Advancing Quantitative DNA Biomarker Detection through Single Molecule Fluorescence Kinetic Fingerprinting

by

Liuhan Dai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2024

Doctoral Committee:

Professor Nils G. Walter, Chair
Professor Ryan Castle Bailey
Professor Neil Marsh
Professor Muneesh Tewari

Liuhan Dai

liuhand@umich.edu

ORCID iD: 0000-0002-9742-236X

Dedication

This Thesis is dedicated to my family and friends.

Acknowledgements

My PhD journey started off even earlier than the beginning of my PhD program, when I was still a senior undergraduate as a visiting scholar in Nils Walter's lab. Mentored by Jieming Li who was a 5-year PhD student at that time, I was able to have my first real taste of scientific research. And it was the visiting experience at his lab that ultimately got me admitted into the University of Michigan-Ann Arbor. Therefore, the very first person I want to thank is Nils Walter, later becoming my PhD advisor in the following five and half years.

Nils really granted me great freedom and financial support to continue my PhD research. Although managing such a big lab, he was always patient enough for my troubleshooting process. Ultimately, I grew from a naïve undergraduate to an independent researcher with probably the most troubleshooting experience in his lab. And I believed these knowledge and skills will well prepare me for future postdoc journey.

I would also like to thank Jieming and Alex, who mentored me during my first three years of PhD. Jieming was really the one who brought me into the single-molecule field. And Alex was the "god father" of the methodology used by me throughout my PhD. He was the big master of our techniques and always willing to answer any questions from career development to experimental questions. Deep in my heart, he was really the model scientist that I dreamed of becoming.

I also want to have special thanks to Kunal and Jingxuan. Kunal was my best friend during my PhD, and I really miss those old days when we stayed together and played video

games. We went grocery shopping every week and talked about various aspects of life. He was really the friend I can rely on during my difficult moments. Jingxuan is my significant other. Interestingly, we met when she was a rotation student in Nils Walter's lab. So, I guess this lab was also my romantic bookmark. Especially when I was writing my thesis, she was always accompanying me and comforting me. She was the one who really appreciated good things in myself and spoke to me when I was so blind to my own difficulties and stuck to my own little world. It was her company that supported me to hold till the end of my PhD journey.

I also want to express gratitude to my lab mates, previous ones, and current ones: Pavel with whom we work together pleasantly on one project; Sujay whom I think first to ask when encountering some scientific problems; my dear SiMREPS microgroup peers, Karen, Tanmay, Shankar, Zi and Paul; Emily E., Rosa, Ameya, Shiba, Yichen, Guoming and Emily S., all grew together and learned to be independent researchers; Andreas, helped me out with molecular cloning when I was a newbie; Robb, was always willing to talk with me about interesting facts in academia and society. I also want to thank my excellent peers in both Chemistry program and Biophysics program, Mochen with whom we played basketball vividly, Ziyuan and Zhenyu whom kindly sent their oral presentations to me for reference, Minjun and Renjie whom invited me for meals endless times, Sicong, Katie and their little baby Lecheng whom brought happiness and hope in life to me, Yihua whom took care of Paddington for three months, Jiameng and Xingyao whom treated me with delicious marinated steak and lamb, etc. It is my pleasure to spend my graduate school years with these wonderful and amazing people.

Finally, I want to thank my parents and family. They support me unconditionally and always encourage me to live up to incoming challenges and to become proud of my past, current

and future. They are my life model and demonstrate great spirit against any adversities, motivating me to always pick myself back and keep pushing forward!

Table of Contents

List of Tables

List of Figures

Abstract

DNA (deoxyribonucleic acid), as the primary genetic material in eukaryotic organisms, is one of the major biomarkers for biological processes, pathological processes and drug responses in clinical diagnostics and prognostics. Genetic variations and DNA methylation are two major types of disease-related DNA biomarker. The gold standard for detecting DNA disease-related biomarkers is PCR (polymerase chain reaction) amplification-based approach, consisting over 50% FDA (U.S. Food and Drug Administration)-approved in vitro diagnostic tests. However, it suffers from two significant technical limitations: biased estimation due to unequal amplification and limited specificity. These issues are largely overlooked within the scientific community.

In this Dissertation, I developed amplification-free detections for DNA methylation cancer biomarkers and DNA mutation cancer biomarkers by using single-molecule fluorescent kinetic fingerprinting (SMFKF) based on single-molecule recognition through equilibrium Poisson sampling (SiMREPS). SMFKF or SiMREPS-based assays offer high-confidence identification of targeted biomarkers as a fluorescent imager repeatedly probes the same molecule. This dissertation consists of two parts: 1) Chapter 2 and 3, developing quantitative SMFKF biosensor for detecting DNA methylation biomarkers; 2) Chapter 4, developing ultrafast quantitative SMFKF biosensor for detecting DNA mutation biomarkers.

Chapter 2 introduced bisulfite Me-SiMREPS (BSM-SiMREPS) that bisulfite-converted methylated 102 nt BCAT1 promoter, a DNA methylation biomarker for colorectal cancer and immobilized it specifically through DNA hybridization with designed probes. Fluorescent DNA

imagers then identify and quantify these immobilized molecules by kinetic filtering. This amplification-free approach yielded a sub-femtomolar limit of detection and 99.9999% specificity for pure DNA methylation biomarker. And eventually, BSM-SiMREPS measured a 31% methylation level of BCAT1 promoter in whole blood DNA, exposing the significant underestimation by PCR-based measurement.

However, DNA degradation and limited conversion efficiency during bisulfite conversion compromise overall sensitivity and specificity. To address these limitations, Chapter 3 introduced a reversible but specific methylation binder – MBD (methyl-binding domain) and achieved direct, amplification-free methylation detection of 55 nt BCAT1 promoter using MBD-SiMREPS. Effects of methylation patterns and sensor structures on methyl-binding kinetics at single-molecule level were investigated with 36 constructs. We discovered that a "branch" motif on the unmethylated reverse strand in a hemimethylated double-stranded DNA facilitated MBD binding, negating the necessity for methylation on the reverse strand. This unexpected MBD behavior could offer novel insights into gene regulation by MBD superfamily in vivo.

In Chapter 4, the potential of fluorogenic imagers was explored. Fluorogenic DNA imagers augment sensitivity of SiMREPS detection by acquiring more fields of view within a similar total acquisition time. We demonstrated fluorogenic SiMREPS (FG-SiMREPS) for detecting three DNA cancer biomarkers: T790M, L858R and HPV. 2-second detections were achieved for all three biomarkers using micromolar fluorogenic imagers. Eventually, we enabled 104-FOV (field of view) scanning detection in just 5 min. However, the inability to distinguish false positives due to significant non-specific interactions was a setback. To address this, further optimizations on surface passivation and imaging conditions are necessary.

Chapter 1 Introduction

DNA is the abbreviation for deoxyribonucleic acid, the genetic materials of all organisms, albeit not for some viruses. It is a polymer consisting of nucleotides as its monomers, connected by phosphodiester bonds. DNA was first discovered and given the name of "nuclein" by Swiss chemist Friedrich Miescher, who extracted it from human white blood cells in 1869[1]. Over the following 60 years, the three major components of each nucleotide were revealed – a phosphate, a deoxyribose and one of four nitrogen-containing nucleobases (adenine [A], guanine [G], cytosine [C] and thymine [T]). In 1919, Russian biochemist Phoebus Levene discovered the order of the three major components of each nucleotide[2]; and in 1950, the American biochemist Erwin Chargaff discovered by thin layer chromatography the so-called "Chargaff rule"[3]. That is, the total amount of purines (A+G) and the total amount of pyrimidines (T+C) are almost always equal, even across species. In 1944, American biochemistry Oswald Avery had already proven that DNA is the hereditary material, by using the transforming principle discovered by British bacteriologist Frederick Griffith[4,5]. Shortly after this proof of a biological role, in 1953 the discovery of the DNA double helix by James Watson and Francis Crick ultimately explained the "Chargaff rule" and allowed researchers to infer the relationship between its structure and function, thus opening the era of molecular biology[6,7].

1.1 DNA biomarkers and personalized medicine

Since DNA encodes genetic information for all higher organisms, any hereditary disease can be traced back to germline abnormalities in DNA sequence or DNA modification. Acquired diseases, caused by somatic alternations due to exposure of environmental factors, can also be

associated with abnormalities in DNA sequence or DNA modification that result in disorder of gene expression and regulation. In fact, DNA abnormality is a common cause of a wide variety of diseases.

As formally defined by the Biomarkers Definitions Working Group[8,9], "A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention." A DNA biomarker in this dissertation consists of two categories: 1) genetic variations such as single nucleotide polymorphisms (SNPs), copy number variations (CPVs) short tandem repeats (STRs), deletions, insertions, or other variation on the DNA sequence level; and 2) epigenetic modifications on DNA such as DNA methylation or other variation in DNA modifications. By identifying and quantifying DNA biomarkers from genomic and epigenomic information obtained from individual patients, personalized medicine approaches can be utilized for specific medical treatments or healthcare decisions for a subgroup of patients with promise to achieve the best treatment outcome.

1.1.1 Genetic variations as disease biomarkers

In 2008, the 1000 Genomes Project was initiated to construct a comprehensive map of human genetic variation by sequencing genomes of thousands of individuals from diverse populations around the world. The final consortium contains reconstructed genomes of 2504 individuals from 26 populations[10]. This project greatly broadens and deepens our knowledge and understanding of the complexity and diversity of genetic variations in the human genome. Evident from the 1000 Genomes Project as well as its succeeding initiatives such as the International Genome Sample Resource (IGSR)[11] and the Genome Aggregation Database (gnomAD)[12], genetic variations can be generally classified by the size of variants in three

categories: 1) single-nucleotide variants (SNVs), a single nucleotide that is altered or substituted in the genome sequence; 2) small insertions-deletions (indel) that cause sequence differences between 1 bp (base pair) – 49 bp; 3) structural variants that encompass at least 50 bp between two individual genomes.

SNVs are the most abundant and well-characterized genetic variations in the human genome in terms of raw numbers. There are 4 to 5 million SNVs in a genome when comparing the DNA sequences of any two unrelated individuals[13]. For a specific SNV with an allele frequency higher than 1% in the population, these variations are called single nucleotide polymorphisms (SNPs). One of the best-known examples of a SNV-related disease is sickle-cell anemia. Sickle-cell anemia is a monogenic inherited disease that is caused by a point mutation and follows Mendelian inheritance. A single T→A mutation in the HBB gene that encodes hemoglobin causes the protein to form an atypical structure when oxygen levels are low [14,15]. This distorts red blood cells into sickle-shaped cells and consequently blocks blood flow, but also confers some protection from malaria and thus is prevalent in sub-Saharan Africa[15]. Monogenic disorders, though easily identified and well-understood, are much rarer compared to polygenic diseases. In fact, the majority of commonly seen diseases such as osteoporosis, diabetes, cardiovascular and inflammatory diseases, psychiatric disorders and most cancers are caused by the combined effects of multiple SNVs in different genes, with each contributing a small effect[16]. Different combinations of SNVs can also give rise to similar phenotypes. To tackle the diversity and heterogeneity of these complex multifactorial disorders, genome-wide association studies (GWAS) screen common SNPs to identify those associated with a disease phenotype. GWAS tests common SNPs across human genome to ask whether the allele frequency of a particular SNP is significantly over- or under-represented among the disease

cohort. The first successful GWAS was published in 2002[17]. It discovered two SNPs that had significantly different allele frequencies in patients with increased risk for myocardial infarction Moreover, the authors performed in vitro functional studies and further unveiled the causal effect of these two SNPs, hinting at a potential molecular mechanism underlying the pathogenesis of the disorder.

Indels are the second most abundant genetic variation in the human genome. There are between 700,000 to 800,000 indel sites per individual genome[13]. Despite the large number of variation sites, indels and their associations with diseases are insufficiently characterized compared to SNVs. Indels can occur in both protein-coding sequences and non-coding sequences. A coding indel that is not a multiple of 3 causes a frameshift mutation and consequently a large change in the polypeptide translated from downstream DNA sequences in the same exon, thus abolishing gene function; by contrast, a coding indel that is a multiple of 3 causes a non-frameshift mutation and results in the insertion or deletion of one or several amino acids in protein sequence. A non-coding indel at regulatory gene elements like promoters, enhancers, splicing or transcription factor binding sites can also modulate transcription. One of the most common genetic diseases in humans, cystic fibrosis, is caused by a coding indel polymorphism (the indel with an allele frequency higher than 1%) in both alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene that eliminates a single amino acid by non-frameshift or in frame deletion[18,19].

Structural variants (SVs) encompass any differences at least 50 bp in length between two individual genomes. They also include insertions, deletions, duplications, inversions, translocations and copy number variations (although these subtypes' definitions sometimes overlap). There are between 23,000 – 28,000 SVs per individual genome and the total length of

SVs covers 0.19% of the entire human genome, being the largest size difference between two haplotypes among the three genetic variations[13]. Disease-related SVs have been historically well characterized in cytogenetics clinics by karyotyping to observe chromosomal aneuploidies or segmental copy-number variations of megabases of DNA[20,21]. Many such disorders are Mendelian sporadic disorders such that a single SV is sufficient to cause the disease. One of the best-known Mendelian SVs is Down syndrome caused by an extra full or partial copy of chromosome 21. The additional full or partial chromosome 21 can either be a separate copy or translocated to a different chromosome.

1.1.2 DNA methylation as cancer biomarkers

Epigenetics is the study of heritable traits or phenotypes that arise from stable distinctions in gene expression and regulation without changes to the underlying DNA sequence. Key players in epigenetic control are non-coding RNAs, DNA modifications and post-translational histone modifications. Different species possess different types of DNA modifications for epigenetic regulation[22,23]. In humans, DNA methylation and its oxidative products are the predominant players in DNA epigenetic modification[24,25].

DNA methylation refers to 5mC, the addition of a 5-methyl group to cytosine in CpG dinucleotides of DNA. The discovery of 5mC as a structural component of natural nucleic acid dates to 1925 when it was found in Tubercle bacillus[26]. Over the past thirty years, researchers have linked DNA methylation to heritable transcriptional repression in vertebrates. DNA methylation-mediated gene silencing plays a crucial role in various biological processes, such as mammalian development, X chromosome inactivation, genomic imprinting, and genome stability, etc.[27–31].

Approximately 40 million to 60 million 5mCs are present in haploid human genome, constituting about 1% of our DNA[32,33]. 5mC almost exclusively occurs at CpG dinucleotides and over 70-80% of CpG dinucleotides are methylated, creating a global ubiquitous methylation landscape[34]. The remaining 20-30% or less unmethylated CpG dinucleotides are clustered in regions known as CpG islands (CGIs)[34]. CGIs, defined as regions over 200 bps with a GC content exceeding 50% and an observed-to-expected CpG ratio greater than 60%, represent GC-rich sequences with elevated CpG-level. Over two thirds of mammalian promoters are CGIs and virtually all housekeeping genes have CGI promoters, the majority of which remain unmethylated at any time throughout normal development and cell division[35]. This local hypomethylation at promoter regions is fundamental for establishing more open chromatin states, signifying an active, or readily activatable, expression status of these genes.

Global ubiquitous methylation with focal unmethylated gaps at CGIs in regulatory elements like promoters and enhancers defines the methylation profiles in normal human cells. Notably, aberrant methylation profiles have been implicated in numerous diseases, particularly cancer, where dysregulation of DNA methylation commonly plays a role in tumorigenesis[36,37]. A defined feature in tumor cells is global hypomethylation (around 40%-60% CpG methylation level) and abnormal focal hypermethylation of CGIs in promoters of cancer-related genes, for example, tumor suppressor genes[38]. In fact, silencing of tumor suppressors by promoter hypermethylation is one of the major drivers in carcinogenesis[34]. Since 2000, thousands of clinical studies have been published to investigate the diagnostic and prognostic potential of DNA methylation biomarkers in numerous diseases, although only a handful of in vitro tests (IVT) have been approved for clinical practice[36]. Promoter methylation of the O6-methylguanine methyltransferase (MGMT) gene is the earliest extensively studied DNA methylation

biomarker[37,39] . Resistance response of malignant gliomas to alkylating agent chemotherapy, the main treatment approach in gliomas, is mainly conferred by MGMT. In 2000, Esteller *et al.* first reported that MGMT promoter methylation is correlated with the regression of tumors, and prolonged overall and disease-free survival in patients treated with carmustine (an alkylating reagent)[29]. A later clinical trial in 2005 by Hegi and colleagues further validated Esteller's findings by showing a clear survival benefit in Temozolomide-treated glioblastoma patients with hypermethylated MGMT promoter methylation[40]. Nowadays, detection of MGMT promoter methylation is a standard practice in predicting resistance response to alkylating agents in glioblastoma patients[36].

1.2 Detection methodology of DNA disease biomarkers

Identifying and quantifying disease-related changes in DNA sequence or methylation generally consist of three steps: 1) biopsy and extraction of DNA from sample matrices such as body fluidics (liquid biopsies) or tissues (tissue biopsies); 2) pretreatment of DNA by amplification-based enrichment, amplification-free enrichment, or no enrichment; 3) signal generation, processing and estimation of the amount of DNA biomarker. In clinical studies, diagnostic or prognostic inference is then achieved based on the results of DNA biomarker quantification.

Detection methods of DNA disease biomarkers can be classified in various ways. For example, depending on whether an assay involves amplification, we can distinguish amplification-based approaches versus amplification-free approaches. Based on different forms of signal response, there exist fluorescence, luminescence sensors, electrochemical, colorimetric, surface plasma resonance (SPR), gravimetric or Ramon spectroscopic sensors, chromatography, and mass spectroscopy, etc.[41–44]. Depending on whether a DNA of interest covers the entire

genome or just a few loci, there are epigenome/genome-wide association studies and locus-specific detection approaches. However, even though numerous disciplines are leveraged in developing detection approaches, few of them are translated into clinical practice. In fact, most clinically approved detection assays of DNA disease biomarkers are based on PCR (polymerase chain reaction).

1.2.1 PCR-based detection methods

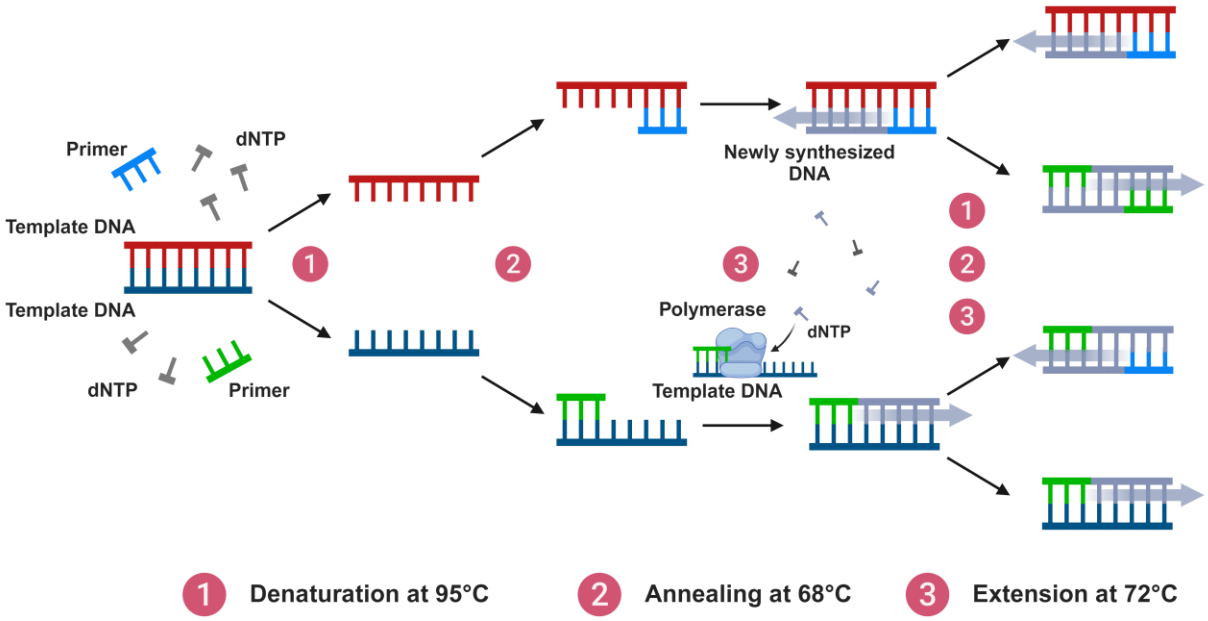

Figure 1.1. Schematic of PCR amplification. dNTP, deoxyribonucleotide triphosphate. Figure made using Biorender.

The most prominent amplification approach is amplification via the polymerase chain reaction (PCR), which was invented by Kary Mullis and won him the Nobel prize in 1993[45,46]. PCR has become an indispensable procedure in numerous assays used in a wide variety of fields, including forensic science, molecular diagnostics, epidemiology, and bioinformatics, etc.

