGFP or mCherry). By using the 5mC selective repeat HD in one of the TALEs, we can detect differences in methylation level, while the universal binder repeat G* in the other TALE is not responsive to 5mC and allows to detect local changes in chromatin compaction. This way it is possible to analyze 5mC independently of potential differences in target accessibility. We applied our method using recombinantly expressed and purified TALE pairs in cellular stains to image SatIII DNA. This pericentromeric DNA is the origin of nuclear stress bodies (nSBs), exhibits aberrant methylation in several cancers and remains challenging to study by conventional methods due to its highly repetitive nature. We proved the applicability of our method to study 5mC differences in user-defined repetitive sequences with single nucleotide and strand resolution. Furthermore, we correlated the methylation status of SatIII with the presence of heat shock factor 1 (HSF1) at its recognition sequence after stress, revealing a role for 5mC in HSF1 recruitment as initial step of nSB formation in a subpopulation of cells. Finally, we constructed and screened a library of size-reduced TALE repeats to identify potential 5mC binders. We found that RVD NH* binds selectively to 5mC, but not C and its application in combination with HD TALEs allows for improved imaging with higher dynamic range. These studies offer a promising imaging tool for studying 5mC function in repetitive sequences and its interplay with other imageable chromatin-interacting proteins with nucleotide, strand, locus and cell resolution.

Álvaro Muñoz López

technische universität dortmund

IMPRS in Chemical and Molecular Biology