PCR is capable of generating billions of copies of DNA sequences of a specific region of interest. In the simplest case of PCR as shown in **Figure 1.1**, we start with a template DNA at a high temperature, which dissociates a duplex DNA into single strands. Secondly, upon cooling down the system a pair of DNA primers hybridizes to specific sites of the two complementary template strands at some distance from one another that defines the amplicon length in between. Upon hybridization, a heat-resistance DNA polymerase starts extending from the primer-binding sites by incorporating individual dNTPs to replicate the template DNA. The final product at the end of one round of amplification are two ssDNA (single-stranded DNA) molecules that have the same sequence as the two template strands. This newly synthesized DNA serves as template for the next round of amplification. Ultimately, the number of amplicon copies will grow exponentially until the starting materials are exhausted.

Amplification-based detection approaches generally have great sensitivity since ideally a very small amount of target DNA can be amplified almost endlessly. Digital droplet PCR (ddPCR) has reliably demonstrated a limit of detection (LOD) lower than 0.5 copies of DNA per microliter of PCR reaction mix, which is equivalent to lower than 1 aM [47]. However, it also significantly suffers from off-target amplification, causing significant false positives[48]. In the end, its specificity is fundamentally limited by the binding selectivity of the primer sequences themselves, which is governed by thermodynamics of primer binding to the DNA of interest relative to off-target DNAs in the analyte matrix.

PCR is one type of nucleic acid amplification reaction where the amplicon, or molecule whose quantity is amplified, is nucleic acid. In detection of DNA biomarkers, nucleic acid amplification may be a way of enrichment or a process of directly generating signal. More recently, contrary to the non-isothermal nature of PCR, a variety of isothermal amplification

techniques, where temperature is held constant throughout the entire process, has been developed, including the hybridization-chain reaction (HCR), loop mediated isothermal amplification (LAMP), rolling circle amplification (RCA), and recombinase polymerase reaction (RPA), etc.[49,50]. Isothermal amplification is thought to be an advance over PCR due to its accessibility since it does not require a thermal cycler to program temperature changes. Nevertheless, PCR is still the gold standard amplification approach due to its broad commercial availability and wealth of knowledge of protocols and guidelines for implementation.

In ddPCR or qPCR (quantitative PCR), the amplification reaction itself generates an amplified signal response for identification and quantification of DNA biomarkers. However, a more common usage of PCR is enrichment, which is then combined with other techniques such as sequencing and microarrays. PCR is a basic procedure for library preparation in both next generation sequencing and microarray techniques. Genomic DNA fragments, uniformly or partially amplified, are immobilized onto a surface for downstream sequencing assays such as pyrosequencing, Illumina sequencing, PacBio sequencing and Nanopore sequencing, or for microarrays, which entail surfaces coated with a matrix of DNA probes with defined locations and sequences. The rate and efficiency of immobilization is always a fundamental limit on the sensitivity of any surface-based assay[51]. Without amplification, DNA fragments of low abundance would never be captured so that only around 1% of genomic sequences will be able to generate signals due to the need for diffusion-limited mass transport to the surface[52]. This is especially important in liquid biopsies where the copy number of circulating nuclear DNAs in patients is below 10 aM in terms of molar concentration[53].

Regardless of whether it is used as a tool of analyte enrichment or direct amplification of signals, it always remains a question whether PCR equally amplifies different DNA sequences

under a specific reaction condition[54,55]. Unbiased amplification is necessary for determining a 5mC methylation fraction since methylation levels are calculated as a ratio of amplified bisulfite-converted methylated and unmethylated sequence. An unbiased treatment in each step including amplification is therefore essential for an accurate measurement. Another challenge is off-target amplification due to limited primer binding specificity. Fundamentally, a thermodynamic difference of primer binding to a sequence of interest and an off-target sequence is the upper bound of selectivity in any PCR reaction when aiming to quantify genetic variations. Small genetic variations, especially SNVs, are difficult to distinguish due to this physical limitation.

1.2.2 Amplification-free detection methods

Exhausting the list of amplification-free detection methods is impossible given the huge number of categories, rendering it out of scope for this dissertation. In this section, we will highlight several cutting-edge detection methods that are comparable to PCR-based approaches in terms of sensitivity and specificity.

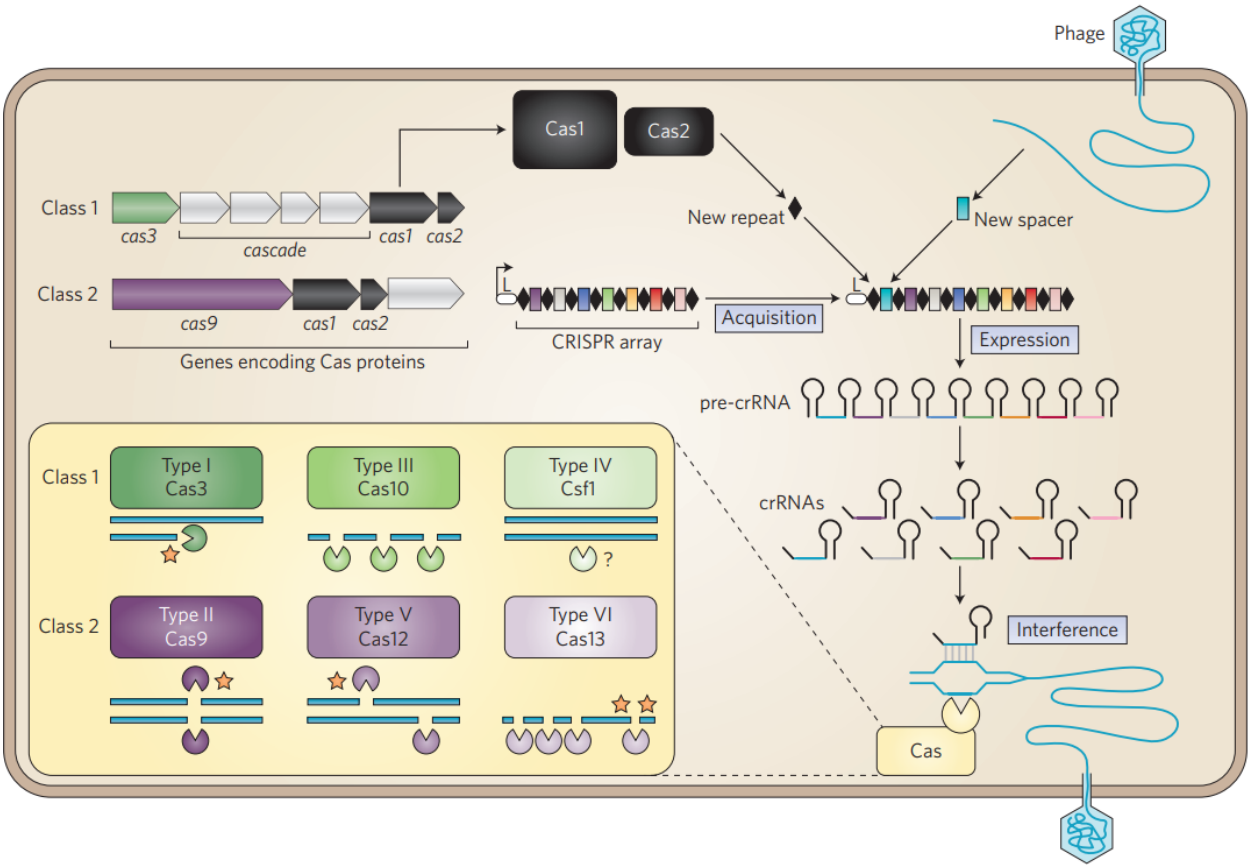

Figure 1.2. CRISPR classification and mechanism of action. Classification of CRISPR-Cas systems hinges on whether a system deploys multiple (class 1) or a single multidomain nuclease (class 2) to clear targeted nucleic acids. In both classes, cas operons (shown as white blocks) are adjacent to CRISPR array. In the above example where a phage virus invades a bacterium, viral genomic DNA is processed as a new spacer and attached with a repeat sequence (black diamond) when incorporated in CRISPR array. CRISPR array can then be transcribed into pre-CRISPR RNAs (pre-crRNAs) and further processed into CRISPR RNA crRNAs that are complementary to viral genomes. Hybridization of crRNAs with newly invading viral DNAs activates nuclease activity of Cas proteins. The cleavage of specific viral DNAs enables protection of microorganisms themselves. The yellow box also shows mechanisms of different Cas nucleases where single blue line represents mRNA transcribed by viral genomes and double blue line represents viral DNAs. Figure taken from Barrangou and Horvath[56].

CRISPR-Cas (CRISPR, clustered regularly interspaced short palindromic repeats; Cas, CRISPR associated proteins) stands out as a revolutionary genome-editing tool and won the Nobel Prize in Chemistry in 2020 for its transformative applications. In nature, it functions as an adaptive immunity system in most bacteria and archaea[56]. The CRISPR locus is a DNA array that stores genetic memories of past viral infections. As shown in **Figure 1.2**, based on DNA fragments in the CRISPR array that originated from prior hostile invaders, Cas proteins generate

crRNAs (CRISPR RNAs) to identify specific invader nucleic acids and neutralize them through the nuclease activities of other Cas proteins or specific domains of the same Cas proteins. So far, 2 major classes, 6 types and 33 subtypes of CRISPR-Cas systems have been discovered, with the class 2 most rapidly expanding[57]. Classification hinges on whether a system deploys multiple (class 1) or a single multidomain nuclease (class 2) to clear targeted nucleic acids. Harnessing its sequence-dependent cleavage activity, CRISPR-based diagnostics (CRISPR Dx) has emerged as a promising tool for nucleic acid detection, highlighted in recent reviews[58,59]. CRISPR Dx often consists of pre-amplification by PCR or isothermal amplification and subsequent cleavage by a Cas enzyme. Notably, there are also amplification-free CRISPR Dx approaches, well-suited for samples with a relatively high concentration of nucleic acids. Class 2 systems dominate CRISPR Dx assays due to the ease of their reconstitution from a single polypeptide chain; of them, Cas12 (type V, cleaving both double-stranded and single-stranded DNAs) and Cas13 (type VI, cleaving single-stranded RNAs) are the most used effectors in pre-amplification-free CRISPR biosensors[60–64]. Both enzymes require a complementary crRNA that bind target sequences and then activate their trans-cleavage activity against reporter sequences. These reporters are often dual-labeled with a fluorophore and a quencher, remaining non-fluorescent until target-dependent cleavage occurs to separate fluorophore and quencher and generate a fluorescence signal. For example, a paper published in 2021 deployed Cas13a for the quantitative detection of SARS-CoV-2 viral RNA and achieved a LOD of 100 copies per µl in patient samples[64]. This advance underscores the continuing development and applications of CRISPR-Cas system in molecular diagnostics. However, CRISPR Dx relies on binding of crRNA to targeted RNA or DNA and thus, its specificity is limited by the thermodynamic binding energy difference when distinguishing epigenomic or genomic variations, especially SNVs.

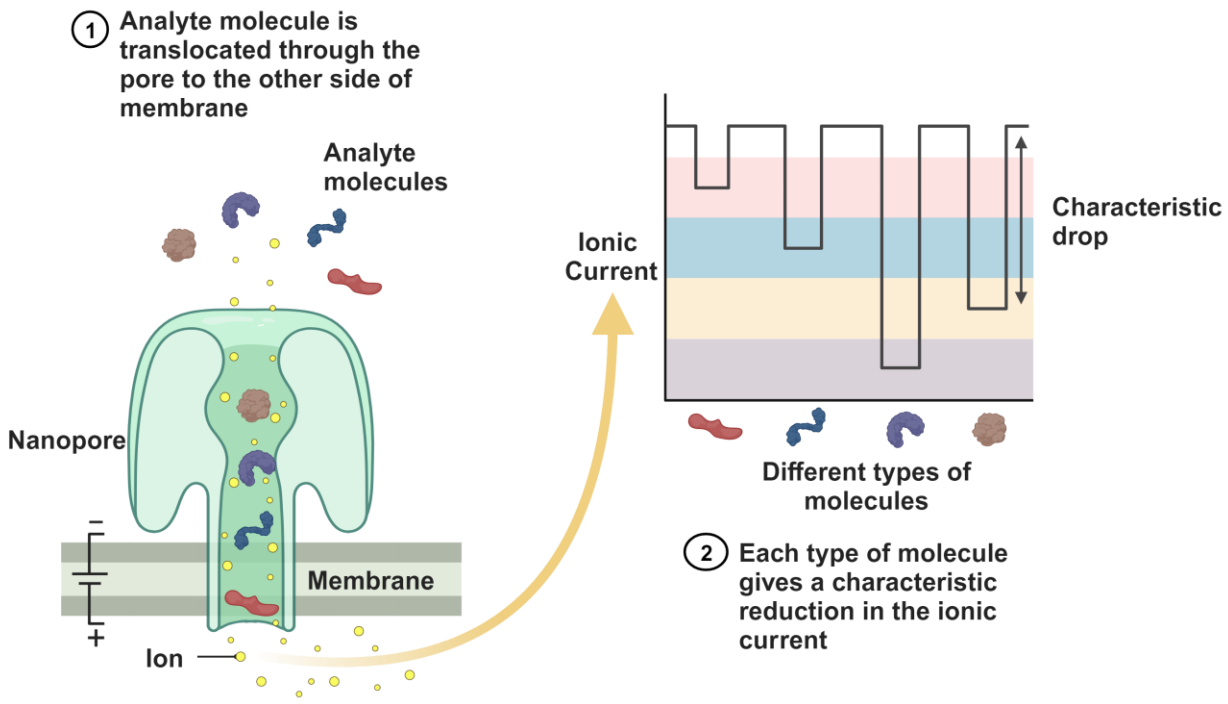

Figure 1.3. Generic schematic of nanopore-based detection. Figure made using Biorender.

Nanopore technology has emerged as another democratizing platform for the detection of nucleic acid disease markers. As its most notable application, nanopore-based single-molecule DNA/RNA sequencing is rapidly supplanting Sanger sequencing as the go-to option for plasmid sequencing. This shift has been catalyzed by increased commercial availability. Due to its superior portability, long-read capability, and great suitability for analyzing repeat sequences, nanopore sequencing has significantly advanced genomics, transcriptomics and epigenomics research[65,66]. In general, a nanopore is a nanometer-sized pore-like or hole-like structure embedded in a dielectric thin material. As shown in **Figure 1.3**, in a typical nanopore measurement, individual analytes enter the nanopore under an applied electric potential, altering the ion flow through its interior volume – a change reflected in a time-dependent current recording with microsecond temporal resolution. Any modulation of amplitude, duration and frequency in ionic current constitutes the single-molecule electric fingerprint of its size, shape,

and conformation. Nanopores can be classified into two categories by their manufacturing materials – biological nanopores (for instance, α-hemolysin embedded in a lipid bilayer) and solid-state nanopores, crafted in thin inorganic or plastic layer like silicon, and graphene, etc. Nanopore-based biosensors do not need any labeling to detect analytes of interest. Since it successively registers single molecules one after the other, an ionic current signature is measured for every molecule, enabling identification of molecules of interest among all translocation events, therefore providing higher sensitivity than an ensemble-level measurement. Consequently, nanopore-based biomarker detection of, e.g., DNA typically yields a sub-picomolar detection limit, even without any enrichment or signal amplification [67]. A recent study uses solid-state nanopores to detect circulating tumor DNA mutations in blood samples by supplementing the electrical signal response with a "cooccurring" fluorescence response generated by a dye-labeled complementary oligonucleotide reporter, highlighting a popular strategy for increasing specificity in nanopore-based detection [68]. However, the emerging trend of electro-optical sensing in nanopore biosensor also underlines a long-existing challenge in nanopore detection: sorting out or preventing non-target current blockades, especially in clinically relevant samples that often contain abundant matrix constituents that can clog the nanopore sensor itself[69].

1.3 Single-molecule fluorescence kinetic fingerprinting (SMFKF)

As a single-molecule electrical biosensor, nanopore-based measurements already demonstrated greater sensitivity and selectivity. However, the same molecule can only be measured once until the end of its translocation. To reach a femtomolar level of sensitivity, the idea of repeatedly measuring the same molecule to ensure a robust identification of target and a

distinct separation of non-target signals, is brought to fruition by an innovative methodology, single-molecule fluorescent kinetic fingerprinting (SMFKF).

1.3.1 Basic principles of SMFKF – SiMREPS

SMFKF is based on the principles of single-molecule recognition through equilibrium Poisson sampling (SiMREPS), a term first coined by Alex Johnson-Buck and Xin Su *et al.* in 2015[70]. In this dissertation, SMFKF is strictly defined as SiMREPS-based methodology. So far, a collection of SMFKF methods has been developed for detecting a wide range of analytes, including RNA, DNA, protein, and small molecules[52,70–78].

To fully understand SiMREPS, let us shift our discussion to a bimolecular association reaction from a single-molecule perspective:

$$A + B \leftrightarrow AB \tag{1.1}$$

In this reversible elementary reaction, molecule A reacts with a molecule B and forms product AB. Let us assume that A is sparsely tethered on a surface and B is labeled with a favorite fluorophore. Imagine that a reservoir of B is added to the surface where a laser is illuminating. By some trick, we restrict our observation only to that illuminated surface. When a B molecule diffuses to an A molecule, B may stay bound with A and eventually form AB. Consequently, we can observe a fluorescence signal lasting on the surface for a certain period. However, because this is a reversible reaction, product AB that was just formed can also dissociate into molecules A and B. B may then diffuse away from the surface and A remains tethered to the same spot. Thus, we observe a dark period at this B-free A spot. Similarly, a different B molecule may repeat the same process as it binds and dissociates from the same A-occupied spot. Ultimately,

different B molecules in the reservoir are "probing" the same A molecule repeatedly, which is reflected by alternating fluorescence on and off signals.

In the above scenario, association of A and B will exactly follow second-order reaction kinetics. From a single-molecule perspective, what this means is that binding and dissociation of B is a Poisson process/sampling. Specifically, the state of B being either bound or unbound to the same molecule A indexed by time is a continuous Poisson process, which can be easily visualized by plotting the fluorescence intensity (F.I.) at a particular A-occupied spot over time:

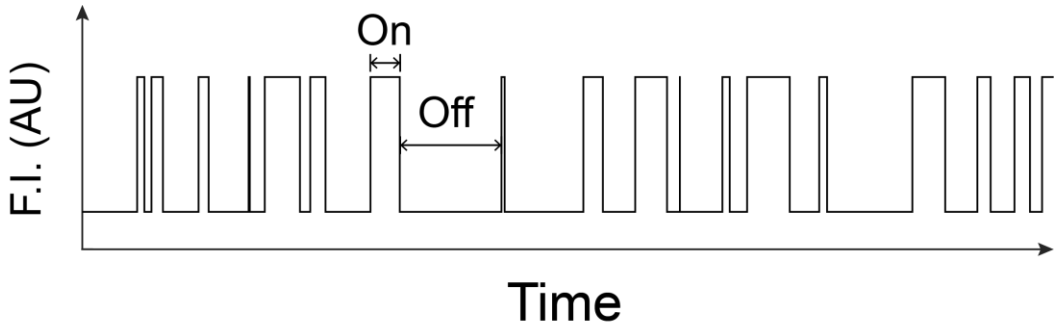

Figure 1.4. Simulation of a two-state continuous-time Poisson process of labeled B interacting tethered A. The fluorescence "on" state represents a binding duration and fluorescence "off" state represents a dissociation duration. AU stands for arbitrary unit.

Figure 1.4 visualizes the fluorescence signal modulation over time due to repeated binding and dissociation of labeled B molecules to the same tethered A molecule. The duration of the "on" state, τ_{on}, and duration of the "off" state, τ_{off}, both follow exponential distributions with distinct lifetimes. The number of transition events, N_{b+d}, follows Poisson distribution. Each binding and dissociation event is a single measurement and the entire fluorescence-time trace consisting of many transition events represents repeated measurements of their interaction kinetics. Therefore, each trace is a unique fingerprinting of the binding kinetics between A and

B. The fluorescence intensity, τ_{on}, τ_{off}, and N_{b+d}, etc. all constitute the single-molecule fluorescence fingerprinting of the analyte's size, shape, and conformation, etc., enabling ultrasensitive and highly specific identification of molecule A.

Mathematically, the statistics of τ_{on} and τ_{off} satisfies the following relationship with ensemble-level reaction rate constants:

$$\langle \tau_{on} \rangle = \frac{1}{k_{off}} \tag{1.2}$$

$$\langle \tau_{off} \rangle = \frac{1}{k'_{on}} \tag{1.3}$$

With,

$$k'_{on} = k_{on}[B] \tag{1.4}$$

Where k_{on} and k_{off} are the association rate constant and the dissociation rate constant, respectively.

Going back to our context of biomarker detection, molecule A is the biomarker molecule of interest and molecule B will be a fluorophore-labeled reporter. In principle, any biomarker molecule can be detected as long as a weak, but specific binding partner is available. An eternal challenge in any analytical assay is separation of target from non-target signals. Next, I will discuss how this SiMREPS principle can be applied to achieve absolute removal of non-target or background signals.

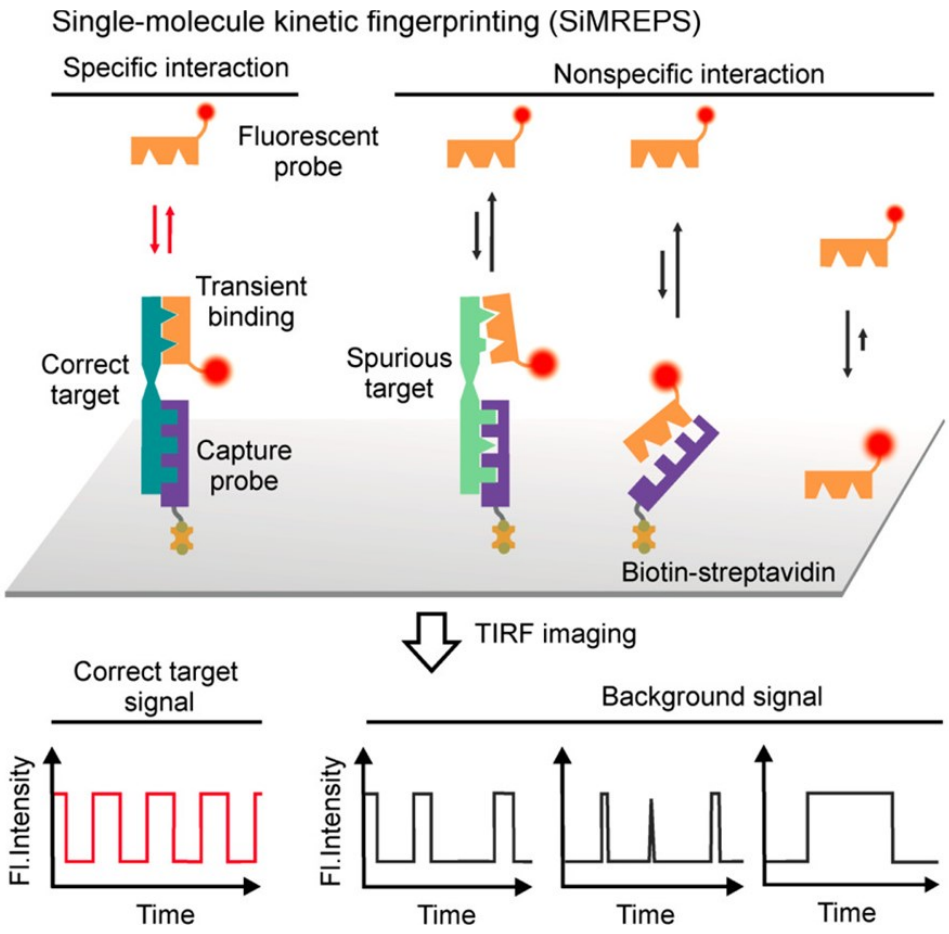

Figure 1.5. Schematic of the principle of single-molecule kinetic fingerprinting (SiMREPS). SiMREPS uses the transient and reversible binding of low-affinity fluorescent probes to immobilized target molecules to generate distinct kinetic fingerprints that permit high-confidence differentiation for specific binding to correct target and nonspecific background binding. Probe binding and dissociation to single molecules are observed in real time microscopy. Figure taken from Mandal and Li et al[79].

As shown in **Figure 1.5**, consider a typical analyte matrix that contains three types of molecules – the biomarker of interest (correct target, e.g., a mutant DNA that drives oncogenesis and is derived from cancer cells), a spurious target sharing some similarity with the correct target (for example, a wildtype DNA derived from normal cells) and background molecules of the same type as the correct target (for example, other genomic DNA fragments). As illustrated in **Figure 1.5**, a surface-tethered capture probe will immobilize both correct target and spurious targets specifically, as well as likely some background molecules nonspecifically. These three

types of molecules will also generate different types of interactions, including specific interactions between the fluorescent probe and the correct target, as well as off-target interactions between the fluorescent probe and spurious targets, capture probes, surface-trapped background molecules or some surface feature itself. Upon illumination, different fluorescence kinetic fingerprints are visualized as distinct on and off patterns, from which kinetic features will be extracted such as fluorescence intensity, τ_{on} or τ_{off}, and N_{b+d}, etc. Those parameters will form a multi-dimensional kinetic space, where each population of signal displays their own distribution.

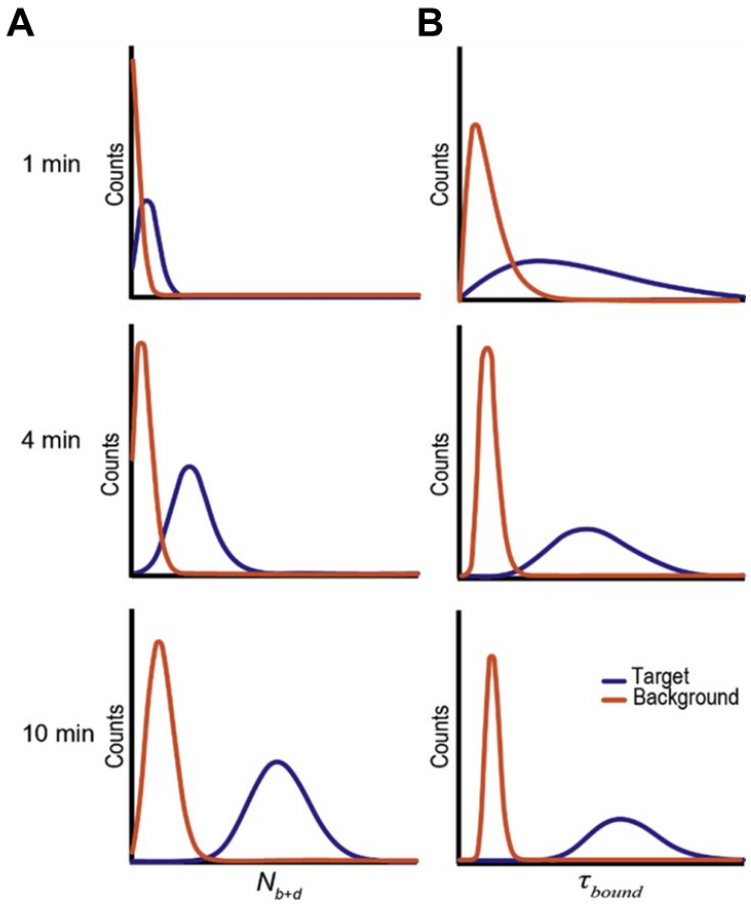

Figure 1.6. Simulated distribution of N_{b+d} and τ_{on} or τ_{bound}. (A) Increasing the acquisition time from 1 min to 10 min yields an increase in separation of the background from the target distribution of N_{b+d}. (B) Increasing the observation time results in observation of more dwell times and a larger shape parameter of the gamma-distributed estimates of bound- or unbound-state lifetime, resulting in a more precise determination of τ_{on} and τ_{off}, and

allowing for more complete separation between target signals and background signals. Figure taken from Chatterjee and Li et al[80].

Figure 1.6 illustrates how statistical characterization of τ_{on} or τ_{off}, and N_{b+d} can completely separate correct target signals from background signals (in this case, background signals refer to signals coming from various off-target interactions) by more repeated measurements in an acquisition window, namely a longer trace. (For mathematical proofs, see Appendix: mathematical framework of kinetic filtering.)

Ultimately, SiMREPS, given a sufficient observation window, can completely filter out both weakly and strongly interacting off-target signals through repeated single-molecule measurement. Apart from τ_{on} or τ_{off}, and N_{b+d}, other features of a trace including signal-to-noise ratio, coefficient of variations and intensity amplitude, etc. can also serve as filtering criteria. Altogether, SMFKF represents a new generation of amplification-free detection approach with ultra-sensitivity and extraordinary specificity that we anticipate will contribute to molecular diagnostics significantly.

1.3.2 General experiment setup

SMFKF is a surface-based detection approach and generally consists of three steps: 1) sample preparation, 2) surface capture, and 3) fluorescence imaging. While sample preparation varies among identities of analyte molecules, surface capture and downstream imaging are consistent among all SMFKF variants[52,70,72–75,81].

As shown in **Figure 1.7**, prior to surface capture, a glass coverslip surface is stringently cleaned, functionalized and passivated with a biotin-doped polyethylene glycol (PEG) surface matrix. The PEG-coated surface is known to effectively reduce non-specific binding of nucleic acids and many proteins. Second, streptavidin or neutravidin is used to coat the surface through

their irreversible interaction with biotin. Following that, a biotinylated capture probe (CP) binds the avidin layer, forming a sandwich structure. Subsequently, samples containing target molecules are incubated with the CPs, which immobilize target molecules specifically and stably through either nucleic acid hybridization or antigen-antibody interaction. The rest of the reagents in the sample matrix are usually washed away prior to fluorescence imaging. The sample chamber where surface capture occurs may as simple as a cut pipette tip glued onto the glass coverslip with Epoxy as shown in **Figure 1.8**A.

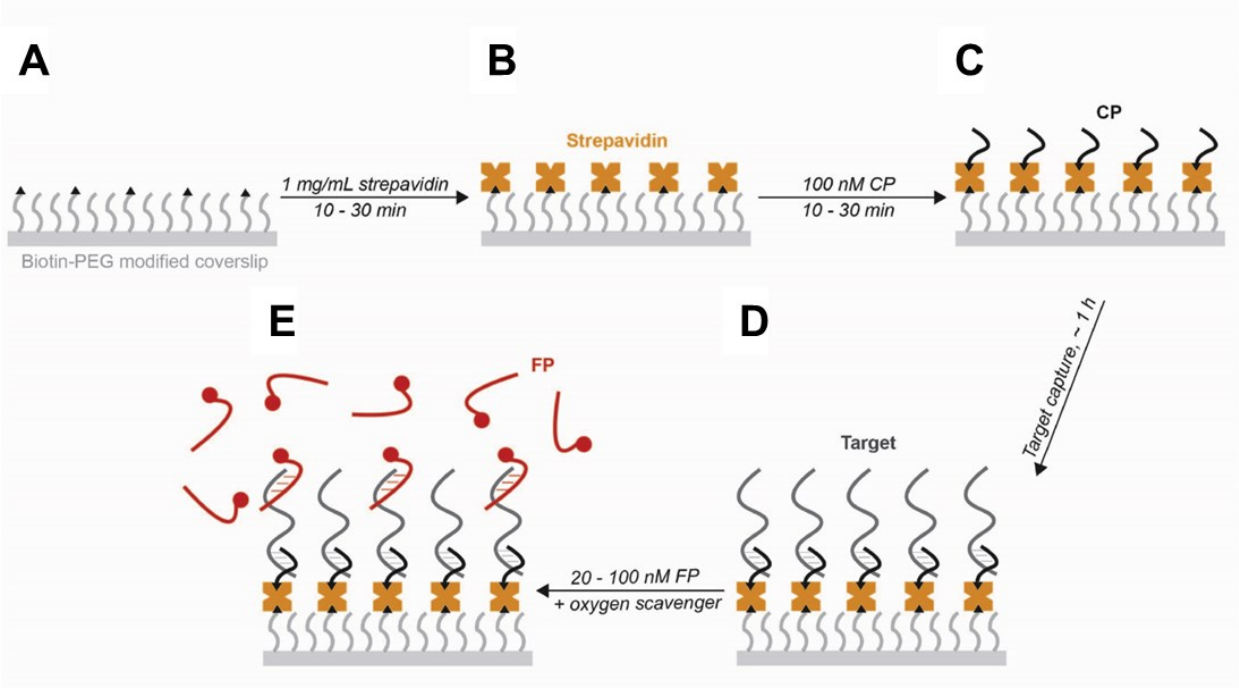

Figure 1.7. A general schematic of surface capture and addition of imaging reporters. (A) A glass coverslip is functionalized with a biotin-containing polyethylene glycol (biotin-PEG). (B) An avidin family protein such as streptavidin or neutravidin is added to provide an anchor point for a biotinylated capture probe (CP) immobilization. (C) CP probes serve as docking sites for analyte of interest. (D) The target solution or sample matrix containing the target molecule is added and incubated to allow sufficient surface capture via nucleic acid hybridization or antigen-antibody interaction. (E) The fluorescent probe (FP) is added in an imaging solution containing an oxygen scavenger system to permit single-molecule kinetic fingerprinting through repetitive binding of the FP to the target. Figure from Chatterjee and Li et al[80].

Following surface capture of molecules of interest, the glass coverslip will be mounted above an objective for TIRF (total internal reflection fluorescence) illumination as shown in

Figure 1.8C. In TIRF, an excitation light continuously illuminates the interface between the glass and aqueous imaging solution at an angle higher than the critical angle, which results in total internal reflection. However, an evanescent field generated by the incident light can penetrate the aqueous phase by hundreds of nanometers when total reflection occurs. The penetration depth, how far this evanescent field can travel into the sample medium, is a function of the relative refractive index, incidence angle and wavelength of excitation light. TIRF has several great benefits for surface-based detection. First, only molecules that are close to the surface within hundreds of nanometers can be illuminated. In other words, we can only observe fluorescence when a probe is bound to the target molecule. Most molecules in the imaging solution are simply "invisible", greatly reducing background signals. Second, the constrained illumination volume minimizes photodamage. In a SMFKF assay, photobleaching of fluorophore attached to the reporter probe is almost never a concern due to an almost "infinite" pool of active fluorophores in the imaging solution. This is especially useful when we want to extend our observation time.

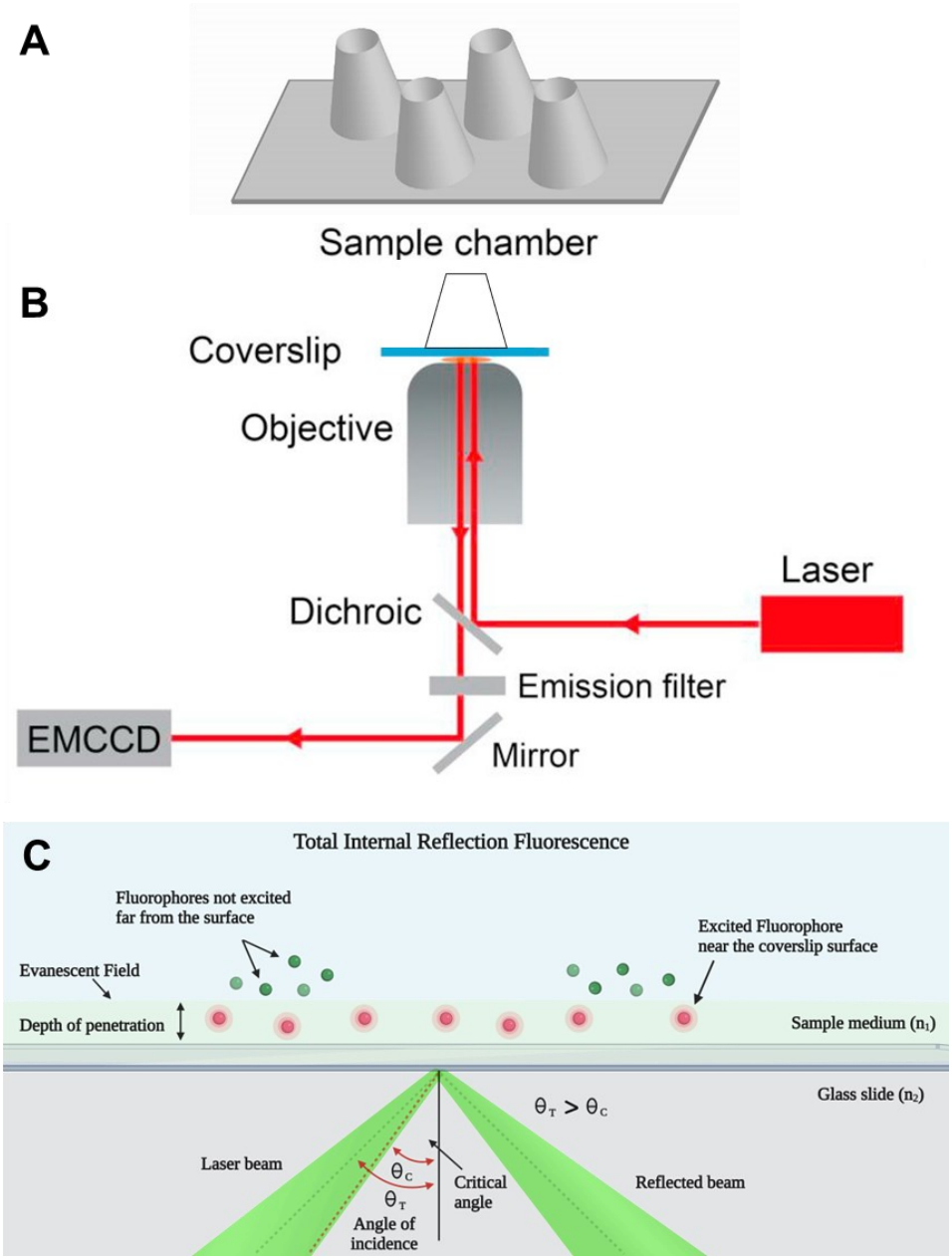

Figure 1.8. Optical and sample setup in fluorescent imaging in SMFKF. (A) Drawing of sample well attached on a glass coverslip. (B) Optical path of an objective-type total internal reflection fluorescence (TIRF) microscope. Figure taken from Johnson-Buck and Li et al[81]. (C) The optical basis of TIRF illumination. Figure taken from Montoya[82]. In TIRF microscopy, the excitation light is reflected on the coverslip/sample interface at the critical angle, θ_c. When the excitation light travels at a high incident angle, θ_τ, which is greater than θ_c, the excitation light is totally reflected from the glass/sample interface and an evanescent field is generated on the opposite side of the interface. The intensity of the evanescent field decreases exponentially with the distance to the interface so only fluorophores close to the surface are significantly excited. Panels were assembled using Biorender.

1.3.3 General data analysis pipeline

Raw data entails consecutive fluorescent images stored in a single movie upon camera acquisition. A movie is a 3D matrix that is basically a time-stack of 2D images. Each image is a 2D matrix where each element stores the raw fluorescence intensity at a specific pixel point of a specific image frame. As shown in **Figure 1.9**, SiMRPS data processing generally consists of these following steps: 1) continuous time-lapse movie collection; 2) identification of local-maximum spots with a certain size, where a "candidate" target molecule resides; 3) after background subtraction, generation of fluorescence intensity-versus-time traces for each candidate spot; 4) kinetic parameters extraction, including τ_{on} or τ_{off}, and N_{b+d} from each idealized trace by hidden Markov modeling using vbFRET[81,83]; 5) acceptance or rejection of candidate molecules based on their kinetic parameters and a set of kinetic filtering criteria. **Figure 1.9** illustrates the data processing pipeline for detection of microRNA miR-16[81]. Finally, a standard curve that plots the number of accepted molecules versus known concentrations is generated and used for quantification and estimation of analytical performance. The limit of detection (LOD) can be derived from the slope and intercept of this standard curve.

The kinetic filtering criteria are optimized using a series of positive and negative signal datasets. Positive datasets contain traces generated from target-only detection, whereas negative datasets contain traces generated from target-free or spurious target detection (for example, a wildtype DNA as opposed to a mutant DNA). A full list of parameters used for kinetic filtering is shown in **Table 1.1**.

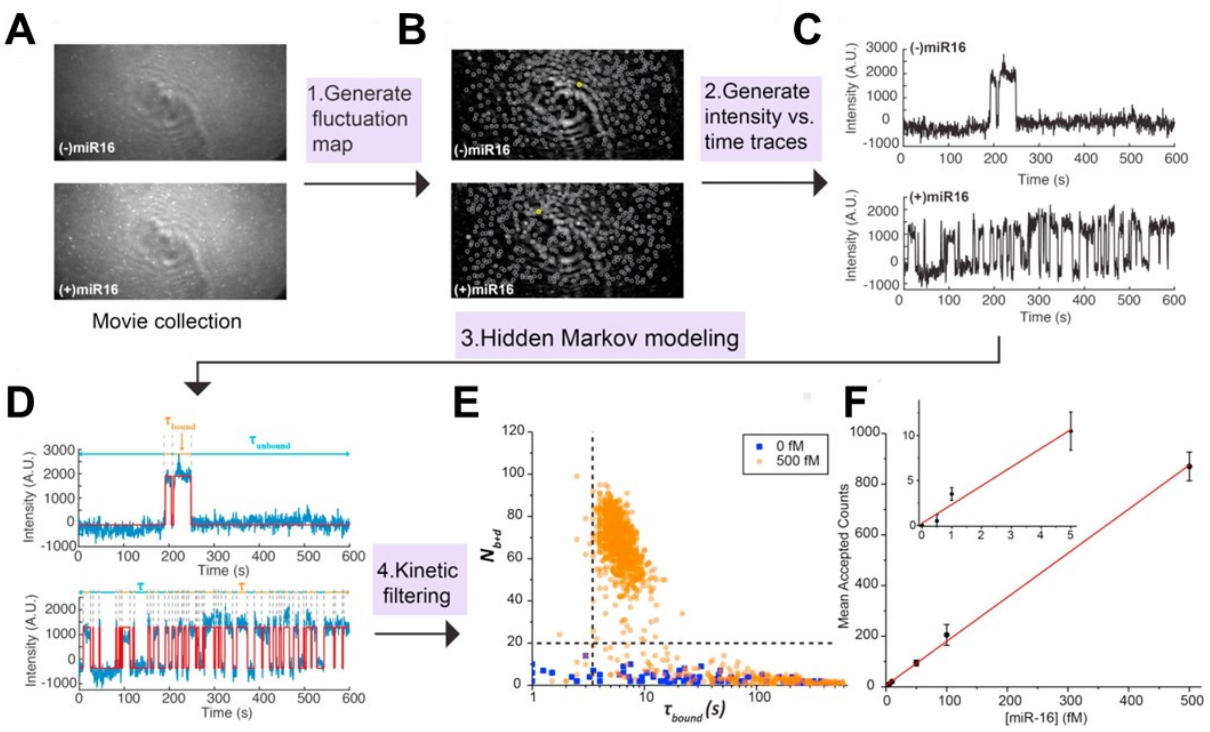

Figure 1.9. Data analysis pipeline. (A) Single-frame images of representative fields of view from TIRF microscopy. (B) Intensity fluctuation maps (calculated by difference in adjacent frames) of the fields of view shown in (A). Grey circles indicate positions of local maxima in the fluctuation map. (C) Representative intensity vs. time traces generated from the spots identified in (B), circled in yellow. (D) HMM idealization (red lines) for each intensity vs. time trace. Bound and unbound-state dwell times (τ_{bound} and $\tau_{unbound}$, respectively) are indicated by the orange and blue horizontal line segments above the idealization. (E) Candidates in the positive (orange circles) and negative (blue squares) controls for miR-16 are well separated by thresholds of $N_{b+d} > 20$ and $\tau_{bound} > 2.5$ s (black dashed lines), permitting discrimination of specific and nonspecific binding at the single-molecule level. Data are pre-filtered for signal-to-noise > 2.5 and intensity > 1000. (F) miR-16 standard curve. n = 3 replicates for blank, 2 replicates for other measurements. Error bars represent 1 standard deviation. Figure taken from Johnson-Buck et al[81].

To find the optimal combination of these parameters, a MATLAB program developed by Alex Johnson-Buck performs automated Monte Carlo-based optimization. A risk function is evaluated for each trial that penalizes for the number of accepted traces in the negative dataset and rewards for those in the positive dataset. In the end, this program outputs a set of filtering criteria that gives the best outcome given the sampled parameter space.

Table 1.1. List of all kinetic filtering parameters and their descriptions.

Kinetic filtering paratemeter	**Description**
Ithresh	Intensity threshold for binding events, used both for determining whether single intensity spikes are to be counted as binding events, as well as for determining whether a trace as a whole will be accepted or rejected.
SNthresh	Signal-to-noise threshold for individual binding events, determining whether or not it is a true binding event
SNthresh_trace	Signal-to-noise threshold for the entire trace to be accepted
min_Nbd	Minimum N_{b+d} value to accept a trace
max_Nbd	Maximum N_{b+d} value to accept a trace
min_tau_on_median (s)	Minimum value of median τ_{on} for a trace to be accepted
min_tau_off_median (s)	Minimum value of median τ_{off} for a trace to be accepted
max_tau_on_median (s)	Maximum value of median τ_{on} for a trace to be accepted
max_tau_off_median (s)	Maximum value of median τ_{off} for a trace to be accepted
max_tau_on_cv	Maximum coefficient of variation (CV) of the τ_{on} to accept a trace
max_tau_off_cv	Maximum coefficient of variation (CV) of the τ_{off} to accept a trace
max_tau_on_event (s)	Maximum value of τ_{on} of an individual event in a trace to be accepted
max_tau_off_event (s)	Maximum value of τ_{off} of an individual event in a trace to be accepted
max_I_low_state	Maximum value of intensity level of the unbound state to accept a trace

1.4 Dissertation outline

This dissertation aims to expand the analytical scope of SMFKF assays to DNA methylation biomarker and break the sensitivity limit by an ultrafast SMFKF detector for DNA mutation biomarkers.

In Chapter 2, I developed a quantitative amplification-free SMKFK assay, termed BSM-SiMREPS, for detecting clusters of DNA methylation in a cancer DNA biomarker, the BCAT1 promoter sequence. BSM-SiMREPS exhibited extraordinary specificity of 99.9999% in most cases and a limit of detection at the sub-femtomolar level, one of the highest sensitivities among all amplification-free approaches described to date. We further demonstrated BSM-SiMREPS measurement of BCAT1 promoter methylation in extracted genomic DNA from whole blood and discovered over 30% methylation, significantly higher than detected by two mainstream PCR-based approaches (whole-genome bisulfite sequencing and methylation EPIC array).

In Chapter 3, I developed MBD-SiMREPS for direct quantification of DNA methylation using a fluorophore-labeled MBD (methyl-binding domain) imager. Methyl-CpG binding activity on both ensemble-level and single-molecule level was observed. Distinguishing methylated and unmethylated DNA was reliably achieved using this MBD imager. I further discovered that a hemimethylated sensor construct was interacting with MBD imager only when a "branch" motif was present at the 5' end of the auxiliary probe. This unique response of MBD imager to architectural changes in the dsDNA substrate was first reported here by SMFKF measurement.

In Chapter 4, I implemented the idea of breaking the limit of detection of SMFKF-based detection by combing total accepted counts of hundreds of FOVs into a single readout. To maintain a similar total detection time, shortening acquisition time for a single FOV was achieved using a fluorogenic imager at micromolar concentrations. In the end, I demonstrated detecting three DNA cancer biomarkers: T790M, L858R and HPV with rationally designed sensor constructs within just 2 seconds.

Finally in Chapter 5, I summarized the conclusions and discussions of this dissertation and present potential future optimizations and applications. A unique but powerful application of MBD imager for in situ whole-genome methylation profiling was put forward and will become one of my major focuses in my postdoc research.

1.5 Appendix: mathematical framework of kinetic filtering

To understand kinetic filtering quantitatively, let us focus on a single parameter, N_{b+d}. As we discussed before, N_{b+d} follows a Poisson distribution and thus our goal is to ensure peak distance between two Poisson populations, N_{b+d}^{target} and $N_{b+d}^{background}$, higher than the sum of their width as follows:

$$\langle N_{b+d}^{target} \rangle - \langle N_{b+d}^{background} \rangle \geq 3\sigma_{b+d}^{target} + 3\sigma_{b+d}^{background} \tag{1.5}$$

Where the peak center position is represented by their mean values, $\langle N_{b+d} \rangle$ and peak width is represented by 3 times the standard deviation, σ_{b+d}. Due to the property of Poisson distribution,

$$\sigma_{b+d} = \sqrt{\langle N_{b+d} \rangle} \tag{1.6}$$

Therefore, the distance between peak centers becomes:

$$\begin{aligned} &\langle N_{b+d}^{target} \rangle - \langle N_{b+d}^{background} \rangle \\ &\quad = \left(\sqrt{\langle N_{b+d}^{target} \rangle} + \sqrt{\langle N_{b+d}^{background} \rangle} \right) \\ &\quad \times \left(\sqrt{\langle N_{b+d}^{target} \rangle} - \sqrt{\langle N_{b+d}^{background} \rangle} \right) \end{aligned} \tag{1.7}$$

This inequality is simplified to:

$$\sqrt{\langle N_{b+d}^{target} \rangle} - \sqrt{\langle N_{b+d}^{background} \rangle} \geq 3 \tag{1.8}$$

The left-hand side of this inequality is essentially the ratio of peak distance and sum of two peak widths. The higher the ratio is, the better the separation between two populations can beachieved. Since there are only two states in our Poisson process, in a single trace, we have two identities that can be used for expressing $\langle N_{b+d} \rangle$ as a function of acquisition time, T:

$$N_b \langle \tau_{on} \rangle + N_d \langle \tau_{off} \rangle = T \tag{1.9}$$

$$N_{b+d} = N_b + N_d \tag{1.10}$$

Where N_b is the number of binding events within a trace and N_d is the number of dissociation events within a trace. Because every binding event always follows a dissociation event (there exist and only exist two states), we have:

$$|N_b - N_d| = 1 \tag{1.11}$$

The difference is often negligible and therefore we can simply get:

$$N_b \approx N_d \tag{1.12}$$

Combining (1.9), (1.10) and (1.12) altogether, in a single trace we have

$$N_{b+d} = 2 \times \frac{T}{\langle \tau_{on} \rangle + \langle \tau_{off} \rangle} \tag{1.13}$$

Finally, across all traces,

$$\langle N_{b+d} \rangle = 2 \times \frac{T}{\langle \tau_{on} \rangle + \langle \tau_{off} \rangle} \tag{1.14}$$

Based on equations (1.2), (1.3) and (1.4), we can express $\langle N_{b+d} \rangle$ as a function of rate constants:

$$\langle N_{b+d} \rangle = 2 \times \frac{T}{\langle \tau_{on} \rangle + \langle \tau_{off} \rangle} = \frac{2T}{\frac{1}{k_{off}} + \frac{1}{k'_{on}}} = 2T \frac{k'_{on} k_{off}}{k'_{on} + k_{off}}$$

$$= 2T \frac{k_{on} k_{off}[B]}{k_{on}[B] + k_{off}} = 2T \frac{k_{on} k_{off}}{k_{on} + k_{off}/[B]} \tag{1.15}$$

From equation (1.15), we can tell that $\langle N_{b+d} \rangle$ is linearly correlated with the acquisition time, T. What essentially discriminates the target and background signal populations is the second kinetic term that includes k_{on} and k_{off}. We can further simplify this expression by introducing equilibrium dissociation constant K_d, which satisfies:

$$K_d = \frac{k_{off}}{k_{on}} \tag{1.16}$$

In the end,

$$\langle N_{b+d} \rangle = 2T \frac{k_{off}}{1 + K_d/[B]} \tag{1.17}$$

In the case where off-target interactions have weaker affinity than target-specific interactions, namely $K_d^{target} < K_d^{background}$, the discrimination power exhibited by N_{b+d} is amplified for background signals with slow dissociation rates. The longer observation we apply, the better separation we can achieve in N_{b+d} distribution until we satisfy the inequality (1.8). However, in the case where fluorescent probe tightly binds some surface features, namely $K_d^{target} \gg K_d^{background}$ and $[B] \gg K_d^{background}$, for background signals we have

$$\langle N_{b+d} \rangle = 2T k_{off} \tag{1.18}$$

This type of background interactions generally exhibited super small k_{off} and in the end $\langle N_{b+d} \rangle$ is only dependent on k_{off} as showed in equation (1.18), resulting $\frac{k_{off}^{target}}{1 + K_d^{target}/[B]} >$

$k_{off}^{background}$. Similarly, a longer observation time T that satisfies the inequality (1.8) can separate correct target signals and background signals. This is an unparalleled advantage over ensemble-level measurements that "statically" remove background signals in blank by simply applying a signal intensity threshold. A fixed signal intensity threshold may remove weakly bound species but is unlikely to remove those tightly bound outliers, where $k_{off}^{background} \ll k_{off}^{target}$.

Another important filtering parameter are the dwell times, τ_{on} and τ_{off}. Individual τ_{on} or τ_{off} is measured for each binding or dissociation event and thus we use the mean or median dwell time to represent this feature collectively for a single trace. For the ease of mathematical deduction, let us choose $\langle\tau_{on}\rangle$ or $\langle\tau_{off}\rangle$ for characterizing an entire trace. $\langle\tau_{on}\rangle$ and $\langle\tau_{off}\rangle$ are described as gamma distributions as shown in **Figure 1.10**[71,80]. However, the gamma distribution only holds true if each trace has the same number of binding events or dissociations events. N_b and N_d follow Poisson distributions just as N_{b+d}. While this assumption in practice is not satisfied, the final distribution is still a compound mixture of many weighted gamma-distributed random variables, each of which has a fixed number of binding events or dissociations events. All of these individual gamma distributions share the same expectation. Although the analytical form of the distribution is not derived yet, we can approximate the final distribution as a single gamma distribution. In fact, as shown in **Figure 1.10**, the distributions of $\langle\tau_{on}\rangle$ and $\langle\tau_{off}\rangle$ match almost perfectly Gamma probabilities. Therefore, we claim that $\langle\tau_{on}\rangle$ and $\langle\tau_{off}\rangle$ both follow Gamma distributions using a shape-and-scale characterization:

$$\langle\tau_{on}\rangle \sim Gamma(\langle N_b\rangle, \frac{1}{k_{off}\langle N_b\rangle}) \tag{1.19}$$

$$\langle \tau_{off} \rangle \sim Gamma(\langle N_d \rangle, \frac{1}{k'_{on}\langle N_d \rangle}) \tag{1.20}$$

With equation (1.12) & (1.18),

$$\langle N_b \rangle = \langle N_d \rangle = \frac{1}{2}\langle N_{b+d} \rangle = Tk_{off} \tag{1.21}$$

By properties of gamma distributions, the standard deviation can be expressed as:

$$\sigma_{\langle \tau_{on} \rangle} = \frac{1}{k_{off}\sqrt{\langle N_b \rangle}} = \frac{1}{k_{off}^{\frac{3}{2}}\sqrt{T}} \tag{1.22}$$

$$\sigma_{\langle \tau_{off} \rangle} = \frac{1}{k'_{on}\sqrt{\langle N_d \rangle}} = \frac{1}{k'_{on}\sqrt{Tk_{off}}} \tag{1.23}$$

Similarly, using $\langle \tau_{on} \rangle$ or $\langle \tau_{off} \rangle$ to distinguish target signal and background signal requires that the distance between two peak centers is higher than the sum of their peak width. In the case of $\langle \tau_{on} \rangle$, we need:

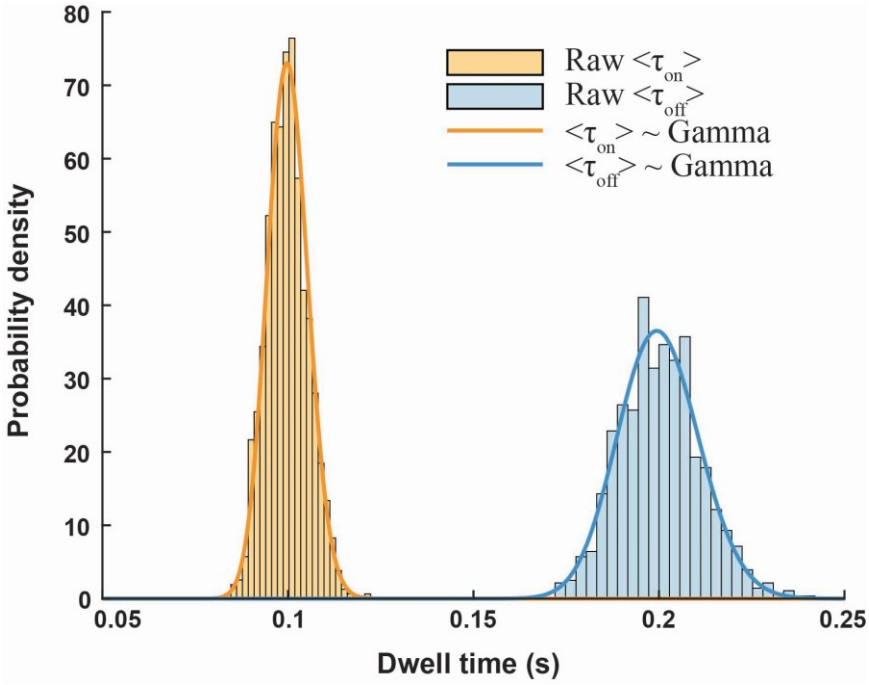

Figure 1.10. Normalized probability density histogram of $\langle \tau_{on} \rangle$ and $\langle \tau_{off} \rangle$ calculated from 1000 simulated traces and probability density curve of gamma distributions generated using $\langle N_{b+d} \rangle/2$ as the shape parameter with scale parameter calculated to keep the mean of $\langle \tau_{on} \rangle$ and $\langle \tau_{off} \rangle$ as $\frac{1}{k_{off}}$ and $\frac{1}{k'_{on}}$ respectively.

$$\left|\frac{1}{k_{off}^{target}}-\frac{1}{k_{off}^{background}}\right| \geq 3\sigma_{\langle\tau_{on}\rangle}^{target}+3\sigma_{\langle\tau_{on}\rangle}^{background}$$

$$\geq \frac{3}{\sqrt{T}}\left(\frac{1}{k_{off,\ target}^{\frac{3}{2}}}+\frac{1}{k_{off,\ background}^{\frac{3}{2}}}\right)$$

$$\geq \frac{3}{\sqrt{T}}\left(\frac{1}{k_{off,\ target}^{\frac{1}{2}}}+\frac{1}{k_{off,\ background}^{\frac{1}{2}}}\right)\left(\frac{1}{k_{off}^{target}}+\frac{1}{k_{off}^{background}}-\frac{1}{\sqrt{k_{off}^{target}k_{off}^{background}}}\right) \quad (1.24)$$

Where the peak center is simply the expectation of a Gamma distribution, and $\frac{1}{k_{off}}$ and peak width are represented by $3\sigma_{\langle\tau_{on}\rangle}$. Dividing both sides by $\left(\frac{1}{k_{off,\ target}^{\frac{1}{2}}}+\frac{1}{k_{off,\ background}^{\frac{1}{2}}}\right)$, we have:

$$\left|\frac{1}{k_{off,\ target}^{\frac{1}{2}}}-\frac{1}{k_{off,\ background}^{\frac{1}{2}}}\right|$$

$$\geq \frac{3}{\sqrt{T}}\left(\frac{1}{k_{off}^{target}}+\frac{1}{k_{off}^{background}}-\frac{1}{\sqrt{k_{off}^{target}k_{off}^{background}}}\right) \quad (1.25)$$

Therefore, extending acquisition time favors a larger extend of separation in $\langle \tau_{on} \rangle$ distribution between target and background. In the scenario where $k_{off}^{background} \gg k_{off}^{target}$ or, alternatively, $\sqrt{k_{off}^{background}} = f\sqrt{k_{off}^{target}}$ & $f \gg 1$, equation (1.25) can be simplified to:

$$\sqrt{Tk_{off}^{target}} \geq 3\left(1 + \frac{1}{f^2 - f}\right) \tag{1.26}$$

$$\sqrt{\langle N_d \rangle} \geq 3\left(1 + \frac{1}{f^2 - f}\right) \geq 3 \tag{1.27}$$

$$\langle N_{b+d} \rangle \geq 18 \tag{1.28}$$

Equation (1.28) suggests that for any off-target molecules from which the probe dissociates very fast, as long as we extend our observation window such that more than 18 transitions on average can be captured within a trace, we can easily distinguish them with high confidence. This principle is of great significance when discriminating against SNVs, where a single mismatch between probe and wildtype target increases k_{off} by more than an order of magnitude, although the thermodynamic difference may be small. This feature is also useful when filtering out weakly interactions between probes and surface matrix including the surface substrate itself, capture probe or surface-captured background molecules, due to non-specific hydrophobic or electrostatic interactions.

Chapter 2 Single Molecule Counting of Bisulfite Converted Methyl CpG

2.1 Introduction

DNA methylation refers to the addition of a 5-methyl group to cytosine in deoxyribonucleic acids. Its discovery as a structural component of natural nucleic acids dates back to 1925 when it was found in Tubercle bacillus species[26]. Over the past thirty years, researchers have linked DNA methylation to heritable transcriptional repression in vertebrates. DNA methylation-mediated gene silencing plays a crucial role in various biological processes, such as mammalian development, X chromosome inactivation, genomic imprinting, and genome stability, etc.[27,28,30,31,84]. Notably, aberrant methylation profiles have been implicated in numerous diseases, particularly cancer, where dysregulation of DNA methylation commonly plays a role in tumorigenesis[34].

DNA methylation is currently being explored as a promising biomarker for early detection of cancer due to several compelling reasons. First, unlike somatic tumor mutation profiles that show significant variation between patients, tumor methylation profiles are highly consistent across individuals[34,85,86]. Moreover, epigenetic alterations tend to occur early during oncogenesis[87–89], making them attractive markers for detecting small cancers at an early stage before they spread. Additionally, DNA methylation status at specific loci are often tissue-specific and therefore carry implicit information about the tissue of origin of a cancer, especially when measuring tumor-derived DNA in the blood as a cancer detection marker[90]. A recent clinical study showed a strong correlation between the methylation level in the BCAT1 and IKZF1

promoter regions in colorectal cancer patients with tumor progression; a growing number of clinical studies of methylated DNA loci as cancer biomarkers are currently underway[91–93].

Ever since its discovery, the detection of DNA methylation has been essential for studies of its regulatory mechanisms. Thin layer chromatography was first used to determine its molecular formula and amount found in genomic DNAs[26,94,95]. Bisulfite treatment followed by polymerase chain reaction (PCR) — that is, methylation-specific PCR (MSP) where sodium bisulfite specifically deaminates unmethylated cytosines to uracils, leaving methylated cytosines intact — revolutionized the methylation field by enabling sequence-specific methylation detection of any CpG site[96]. Apart from locus specific methylation quantification, development in high throughput sequencing and microarray detection enables methylation profiling of an entire gene and even the entire genome ("methylome")[97–100]. These amplification-based approaches generally have high sensitivity, but their specificity suffers from target-independent amplification — a source of non-specific signal[48,101]. Because the methylation level is calculated as a ratio of signals of amplified methylated and unmethylated sequences, unbiased treatment in each step, including amplification, is essential for accurate measurement. Furthermore, bisulfite treatment causes DNA fragmentation and damage (e.g., through depurination and depyrimidation) that compromises primer binding efficiency and accuracy, as well as replication fidelity of polymerases. Despite these important technical concerns, bisulfite sequencing has been considered the "gold standard" in methylation detection for decades.

Recently, our group developed an amplification-free single molecule kinetic fingerprinting technique — single-molecule recognition through equilibrium Poisson sampling (SiMREPS) that measures the Poisson statistics of weak but specific probes binding to individual target molecules through real-time observation of transient, repeated interactions. Applying

kinetic filtering of each individual time trace not only eliminates background signals almost completely, but also discriminates signals of spurious targets, for example, single nucleotide variants (SNVs). Background-free assays with around 99.9999% specificity were demonstrated in the detection of miRNAs, somatic mutant DNAs, and proteins[52,70,72,73]. This advancement holds great promise for the accurate and reliable detection of DNA methylation and other biomarkers in various applications.

2.2 Materials and methods

Table 2.1. Lists of DNA strands, their code names, sequences and descriptions.

Code name	Sequences	Description
BCAT1 Forward	GTCTTCCTGCTGATGCAATCCGCTAGGTCGC GAGTCTCCGCCGCGAGAGGGCCGGTCTGCAA TCCAGCCCGCCACGTGTACTCGCCGCCGCCT CGGGCACTG	Full-length BCAT1 promoter forward strand, directly purchased from IDT
BCAT1 Reverse	CAGTGCCCGAGGCGGCGGCGAGTACACGTGG CGGGCTGGATTGCAGACCGGCCCTCTCGCGG CGGAGACTCGCGACCTAGCGGATTGCATCAG CAGGAAGAC	Full-length BCAT1 promoter reverse strand, directly purchased from IDT
dsBCAT1	NA	Prepared by mixing equal amount of BCAT1 Forward and Reverse in PBS buffer
Me-BCAT1 Forward	GTCTTCCTGCTGATGCAATC/iMe-dC/GCTAGGT/iMe-dC/G/iMe-dC/GAGTCTC/iMe-dC/GC/iMe-dC/G/iMe-dC/GAGAGGGC/iMe-dC/GGTCTGCAATCCAGCC/iMe-dC/GCCA/iMe-dC/GTGTACT/iMe-dC/GC/iMe-dC/GC/iMe-dC/GCCT/iMe-dC/GGGCACTG	Full-length methylated BCAT1 promoter forward strand, directly purchased from IDT
Me-BCAT1 Reverse	CAGTGCC/iMe-dC/GAGG/iMe-dC/GG/iMe-dC/GG/iMe-dC/GAGTACA/iMe-dC/GTGG/iMe-	Full-length methylated BCAT1 promoter reverse

	dC/GGGCTGGATTGCAGAC/iMe-dC/GGCCCTCT/iMe-dC/G/iMe-dC/GG/iMe-dC/GGAGACT/iMe-dC/G/iMe-dC/GACCTAG/iMe-dC/GGATTGCATCAGCAGGAAGAC	strand, directly purchased from IDT
dsMe-BCAT1	NA	Prepared by mixing equal amount of Me-BCAT1 Forward and Reverse in PBS buffer
10Me-BCAT1 Forward	GTCTTCCTGCTGATGCAATCCGCTAGGTCGCGAGTCTC/iMe-dC/GC/iMe-dC/G/iMe-dC/GAGAGGGC/iMe-dC/GGTCTGCAATCCAGCC/iMe-dC/GCCA/iMe-dC/GTGTACT/iMe-dC/GC/iMe-dC/GC/iMe-dC/GCCT/iMe-dC/GGGCACTG	Full-length methylated BCAT1 promoter (except the first three CpGs) forward strand, directly purchased from IDT
	Real target	
102 nt MBC	GTUTTUUTGUTGATGUAATU/iMe-dC/GUTAGGT/iMe-dC/G/iMe-dC/GAGTUTU/iMe-dC/GU/iMe-dC/G/iMe-dC/GAGAGGGU/iMe-dC/GGTUTGUAATUUAGUU/iMe-dC/GUUA/iMe-dC/GTGTAUT/iMe-dC/GU/iMe-dC/GU/iMe-dC/GUUT/iMe-dC/GGGUAUTG	The forward strand sequence of product after treating dsMe-BCAT1 with Methylation-Lightning™ Kit, since in principle only the forward sequence will be captured. However, the product will always contain the reverse strand as well.
102 nt UBC	GTUTTUUTGUTGATGUAATUUGUTAGGTUGUGAGTUTUUGUUGUGAGAGGGUUGGTUTGUAATUUAGUUUGUUAUGTGTAUTUGUUGUUGUUTUGGGUAUTG	The forward strand sequence of product after treating dsBCAT1 with Methylation-Lightning™ Kit, since in principle only the forward sequence will be captured. However, the product will always contain the reverse strand as well.
102 nt rMBC	UAGTGUU/iMe-dC/GAGG/iMe-dC/GG/iMe-dC/GG/iMe-dC/GAGTAUA/iMe-dC/GTGG/iMe-dC/GGGUTGGATTGUAGAU/iMe-dC/GGUUUTUT/iMe-dC/G/iMe-dC/GG/iMe-dC/GGAGAUT/iMe-	Sequence of product after treating Me-BCAT1 Reverse with Methylation-Lightning™ Kit

	dC/G/iMe-dC/GAUUTAG/iMe-dC/GGATTGUATUAGUAGGAAGAU	
102 nt fMBC	GTUTTUUTGUTGATGUAATU/iMe-dC/GUTAGGT/iMe-dC/G/iMe-dC/GAGTUTU/iMe-dC/GU/iMe-dC/G/iMe-dC/GAGAGGGU/iMe-dC/GGTUTGUAATUUAGUU/iMe-dC/GUUA/iMe-dC/GTGTAUT/iMe-dC/GU/iMe-dC/GU/iMe-dC/GUUT/iMe-dC/GGGUAUTG	Sequence of product after treating Me-BCAT1 Forward with Methylation-Lightning™ Kit
	Mimic	
102 nt MBC Mimic	GTUTTUUTGUTGATGUAATUCGUTAGGTCGCGAGTUTUCGUCGCGAGAGGGUCGGTUTGUAATUUAGUUCGUUACGTGTAUTCGUCGUCGUUTCGGGUAUTG	The mimic for 102 nt MBC without methylation sites, directly purchased from IDT
102 nt UBC Mimic	GTUTTUUTGUTGATGUAATUUGUTAGGTUGUGAGTUTUUGUUGUGAGAGGGUUGGTUTGUAATUUAGUUUGUUAUGTGTAUTUGUUGUUGUUTUGGGUAUTG	The mimic for 102 nt UBC without methylation sites, directly purchased from IDT
55 nt MBC Mimic	GTUTTUUTGUTGATGUAATUCGUTAGGTCGCGAGTUTUCGUCGCGAGAGGGUCGG	The mimic for 55 nt MBC without methylation sites, directly purchased from IDT
55 nt UBC Mimic	GTUTTUUTGUTGATGUAATUUGUTAGGTUGUGAGTUTUUGUUGUGAGAGGGUUGG	The mimic for 55 nt UBC without methylation sites, directly purchased from IDT
42 nt MBC Mimic	TGUAATUCGUTAGGTCGCGAGTUTUCGUCGCGAGAGGGUCGG	The mimic for 42 nt MBC without methylation sites, directly purchased from IDT
42 nt UBC Mimic	TGUAATUUGUTAGGTUGUGAGTUTUUGUUGUGAGAGGGUUGG	The mimic for 42 nt UBC without methylation sites, directly purchased from IDT
	Sensor construct	
Aux1	CTTATCTGTTTTCGCGACCTAACGAATT	Auxiliary probe 1, providing docking site for FP1, 1b and 1c

Aux2	CGACCCTCTCGCGACGAtttATAGCATGTTT	Auxiliary probe 2, providing docking site for FP2 and 2c
Biotin-Aux2	CGACCCTCTCGCGACGATTTATAGCATGTTT /3bioTEG/	Biotinylated auxiliary probe 1, providing docking site for FP2 and also serving as a capture probe in Figure 2.
CP1	/5Biosg/ATAATTAATAACATCAACAAAAA AAC	Capture probe 1, used in Figure 2 and Figure 3.
CP2	/5Biosg/TAATTAATACACGTAACGAACTA AATTACA	Capture probe 2, used in Figure 4 and 5.
Block1	TCACAACCTAACAAATTACA	Blocker 1, a short strand complementary to UBC or UBC Mimic (any sizes) at the Aux1-binding region.
Block2	CAACCCTCTCACAACAA	Blocker 2, a short strand complementary to UBC or UBC Mimic (any sizes) at the Aux2-binding region.
	Imager	
FP1	/5Cy3/CAGATAAG	Fluorescent probe 1, interacting with Aux1 with 8-nt binding region
FP1b	/5Cy3/ACAGATAAG	Fluorescent probe 1b, interacting with Aux1 with 9-nt binding region
FP1c	/5Cy3/TTACAGATAAG	Fluorescent probe 1c, interacting with Aux1 with 9-nt binding region
FP2	CATGCTAT/3Cy5Sp/	Fluorescent probe 2, interacting with Aux2, with 8-nt binding region
FP2c	CATGCTATTTT/3Cy5Sp/	Fluorescent probe 1, interacting with Aux1, with 8-nt binding region

Genomic DNA		
WGA	NA	(Directly purchased from Zymo Research, cat. no. D5013-1) Human HCT116 DKO Non-methylated DNA, generated using phi29 DNA polymerase based whole genome amplification techniques from HCT116 DKO cell line derived genomic DNA
+Me WGA	NA	(Directly purchased from Zymo Research, cat. no. D5013-2) Human WGA Methylated DNA, generated using human WGA Non-methylated DNA that has been enzymatically methylated at all double-stranded CG dinucleotides using M.SssI methyltransferase.
Blood DNA	NA	(Directly purchased from Enzo Lifesciences, ENZ-GEN117-0100) Human Genomic DNA, Male, comprised of a pool of ≥ 8 normal donors. It is intact DNA extracted from freshly harvested whole blood of healthy males. DNA is treated with DNase free-RNase to remove residual contaminant RNA.
BS WGA	NA	Bisulfite conversion product of WGA
BS +Me WGA	NA	Bisulfite conversion product of +Me WGA
BS Blood DNA	NA	Bisulfite conversion product of Blood DNA

2.2.1 Assay pipeline and working principle of BSM-SiMREPS

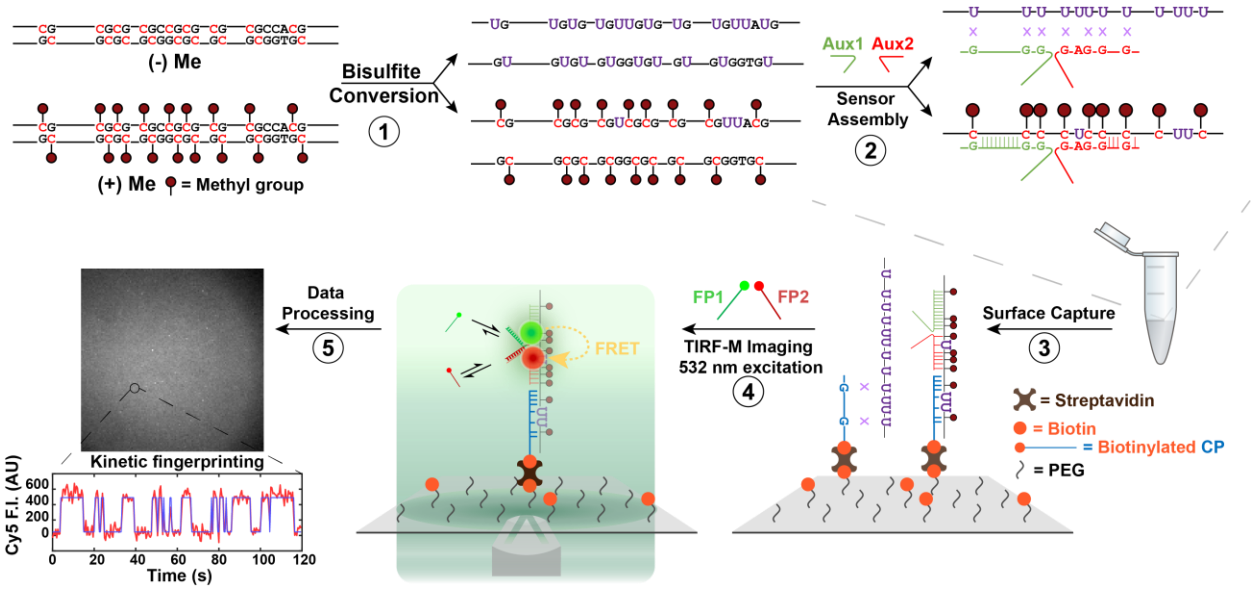

Figure 2.1. Schematic of BMS-SiMREPS pipeline. (I) Full length BCAT1 promoter (methylated or unmethylated) of 102 nt undergoes bisulfite treatment where first of all, originally fully complementary double stranded target becomes single stranded due to 44 mismatches introduced. Secondly, unmethylated cytosine is converted to uracils and methylation cytosine stays intact. (II) The product of bisulfite treatment is mixed with auxiliary probes (Aux1 and Aux2) which serves as docking sites for fluorescent probes and only methylated forward strand is bound. (III) After sensor assembly, the mixture is incubated with a capture probe coated streptavidin-pegylated coverslip. (IV) And finally, an imaging buffer containing two fluorescent oligos is added to the sample well and the surface is excited by green laser under TIRF illumination. Cy5 fluorescence is collected as readout. (V) During data processing, individual fluorescent spots are identified and characterized for intensity-time traces which are fitted by hidden Markov model to identify "on" and "off" states that will be used to separate target signals from background signals. AU stands for arbitrary unit; F.I. stands for fluorescence intensity.

Here, we applied the principle of SiMREPS to methylation detection by coupling it with bisulfite treatment, termed bisulfite Me-SiMREPS (BSM-SiMREPS). We chose 102 nt BCAT1 promoter, a previously reported DNA methylation biomarker for colorectal cancer to demonstrate BSM-SiMREPS[85,91]. The assay starts with bisulfite treatment, which converts cytosines to uracils in the target sequences shown in **Figure 2.1**. Throughout the entire target sequence there are 13 methylation sites, occurring exclusively at cytosines of CpG dinucleotides

(only 9 methylation sites are shown in **Figure 2.1**). Bisulfite treatment not only deaminates cytosines but also introduces mismatches between original fully complementary target strands. Each unmethylated cytosine introduces one mismatch resulting in 44 GU mismatches in bisulfite-converted methylated BCAT1 promoter (102 nt MBC) and 70 GU mismatches in bisulfite-converted unmethylated BCAT1 promoter (102 nt UBC). This guarantees single-stranded form of the DNA products at room temperature. Next, two auxiliary probes (Aux1 and Aux2) are mixed with single-stranded DNAs; they only stably bind bisulfite converted methylated BCAT1 promoter (102 nt MBC). The presence of 2 μM dT10, together with blocker1 and blocker 2 (see **Table 2.1**) was used to prevent weak binding of auxiliary probes to bisulfite converted unmethylated BCAT1 promoter (102 nt UBC). The sensor mixture is then added to sample wells, where a biotinylated capture probe (CP) is precoated on a streptavidin-coated PEG surface through biotin-streptavidin interaction. Similarly, CPs specifically recognize 102 nt MBC, whereas the weak interaction of the capture probe with 102 nt UBC is suppressed by dT10 addition. An imaging buffer containing a pair of FPs labeled by Cy3 and Cy5, respectively is added and the sample well is imaged under TIRF (total internal reflection fluorescence) illumination using an oil immersion objective. Transient interactions of the FP pair with the two docking sites on the auxiliary probes generate kinetic fingerprints upon Cy3 excitation through FRET. The Cy5 emission signal is collected for downstream data analysis.

Background signals are significantly reduced to nearly zero in BSM-SiMREPS by three mechanisms: 1) ensemble fluorescence originating from FPs in solution is minimized by the constrained excitation volume under total internal reflection conditions; 2) fluorescence resulting from diffusion of FPs to the surface upon excitation is minimized by FRET, as the probability that both FPs locate together within FRET distance on the surface is rather low; 3) a small

fraction of background signals with intensity fluctuations distinct from the above signals present as a Poisson process but is rejected by kinetic filtering based on the unique fingerprints of FPs to the target complex. Optimizing FP sequences and assay conditions enables a shorter acquisition time per field of view (FOV) and higher sensitivity by collecting multiple FOVs.

2.2.2 Oligonucleotides

All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT, www.idtdna.com) with standard desalting purification, unless otherwise noted. All fluorescent probes (FPs) with either 5' Cy3 or 3' Cy5 modifications were purchased from IDT with high-performance liquid chromatography (HPLC) purification. The full-length promoter sequence of branched-chain amino acid transaminase 1, BCAT1 was chosen as our detection target — its genomic coordinates were Chr12: 24,949,058 - 24,949,159 (genome build: UCSC Genome browser GRCh38/hg38 version)[85,91]. All different lengths of methylation mimics had the same sequences as bisulfite-converted methylated BCAT1 promoter with corresponding sizes but missed methylation modification; all different lengths of non-methylation mimics had the same sequences as bisulfite-converted unmethylated BCAT1 promoter with corresponding sizes. All mimics were purchased directly from IDT and used as model targets for assay testing and optimizations. All full-length targets or strands of 102 nt were purchased from IDT with ultramer oligo synthesis and standard desalting purification. See **Table 2.1** for descriptions of each target and their acronyms.

2.2.3 Bisulfite treatment

Bisulfite treatment followed the manufacturer's protocol (EZ DNA Methylation-Lightning™ Kit, Zymo Research). Double-stranded DNA substrates were prepared by annealing

complementary single-stranded oligonucleotides at around 1 μM final concentration in 4X PBS (Phosphate-buffered saline) (40 mM Na_2HPO4, 7.2 mM KH_2PO_4, pH 7.4, 548 mM NaCl, 10.8 mM KCl), heating at 90°C for 5 min, cooling to 37°C for 5 min, and finally holding at room temperature for 10 min before storage at − 20°C for further use. PBS buffers were either diluted from 10X PBS stock solution or directly purchased from Fisher Scientific (Invitrogen™ UltraPure™ DNase/RNase-Free Distilled Water or Gibco™ PBS, pH 7.4, or Gibco™ PBS (10X), pH 7.4). Briefly speaking, DNA substrates once mixed with Lightning Conversion Reagent were first heated at 98°C for 8 min, annealed to 54°C for 1 h, and finally held at 4°C before desulphonation. Desulphonation lasted 15 min before column purification. The eluted product was about 20 μl, and its concentration was determined by Qubit (Qubit™ ssDNA Assay Kit, ThermoFisher) or Nanodrop (NanoDrop 2000, Thermofisher) from 3 independent measurements.

For bisulfite conversion of Blood DNA, WGA, and +Me WGA, each 150 μl reaction only takes 600 ng genomic DNAs at most to ensure optimal conversion efficiency and specificity, and following elution from purification column, elutes from all individual reactions were combined and concentrated using a vacuum centrifuge. This process was termed parallel bisulfite conversion. All quantifications of genomic DNAs were using parallel bisulfite conversion. Overloading of genomic DNAs has been observed to cause significant false positives from incomplete conversion (data not shown here).

2.2.4 Design of BSM-SiMREPS probes

Auxiliary probes and capture probes are designed to stably bind target sequences. There are three considerations in their sequence design: 1) melting temperatures (T_m) of their hybridizations with target should be 10°C above the imaging temperature or capture temperature;

2) ideally upon hybridization with target, binding regions of two neighboring strands should be at least 3 nt away on the target sequence to avoid steric hindrance; 3) biotinylation preferably occurs at 5' end of capture probes and thus only full-length oligo product after synthesis can stay on the surface. T_m (melting temperature) between auxiliary probes or capture probes and target was estimated by the NUPACK Web application (http://www.nupack.org/partition/new) using the following parameters: number of strand species = 2, maximum complex size = 2, computer melt = checked, minimum temperature = 10°C, increment = 3°C, maximum temperature = 70°C, target DNA concentration = 1 pM, probe concentration = 1 pM, Na+= 646 mM. A melting curve is first generated by Nupack and then we take the first derivative of fractions of unpaired bases. The peak position of the first derivative of melting curve indicates Tm of probe binding at this specific condition (see **Figure 2.2**).

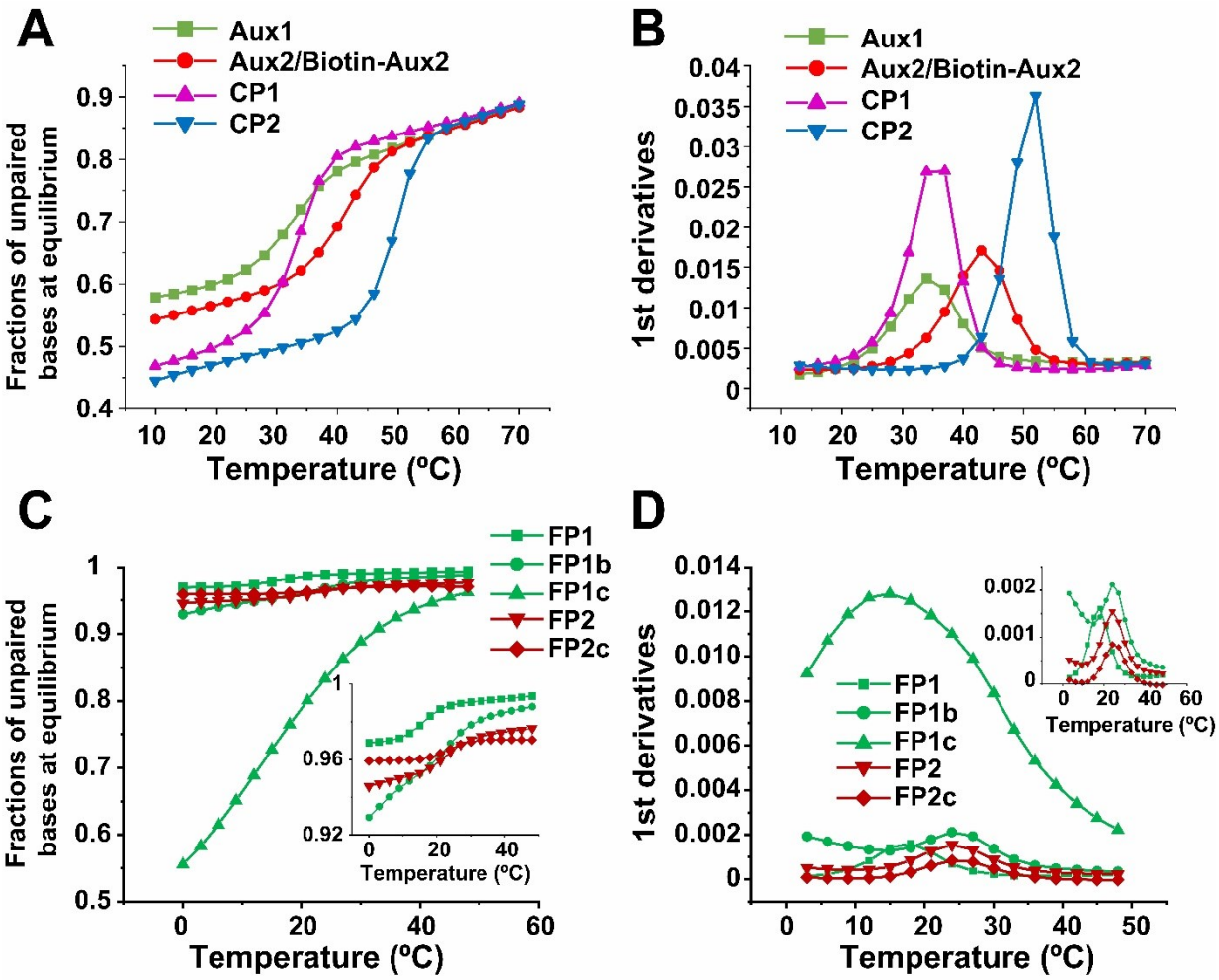

Figure 2.2. Calculations of T_m by melting curves. (A) (B) are melting curves and first derivative of unpaired base fraction with respect to temperature respectively for bindings of capture probes and auxiliary probes to 102 nt MBC Mimic. (C) (D) are melting curves and first derivative of unpaired base fraction with respect to temperature respectively for bindings of different imagers to auxiliary probes. All data are generated by Nupack prediction and replotted.

The design of FPs was assisted by Nupack (https://www.nupack.org/). Their sequences were much less restricted than those of auxiliary probes and capture probes since they were target-independent, but there were several considerations to obtain optimal FPs' sequences: 1) no secondary structure should be observed by Nupack prediction; 2) no self-hybridization should be observed by Nupack prediction; 3) ideally, no additional binding region that is longer than 4 nt should be observed on any probes' or target's exposed sequences after construct assembly to

avoid unwanted background signals and kinetic fingerprinting. Melting temperature of FPs' hybridizations to docking sites on auxiliary probes should be close to room temperature estimated by Nupack in a similar way described above, except that a few parameters are adjusted as follows: minimum temperature = 0°C, maximum temperature = 50°C, auxiliary probe concentration = 1 nM, fluorescent probe concentration = 100 nM (see **Figure 2.2**).

2.2.5 BSM-SiMREPS assay

Sample cells made of cut P20 pipette barrier tips were attached to glass coverslips passivated with a 1:10 mixture of biotin-PEG and mPEG. A detailed protocol of slide preparations is discussed in previously published papers[81]. Sample cells were first washed with T50 buffer (10 mM Tris-HCl [pH 8.0 at 25°C], 50 mM NaCl) and then incubated with 40 µl 0.25 mg/ml streptavidin in T50 buffer for 10 min. Following a wash with 1X PBS for 3 times, 100 nM capture probe in 1X PBS that was preheated at 90°C for 5 min in a metal bath, annealed at 37°C for 5 min in a water bath, and cooled down to room temperature, was then added to the sample well. The sample well was incubated for 10 min and washed with 4X PBS for 3 times waiting for the target strand. A mixture of sensor components was prepared by adding either bisulfite-converted methylated or unmethylated double-stranded BCAT1 promoter or single-stranded mimics to a PCR tube that contained 10 nM Aux1, Aux2, blocker 1, and blocker 2 in 4X PBS / 2 µM poly-T oligodeoxyribonucleotide (dT10) carriers. All dilutions of targets were performed in the presence of 2 µM dT10 in GeneMate low-adhesion 1.7 mL microcentrifuge tubes (VWR, Cat No. 490003-230). PCR tubes that contained sensor components including targets are then heated at 73°C for 3 min, annealed at 46.6°C for 5 min, were then annealed at 40°C for another 5 min and finally cooled down to 25°C. This sensor assembly process was performed in a thermocycler. The sensor construct that was properly assembled was added to the sample cell

and then incubated for 1 h at room temperature. After target capture, sample cells were washed 3 times with 4X PBS and 100 µl imaging buffer containing the desired concentration of FPs in the presence of an oxygen scavenger system (OSS) — 1 mM Trolox, 5 mM 3,4-dihydroxybenzoate, 50 nM protocatechuate dioxygenase — was added and then imaged by objective-TIRF microscopy.

2.2.6 Single-molecule fluorescence microscopy

Initial optimizations on sensor design, sequences of FP pairs, and imaging temperature (**Figure 2.3**, **Figure 2.4**B&C) were performed using an Olympus IX-81 objective-type TIRF (O-TIRF) microscope with a 60X oil-immersion objective (APON 60XOTIRF, 1.45 NA) equipped with both a cell^TIRF^{TM} and a z-drift control module (Olympus IX2-ZDC2). An EMCCD (electron-multiplying charge-coupled device) camera (Andor IXon 897, EM gain 150) was used to record the movies. For recording Cy5 emission by FRET with optimal signal-to-noise ratio (S/N), FP pairs were excited by a 532 nm at a power of 15 mW (OBIS 637 nm LX, 100 mW) after passing through a dichroic mirror (ZT405/488/532/640rpc, Chroma), and an emission filter (ET705/100m, Chroma) and the TIRF angle was adjusted to achieve a calculated evanescent field penetration depth of 80 nm. The signal integration time (exposure time) per frame was 500 ms unless otherwise noted, movies of 2-10 min were collected per field of view (FOV). If needed, an objective heater (BIOPTECHS) was used to raise the imaging temperature after calibration with an infrared thermometer (Lasergrip 800, Etekcity). Quantification experiments in genomic DNAs are also conducted at this microscope (**Figure 2.11**).

For further optimizations on FP concentrations as well as calibrations (**Figure 2.4**D&E, **Figure 2.9**), the same O-TIRF microscope was equipped with both cell^TIRF^{TM} and Z-drift control modules (ASI CRISP). An EMCCD camera (Evolve 512, Photometrics) was used to

record movies. FP pairs were excited by a 532 nm at power of 15 mW after after passing through a Cy3-A647 FRET dichroic mirror (ZT40DRC-UF2, Chroma) and an emission filter (ET655LP-TRF filter, Chroma). If needed, the same objective heater was used to raise the temperature following the same calibration procedure.

Consecutive Multiple-FOV detection was achieved by a journal programmed in Metamorph[74]. A total of 10 FOVs were collected for all quantification experiments (**Figure 2.9** and **Figure 2.11**).

2.2.7 Processing and analysis of objective-type TIRF data

A set of custom MATLAB codes were used to identify spots with significant intensity fluctuation within each FOV, generate intensity-versus-time traces at each spot, fit these traces with two-state hidden Markov modeling (HMM) algorithm to generate idealized traces, and eventually identify and characterize transitions with idealized traces. A set of filtering criteria were generated to distinguish target-specific signal and non-specific signal by feeding traces from no target control experiments and unmethylated target-only experiments as negative dataset and traces from methylated target-only experiments as positive dataset into a SiMREPS optimizer (see **Table 2.2**). A detailed discussion of data analysis pipeline can be found in papers previously published in our group[81].

Table 2.2. Optimized parameter sets for trace generation and analysis.

Trace Generation Parameters	
use fluctuation map?	2
Stdfactor	15
start frame	1
end frame	240

edgePx	20
Percentilecut	0.95
ROI size (pixels)	3

Trace Analysis Parameters (KFC)	
start frame	1
end frame	240
exposure time (s)	0.5
Smoothframes	1
remove_single_frame_events	FALSE
Ithresh	151.6923572
SNthresh	2
SNthresh_trace	2.5
min_Nbd	12
max_Nbd	65
min_tau_on_median (s)	0.5
min_tau_off_median (s)	0.5
max_tau_on_median (s)	7
max_tau_off_median (s)	35
max_tau_on_cv	1.8
max_tau_off_cv	Inf
max_tau_on_event (s)	8
max_tau_off_event (s)	80
max_I_low_state	194000
vary_I_vals	FALSE
num_intensity_states	2
ignore_post_bleaching	FALSE
bleaching_wait_time (s)	Inf
use_FRET_threshold	FALSE
FRET_threshold	0

2.2.8 Compilation of blood DNA methylation using recountmethylation

Recountmethylation is a R/Bioconductor package with 12537 uniformly processed EPIC and HM450K blood samples on GEO[102,103]. All data on GEO until December, 2022 measured illumina EPIC array and illumina HM450K array are compiled and made available on its public data server (https://recount.bio/data/). To extract all DNA methylation beta values on 102 nt

BCAT1 promoter measured on illumina EPIC array, a custom-written R code was used to extract and analyze them from the recountmethylation data server (See https://github.com/dai905/My_Recountmethylation/blob/ed415fac70776f33e1d57e098989a278f8e216f1/Blood_EPIC_v1.Rmd). For specific file url on the online server, see https://recount.bio/data/remethdb_h5se-gm_epic_0-0-2_1589820348/ and https://recount.bio/data/remethdb_epic-hm850k_h5se_gm_1669220613_0-0-3/).

2.3 Results

2.3.1 Initial optimization of sensor design and FP pair sequences

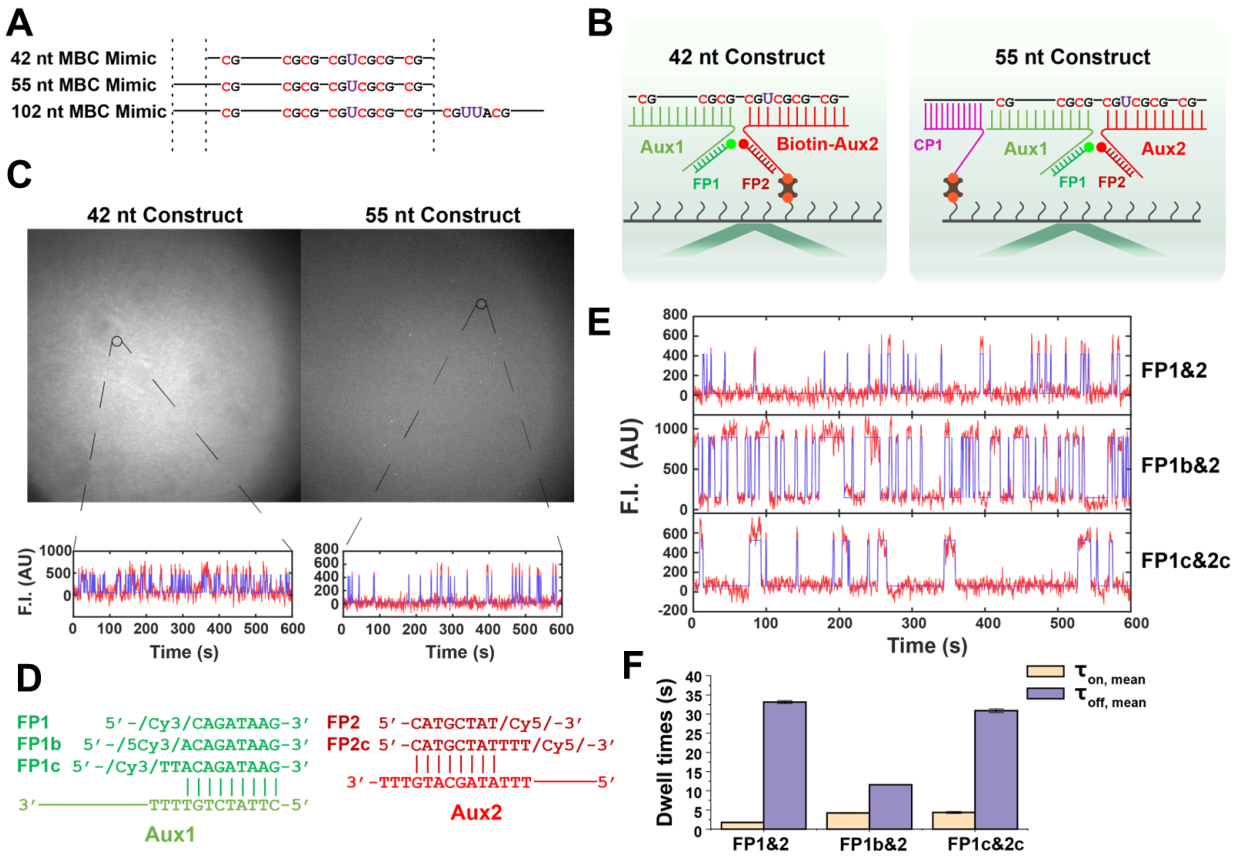

Figure 2.3. Optimization of sensor constructs and imager sequences. (A) Sequences of different lengths of MBC Mimic and their alignment result. (B) 42 nt construct is the sensor design for detecting 42 nt MBC Mimic and Aux2 serves both as an auxiliary probe and a biotinylated capture probe. 55 nt construct is the sensor design for detecting 55 nt MBC Mimic and the main difference is incorporating a biotinylated CP1 additionally. (C) Screenshot of raw movies using different constructs and their associated intensity-time traces (red lines) fitted by HMM (blue lines).

(D) Sequences of different imagers and their binding sites on Aux1 or Aux2 respectively. (E) Representative intensity-time traces (red lines) fitted by HMM (blue lines) for different imager pairs. (F) Comparison of mean τ_{on} and τ_{off} of each pair of imagers. Mean τ_{on} and τ_{off} are calculated by fitting dwell times of individual events in all traces with exponential decay. Error bars represent fitting errors.

Before we tested with 102 nt BCAT1 promoter, we generated a 42 nt MBC Mimic (see **Table 2.1** and **Figure 2.3**) as a model target for our initial optimization. Instead of using three probes as in our later sensor construct (**Figure 2.3**B), a biotinylated auxiliary probe 2 (Biotin-Aux2, see **Table 2.1** and **Figure 2.3**B) served both as a capture probe and docking site for FP2, whereas auxiliary probe 1 (Aux1) provided a docking site for FP1 (**Figure 2.3**B). However, the result was undesirable. Blinking spots were hard to resolve due to high background in the raw image. Kinetic fingerprinting also showed poor signal-to-noise ratio (S/N) (**Figure 2.3**C) and it was unclear whether the signal fluctuation was the result of FP binding and dissociation or simply due to fluctuation of the background level. Based on the sensor construct design, one reasonable source of high background was the interaction of FP2 with Biotin-Aux2 regardless of the presence of target. The surface was saturated by Biotin-Aux2 and thus FP2 could interact almost anywhere on the surface, generating a high background of Cy5 fluorescence even upon Cy3 excitation. Therefore, we designed a new construct by introducing a third probe — a separate biotinylated capture probe 1 (CP1) for detecting 55 nt MBC Mimic (see **Table 2.1**, **Figure 2.3**B). In the new design, FP2 should not interact with CP1 directly. Instead, the presence of target was necessary to bring both Aux1 and Aux2 together, thus making it possible that FP1 and FP2 interact with their docking sites in close proximity. Upon Cy3 excitation, this allowed Cy5 fluorescence emission via FRET. The raw image showed spatially distinct bright spots that were well separated from the signal in surrounding pixels, and their kinetic fingerprints demonstrated a good S/N (**Figure 2.3**C).

Following this design principle, different FP pairs were tested to obtain optimal dwell times (the time spent in either the high-intensity state — yielding bound time, τ_{on}; or in the low-intensity state — yielding unbound time, τ_{off}). **Figure 2.3**D shows different FP pair sequences and their shared docking sites on Aux1 and Aux2. FP1b and FP2 showed the similar τ_{on} and τ_{off} (**Figure 2.3**F), which were most favorable since equivalent dwell times generate the most transitions given a certain acquisition time, providing a better separation from background signal[73,104] It also facilitated downstream optimization of the kinetics by allowing shorter acquisition times (**Figure 2.3**D-F). Comparing dwell times of FP1&2 and FP1b&2, the only difference that the binding region of FP1b had was that one more nucleotide than FP1 changed the kinetics dramatically — both extending τ_{on} and shortening τ_{off}; FP1c&2c were initially designed to bring the two fluorophores closer by extending the length of overhangs to which the two fluorophores were attached. However, we did not observe an increase in S/N. In fact, when one FP first binds the docking site of auxiliary probes, its additional thymines might extend the fluorophore too far due to the presence of carbon linker attached to the fluorophore moiety and prevent the other FP from approaching their docking sites. Due to biased dwell times with FP1c&2c, FP1b&2 remained the best candidate.

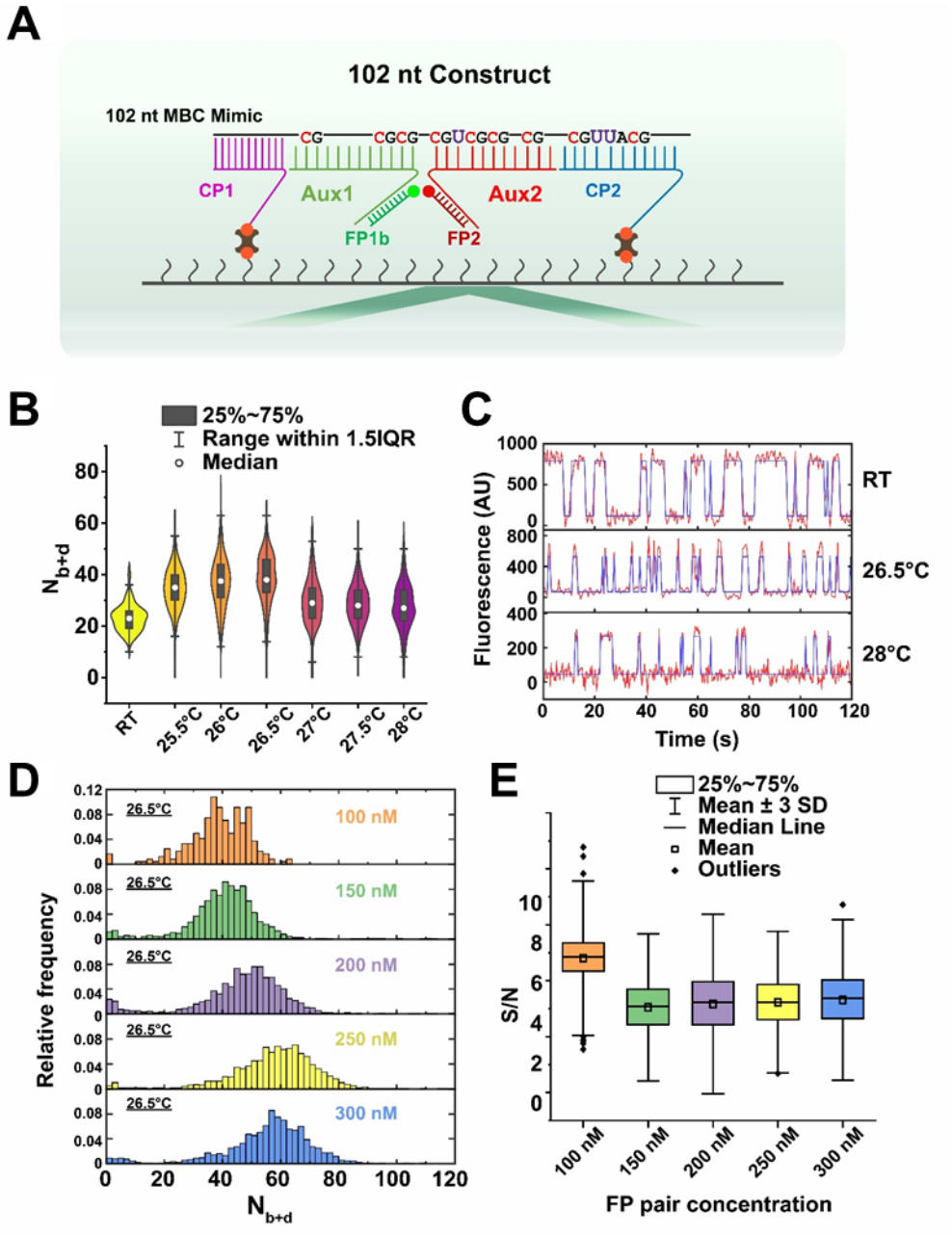

Figure 2.4. Optimization of imaging conditions. (A) Sensor construct used in optimizations. 102 nt construct is designed for detecting 102 nt MBC Mimic with 2 biotinylated capture probes, CP1 and CP2. (B) Optimizations of imaging temperature. Effects of imaging temperatures on N_{b+d} distribution are shown in violin plots. (C) representative intensity-time traces (red lines) fitted by HMM (blue lines) at different temperatures. Panel B-C are using 100 nM FP1b&2. (D) Optimizations of imager concentrations. Effects of imager concentrations on N_{b+d} distributions are shown in histograms. € Effects of imager concentrations on signal-to-noise ratio (S/N) using the same dataset as in panel D.

2.3.2 Further optimization of imaging temperature and FP concentration

In order to achieve high sensitivity, the detection of multiple FOVs within a reasonable amount of time is desirable (**Figure 2.9**A). Sequential data acquisition of multiple FOVs enabled treating total molecule counts as a single readout, thus "amplifying" the signal approximately by the number of FOVs collected, provided that background was not also proportionately amplified. To achieve multiple FOVs without drastically extending the imaging time, rapid kinetic fingerprinting was achieved by optimizing the imaging temperature and FP concentrations with the 102 nt MBC Mimic using both capture probes CP1&CP2 (**Figure 2.4**). In SiMREPS, the number of binding and dissociation (N_{b+d}) is a good measure of the kinetics given a certain observation window. We first optimized imaging temperatures from room temperature (RT) up to 28°C (**Figure 2.4**B&C). Theoretically speaking, heat should shorten both τ_{on} and τ_{off} by accelerating diffusion/disrupting self-structure and disrupting the bound state, respectively. From RT to 26.5°C, N_{b+d} increased with the temperature through a shorter τ_{on} since heat disrupts binding of FPs to their docking sites (**Figure 2.4**B&C). However, after passing 26.5°C, N_{b+d} started to decline due to a longer τ_{off}, which could be explained by partial dissociation of auxiliary probes at high temperatures since the Tm of Aux1 is only 33°C (**Figure 2.4**B&C). In other words, there was a certain probability that auxiliary probes partially dissociated from target even when both FPs were binding their docking sites — a complication that contributed to τ_{off} at high temperature. Another aspect of optimization is the FP concentration, which mainly impacts the association rate constant — or reciprocal of τ_{off}. **Figure 2.4**D suggested an upshift in the N_{b+d} distribution as the FP concentrations increased. However, we decided not to go for concentrations higher than 100 nM since the S/N declined significantly once the concentrations surpassed 100 nM (**Figure 2.4**E). In order to detect as many molecules as at lower FP concentrations, the S/N threshold had to be relaxed and more false positives were be accepted in

no-target control experiments (**Figure 2.5**), which ultimately defeated the purpose of increasing sensitivity.

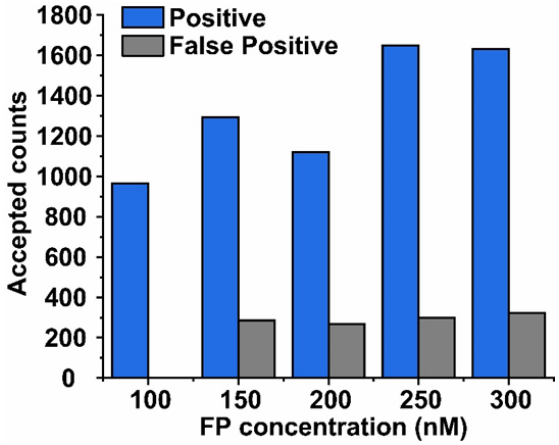

Figure 2.5. Specificity comparison among different concentrations of FP pair, FP1b+2. At each concentration, a particular kinetic filtering is applied to both positive dataset and negative dataset to generate positive counts and false positive counts respectively. Positive dataset consists of traces for detecting 1 pM 102 nt MBC Mimic at corresponding FP concentration. Negative dataset consists of three sets of traces: 1) no target control; 2) detection of 1 nM 102 nt UBC Mimic; 3) detection of 5 nM 102 nt UBC Mimic. This particular kinetic filtering criterium is generated separately by SiMREPS optimizer for each concentration in order to maximize positive counts and minimize false positive counts by feeding the same two datasets. Data from a single FOV is collected at 26.5°C and analyzed for each condition. Although a marginal increase in positive counts observed across concentrations higher than 100 nM, a significant amount of false positives also inevitably occur. Note that over 200 false positives come from just a single FOV. In other words, a marginal increase in signal amplitude cannot compensate for the significant decline in specificity.

Apart from the imaging conditions, sensitivity also depends on capture efficiency, a function of capture probe design and capture strategy. In fact, in addition to CP1, capture probe 2 (CP2) was designed to bind target at its 5' end (**Figure 2.4**A). Theoretically, both capture probes, when mixed, could suppress effects of DNA damage on capture efficiency caused by bisulfite treatment since target is captured as long as either of the two binding sites on the target is sufficiently undamaged. By contrast, primer-dependent PCR amplification would not work then. Another advantage of CP2 is that its binding region spreads across two methylation sites for another layer of specificity. The double capture was therefore utilized in all optimizations of the imaging conditions (**Figure 2.4**). However, quantification suggested no significant improvement

of sensitivity or specificity upon double capture (**Figure 2.6**), whereas CP1 occasionally generated inconsistent counts (**Figure 2.6**D). Consequently, we decided to use single capture by CP2 for our final design, as well as for the following characterization.

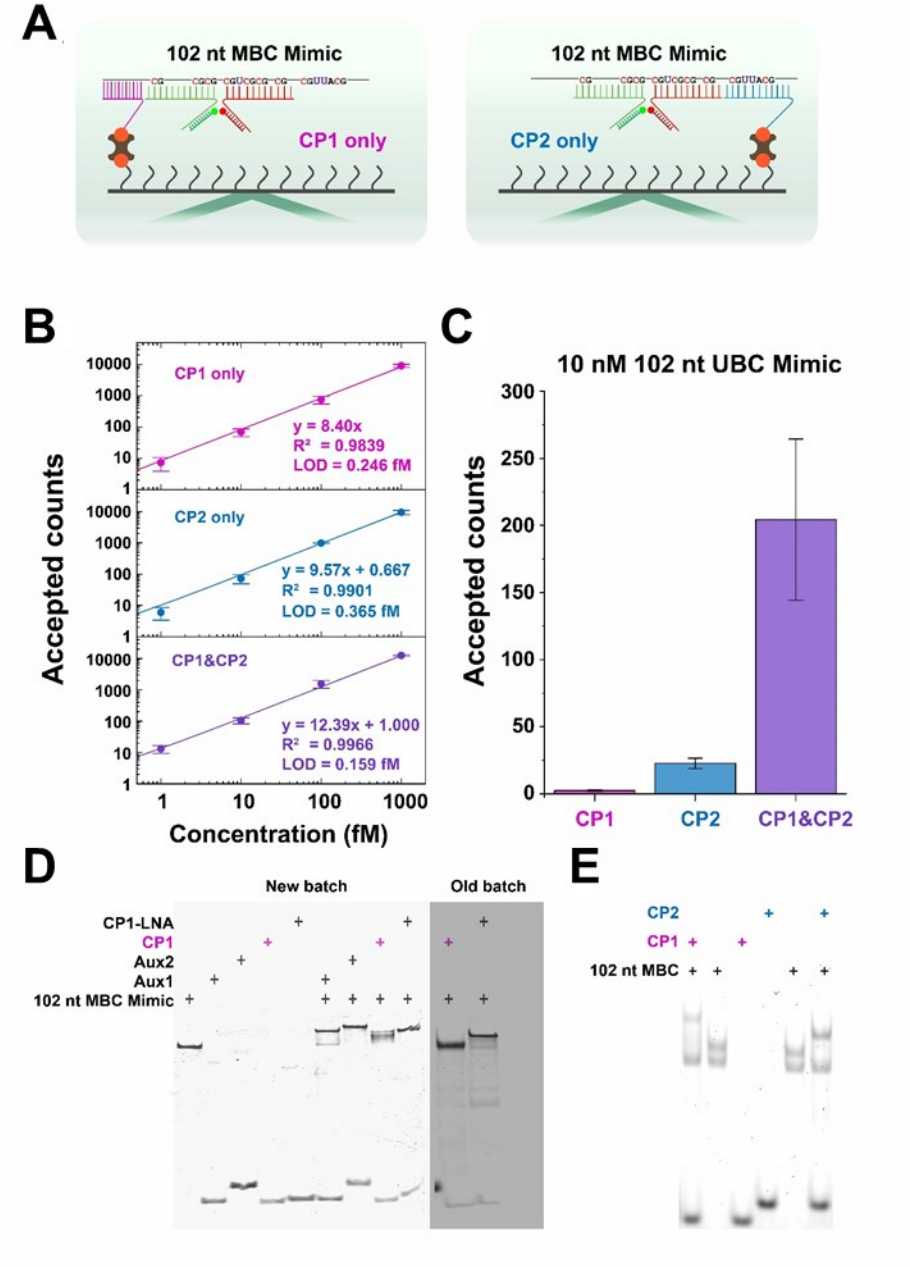

Figure 2.6. Analytical performances using different capturing approaches. (A) Sensor constructs using individual capture probes. Imaging conditions using the same as in **Figure 2.7**. (B) Standard curves used for detecting 102 nt MBC Mimic using different capturing approaches. (C) False positive counts using different capturing approaches. (D) 12% native PAGE to show degradation of CP1 under long-term storage. Each strand of new batch was loaded

with the same amounts as of old batch. (E) 5% native PAGE to show binding of CP1 and CP2 to 102 nt MBC. Both gels in Panel D&E are stained by SYBR-Au and visualized using Cy2 fluorescence

2.3.3 Side-by-side comparison of assay performance with MBC and MBC mimic.

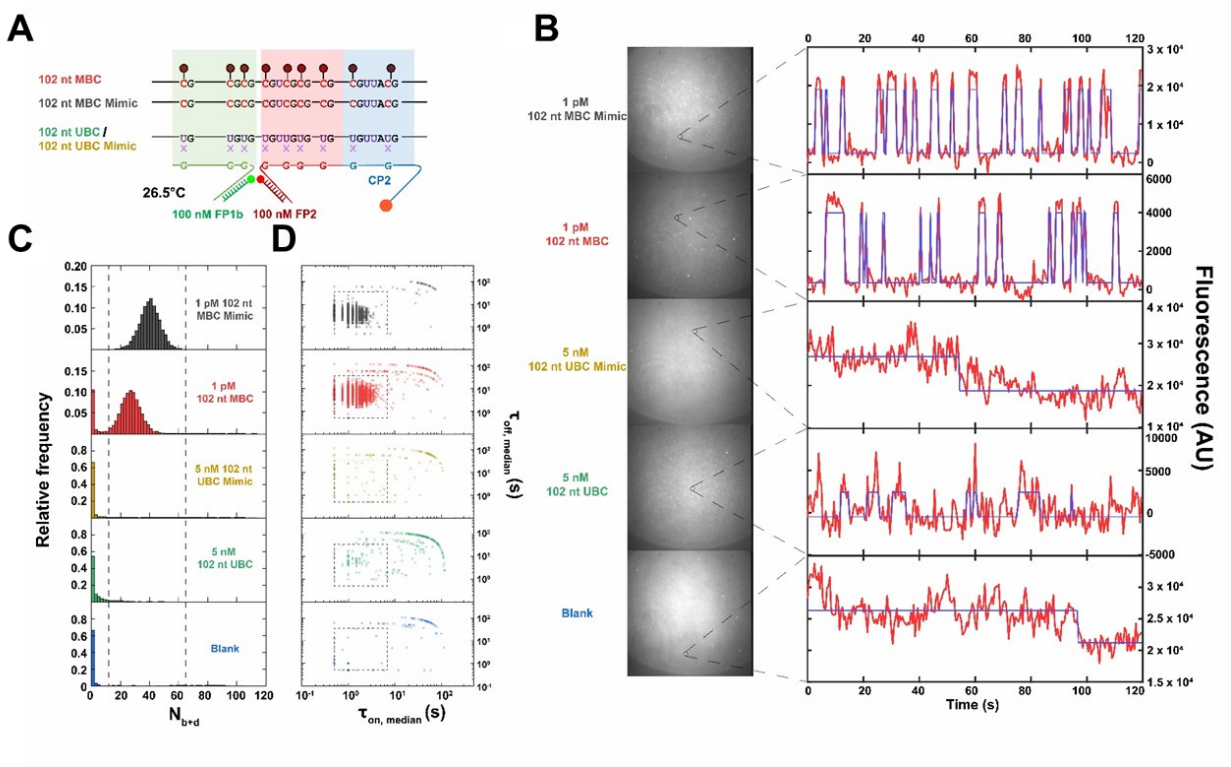

Figure 2.7. Detection of different types of samples using BSM-SiMREPS. (A) Sequence differences and methylation differences of different types of samples. 102 nt MBC Mimic shares the same sequence as 102 nt MBC but without methyl group and without bisulfite-converted methylated reverse strand. 102 nt UBC Mimic shares the same sequence as 102 nt UBC but without bisulfite-converted unmethylated reverse strand. Shaded regions are binding sites for each auxiliary probe and capture probes. All other panels use the same capturing and imaging conditions as in texts. (B) Screenshot of raw movies of different types of samples and their associated representative intensity-time traces (red lines) fitted by HMM (blue lines). (C) Distribution of N_{b+d} for different types of samples. Dashed lines represent threshold to distinguish MBC or MBC Mimic against UBC or UBC Mimic or blank. (D) Distribution of median τ_{on} andτ_{off} of each trace for different types of samples. Dash lines represent threshold to distinguish MBC or MBC Mimic against UBC or UBC Mimic or blank.

The ultimate sensor construct included Aux1, Aux2 and CP2 to immobilize and discriminate methylated and unmethylated species with downstream imaging at 26.5°C, 100 nM FP1 and 100 nM FP2 (**Figure 2.7**). The design schematic based on the 102 nt MBC Mimic or 102 nt MBC is shown in **Figure 2.7**A. To test the robustness of our assay with both mimic and

real target (see **Table 2.1**), we conducted three control experiments to examine their performance with 1 pM 102 nt MBC Mimic or MBC as methylation positive control, 5 nM 102 nt UBC Mimic or UBC as methylation negative control and no target as blank control (**Figure 2.7**B-D). **Figure 2.7**B indicates a clear distinction in kinetic fingerprints among the three control experiments for both mimic and real target. **Figure 2.7**C illustrates a clear separation of target-specific and non-target signals in their N_{b+d} distribution for both mimic and real target. **Figure 2.7**D further highlights the distinct populations representing the MBC and MBC Mimic signals, which were well separated from UBC, UBC Mimic and blank. Notably, the N_{b+d} distribution in the case of MBC shifted to the left compared to MBC Mimic due to longer τ_{off} (**Figure 2.7**C and **Figure 2.8**A). It is plausible that DNA damage such as depurination or depyrimidation after bisulfite treatment disrupted the sensor assembly in the former case. That is, we suspect that there was a certain probability that the auxiliary probes partially dissociated from the target even when both FPs are binding their docking sites — a disassembly that would extend τ_{off}. **Figure 2.8**B also showed a clear decrease in the "on" state fluorescence intensity, suggesting a decrease in FRET efficiency in the case of detecting MBC, further supporting this mechanistic explanation. To further test our hypothesis, we examined the sensor assembly of both MBC and MBC Mimic by native PAGE, but not observe dissociation of Aux1 or Aux2 in **Figure 2.8**C, perhaps because it does not extend to a gel with its caging effect.

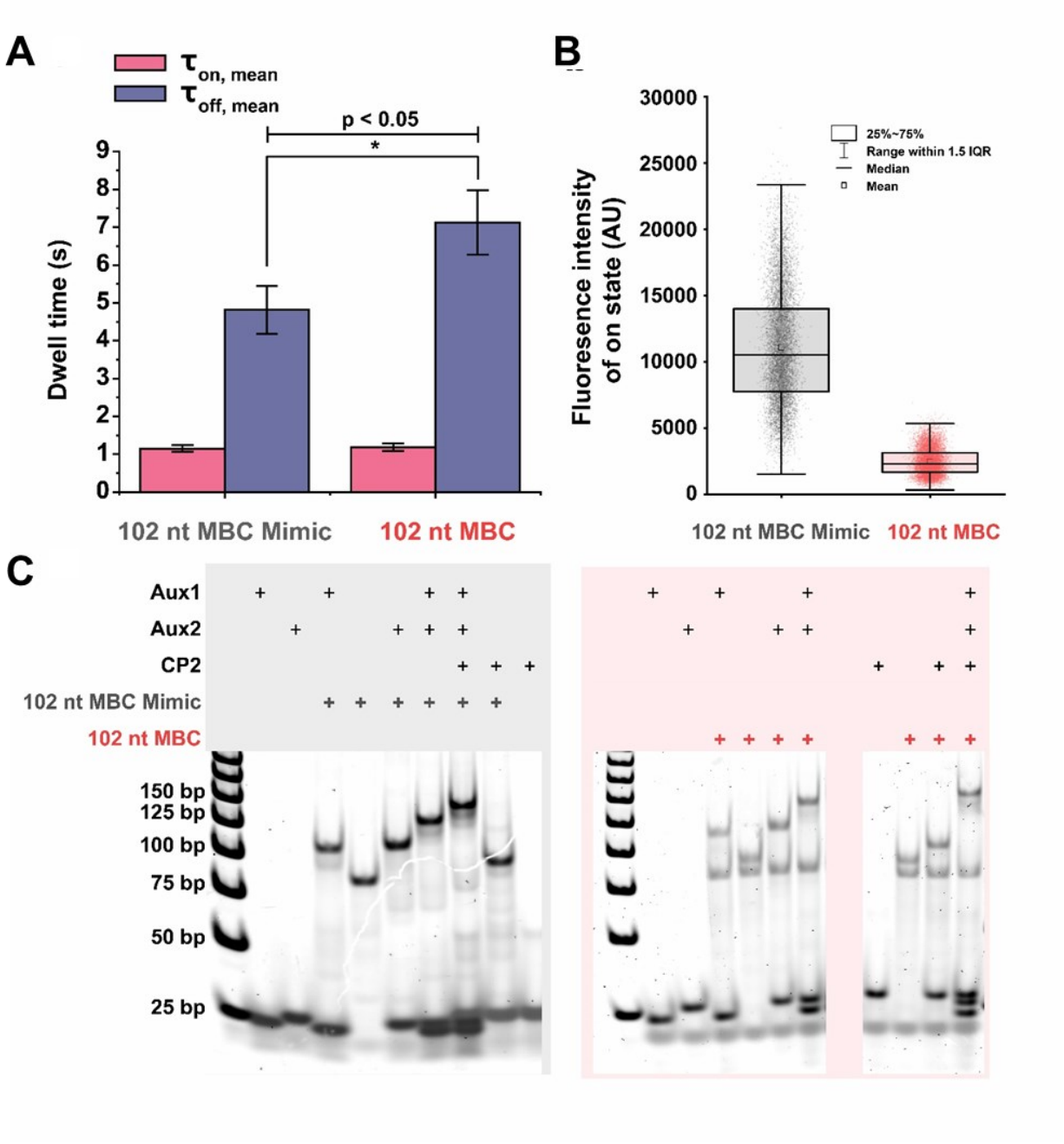

Figure 2.8. Differences between 102 nt MBC Mimic and 102 nt MBC. (A) Mean dwell times differences between detecting 102 nt MBC Mimic and 102 nt MBC. Mean τ_{on} and τ_{off} are calculated by fitting dwell times of individual events in all traces with exponential decay. Datapoints are presented as mean ± s.d., where n = 3 independent experiments. Confidence levels as assessed using a single-tailed, unpaired t-test. (B) Distribution of intensity level of on state for detecting 102 nt MBC Mimic and 102 nt MBC. Panel A&B are using the same capturing and imaging conditions as in **Figure 2.7**. (C) 5% native PAGE to show probes' binding to 102 nt MBC Mimic. (D) 5% native PAGE to show probes' binding to 102 nt MBC.

2.3.4 Sensitivity and specificity of BSM-SiMREPS for detection of mimic and real target

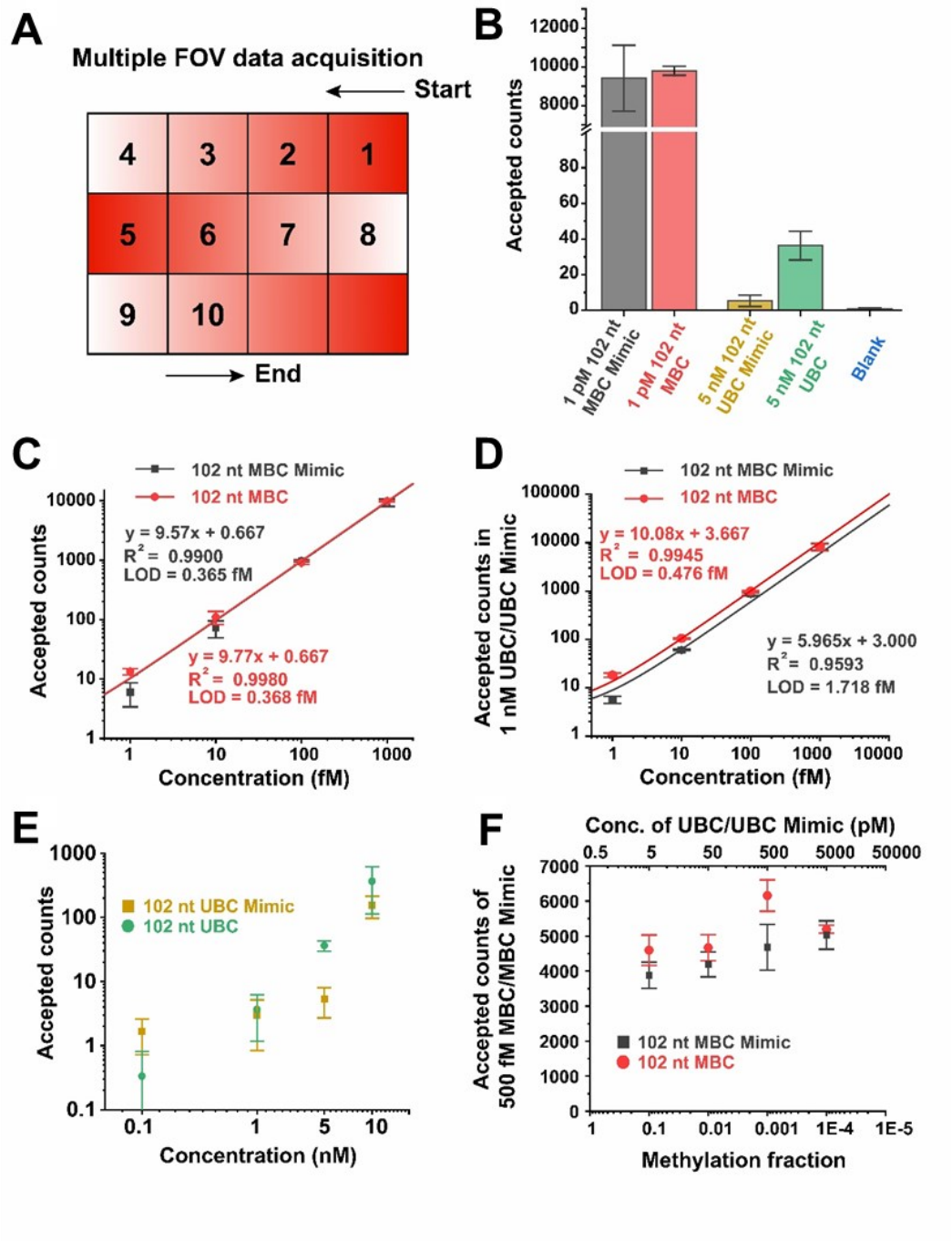

Figure 2.9. Quantification of MBC and MBC Mimic and analytical performances. (A) Illustration of multiple FOV data acquisition. The acquisition starts from top right corner and moves to the next adjacent FOV in a serpentine manner. (B) Accepted counts of different types of samples. (C) Standard curves of detecting 102 nt MBC Mimic and 102 nt MBC. (D) Standard curves of detecting 102 nt MBC Mimic in a background of 1 nM 102 nt UBC Mimic and detecting 102 nt MBC in a background of 1 nM UBC. (E) False positive counts of 102 nt UBC Mimic and UBC at different concentrations. (F) accepted counts of 500 fM 102 nt MBC in different concentrations of UBC and accepted counts of 500 fM 102 nt MBC Mimic in different concentrations of UBC Mimic. (Datapoints in panel A-F are presented as mean ± s.d., where n = 3 independent experiments. Ten FOVs are collected for each condition in panel A-F.

As indicated in **Figure 2.9**A, the total accepted counts from 10 FOVs produced a single readout. **Figure 2.9**B shows dramatic differences in accepted counts for detecting 1 pM MBC Mimic, 1 pM MBC, 5 nM UBC Mimic, 5 nM UBC, and blank. Almost zero false positive counts were achieved in the blank control with optimized kinetic filtering. **Figure 2.9**C shows an almost perfect alignment between two calibration curves of MBC Mimic and MBC with approximately the same LODs. In a background excess of 1 nM UBC Mimic or UBC, both calibration curves still maintain good linearity but with a slightly lower sensitivity for MBC detection and a significant interference for MBC Mimic detection. **Figure 2.9** shows measurements of false positives at different concentrations of UBC Mimic or UBC. At 10 nM UBC Mimic or UBC, accepted counts decreased over time across different FOVs (**Figure 2.10**), suggesting competitive surface dissociation of sensors incorporating non-methylation species. Therefore, we considered no more than 5 nM of UBC Mimic or UBC as a more reliable condition for testing specificity since the accepted counts were consistent across different FOVs measured over time. **Figure 2.9**F measured robustness of our assay against different concentrations of UBC Mimic and UBC when detecting 500 fM MBC Mimic and MBC, respectively. A slight increase in accepted counts was consistently observed as the concentration of non-methylation species increased. The accepted counts seemed to peak at 0.1% methylation fraction. No significant difference in kinetic fingerprint was observed among the different conditions. Therefore, we hypothesize that this increase was related to surface capture efficiency. One explanation may be that an excess of UBC or UBC Mimic helps prevent non-specific adsorption of MBC or MBC Mimic to the surface or the wall of our sample wells.

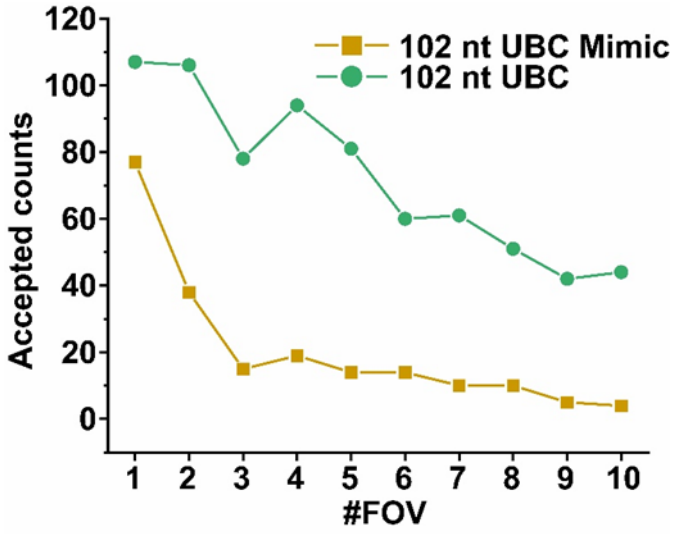

Figure 2.10. Accepted counts across different FOVs for detecting 10 nM 102 nt UBC Mimic and 10 nM 102 nt UBC. Different FOVs are collected sequentially in a timely order. In both cases of mimic and real target, a clear decrease is observed in accepted counts over time, indicating gradual dissociation of construct containing UBC Mimic or UBC on the surface.

Table 2.3. Maximum specificity imposed by thermodynamics and apparent specificity as well as discrimination factors calculated at different experiment conditions.

Concentration		ΔΔG (kJ/mol)	$Q_{max,therm}$ (×10^6)	Q_{app} (×10^6)	$Q_{app}/Q_{max,therm}$	Specificity
UBC Mimic	**MBC Mimic**					
5 nM	500 fM			9.42	0.113	99.9999894 %
1 nM	1 pM	-45.44	83.43	2.75	0.033	99.9999637 %
	100 fM			2.87	0.034	99.9999652 %
	10 fM			1.92	0.023	99.9999480 %
	1 fM			0.89	0.011	99.9998875 %
UBC	**MBC**					
5 nM	500 fM			1.03	0.012	99.9999028 %
1 nM	1 pM	-45.44	83.43	2.25	0.027	99.9999556 %
	100 fM			2.71	0.032	99.9999631 %
	10 fM			2.77	0.033	99.9999639 %
	1 fM			4.00	0.048	99.9999750 %

Table 2.3 summarizes the calculated LOD values, discrimination factors and specificity under different experimental conditions. The detailed calculation protocols were reported previously[72]. In the end, we achieved an LOD of 0.365 fM for detecting 102 nt hypermethylated BCAT1 promoter with BSM-SiMREPS, combined with a specificity of 99.9999%.

2.3.5 Detection of BCAT1 promoter methylation in a background of genomic DNA.

Next, we detected BCAT1 promoter methylation in a background of genomic DNA as shown in **Figure 2.11**. Different concentrations of 102 nt MBC were spiked into two different genomic DNA matrices – whole-genome amplified DNA (BS WGA) from the DKO HCT116 cell line, and DNA extracted from human male whole blood (BS Blood DNA; see **Table 2.1**). We bisulfite-converted genomic DNA separately and measured its molar concentration by treating each copy of haploid genome as a single molecule and then mixed converted genome DNA with 102 nt MBC at different ratios – MBC: genomic DNA = 0 fM: 20 fM, 1 fM: 20 fM, 5 fM: 20 fM, 10 fM: 20 fM, and 20 fM: 20 fM. Both genomic DNA preparations are commonly used as methylation-negative control samples. In the presence of genomic DNA, a higher fluorescence background arose due to non-specific interactions of FPs with genomic DNA fragments adsorbed to the surface, as seen in **Figure 2.11**A. However, sensor molecules that incorporated MBC or the BCAT1 promoter in genomic DNA fragments could still be distinguished due to their unique kinetic fingerprints, distinct from all background signals. Interestingly, we also detected a significant BCAT1 methylation level in blood DNA, whereas almost zero counts were detected in WGA (**Figure 2.11**B); BS WGA from the HCT116 DKO cell line is known to be an absolute methylation-negative control due to its derivation from PCR amplification. The fact that 20 fM haploid +Me WGA (see **Table 2.1**) gave almost the same average number of counts as 20 fM haploid WGA + 20 fM MBC supports the conclusion that no

over- or undercounting occurred of BCAT1 methylation sites in WGA, further validating the robustness of our assay. By establishing calibration curves for detecting MBC in both BS Blood DNA and BS WGA (**Figure 2.11**C), we were able to determine an LOD of 1. 33 fM and 1.62 fM, respectively. The slight decrease in sensitivity for BS WGA may result from both compromised capture efficiency due to competition from other genomic DNA fragments or high background due to nonspecific interactions between FPs and genomic DNA fragments. Based on the linear fitting of the calibration curve for BS WGA, the concentration of methylated BCAT1 promoter in 20 fM haploid Blood DNA was estimated to be 6.1 fM, corresponding to a 31% methylation level.

We further compared our measurement results from BSM-SiMREPS with both a microarray-based assay and NGS. Using recountmethylation (a R package that allows access to compiled methylation beta values from GEO database. See Materials and methods), we compiled methylation beta values at BCAT1 promoter measured on Illumina Infinium MethylationEPIC microarray (EPIC array) from 12,392 blood samples available on GEO[103]. These blood samples are a subset of over 38,000 studies using the EPIC array (GEO accession ID: GPL21145) and consist of most studies available until December 2022. There were three degenerate CpG targeting probes available that covered the BCAT1 promoter as shown in **Figure 2.11**D. The distributions of methylation beta values of blood samples and whole blood samples are shown in **Figure 2.11**E. All three CpG probes generated signals comparable to each other, with a median value of ~5%, well below the 31% found via BMS-SiMREPS. In terms of NGS data, as far as we could find, we manually combined 72 tracks at BCAT1 promoter derived from the bisulfite sequencing studies found on the UCSC genome browser. These 72 tracks originated from

different subtypes of blood cells. **Figure 2.11**F shows their methylation distributions; the values at all CpG islands had a median value around 2%, again well below 31%.

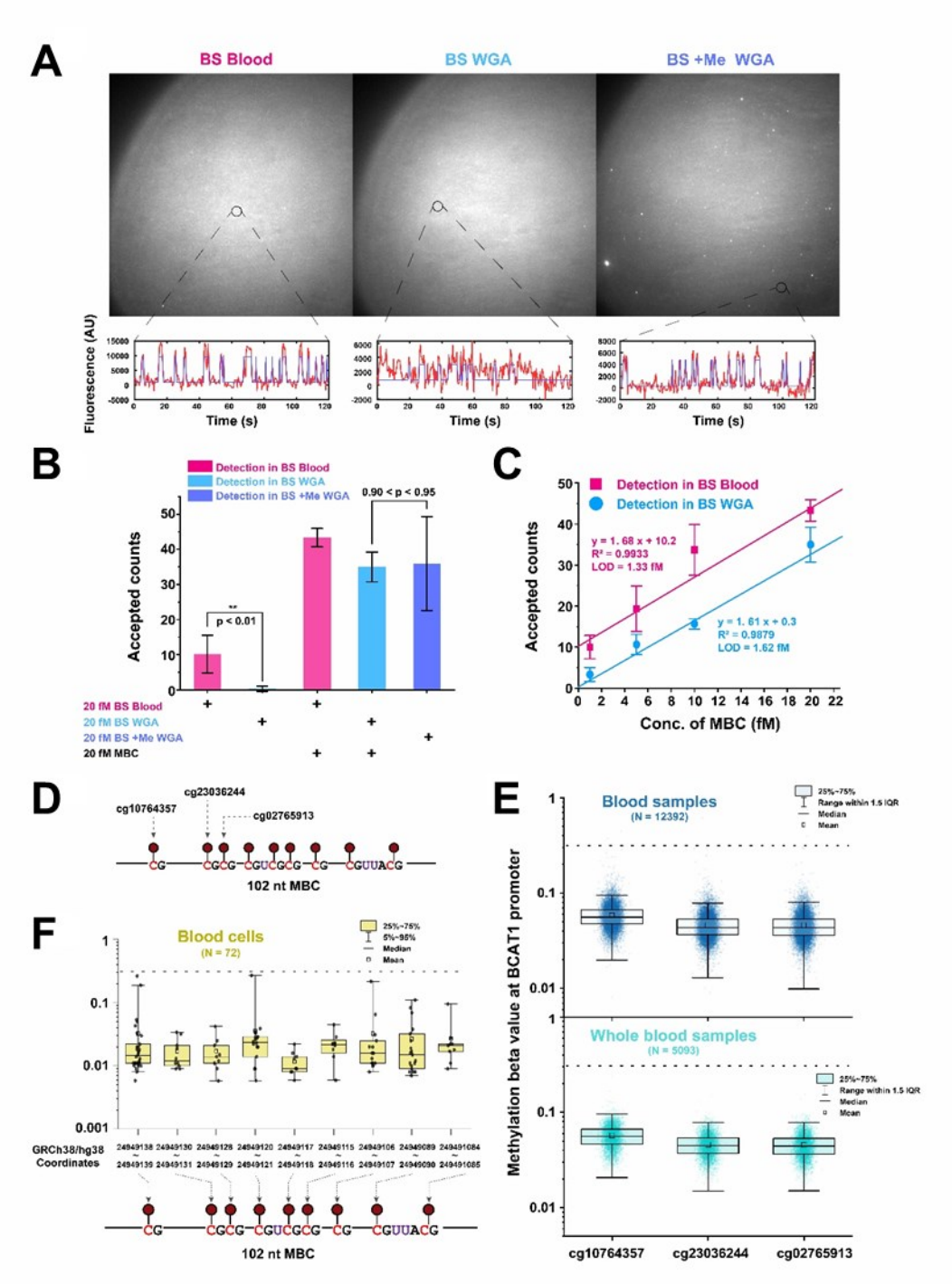

Figure 2.11. Quantification of 102 nt MBC in a background of two types of genomic DNAs and comparison with NGS and illumina Infinium MethylationEPIC microarray (EPIC array) at BCAT1 promoter. (A) Screenshot of raw movies for detecting three different types of genomic DNAs and their associated representative intensity-time traces

(red lines) fitted by HMM (blue lines). (B) Quantification of different genomic DNAs at 20 fM. Confidence levels as assessed using a single-tailed, unpaired t-test. (C) Standard curves for detecting 102 nt MBC in two types of genomic DNAs. (D) Illustration of three cg probes covering BCAT1 promoter used in EPIC array (GEO accession ID: GPL21145). (E) Distribution of methylation beta values at BCAT1 promoter region for blood and whole blood samples using EPIC array. Data are compiled for studies on GEO until December 2022 using recountmethylation. (F) Distribution of methylation fractions using bisulfite-sequencing at each CpG site in BCAT1 promoter in blood cells. (Dash lines in panel E-F represent methylation levels of whole blood DNA at BCAT1 promoter measured by standard curves in panel D.)

2.4 Discussion

In this study, we developed a methylation-specific single molecule fluorescence kinetic fingerprinting biosensor by combining SiMREPS with bisulfite treatment, and demonstrated detection of hypermethylation of BCAT1 promoter, a methylation biomarker associated with colorectal cancer. Initial optimization of sensor design almost completely suppressed background signals due to the constrained detection volume of TIRF microscopy imaging and the strict signal generation condition imposed by FRET between adjacently bound donor and acceptor carrying FPs. Further optimizations of FP pair sequences, imaging temperature and FP concentrations not only achieved an ultra-low background detection, but also enhanced sensitivity by speedy acquisition of multiple FOVs for a single readout. In the end, we accomplished sub-femtomolar LODs with a specificity of 99.9999% in the detection of both mimics and real targets (see **Table 2.1**) of the BCAT1 promoter. Furthermore, we demonstrated reliable quantification in a background of genomic DNA while avoiding both over- and undercounting of BCAT1 methylation sites. Finally, we observed 31% methylation level at BCAT1 promoter in human male whole blood DNA, much higher compared to published measurements from the EPIC array or NGS. This difference may be explained by PCR bias – that is, during library preparation, unequal amplification of the bisulfite-converted, damaged methylated and unmethylated targets.

Our result also suggested that isolating blood cells is an essential step in detecting DNA methylation biomarkers in liquid biopsies. Whole blood DNA mainly consists of DNAs from blood cells and less than 1% of circulating tumor DNAs in cell-free DNAs are derived from tumor cells in early-stage cancer patients[105,106]. For BCAT1 promoter, direct measurement of whole blood samples for DNA methylation shall not give any differences between healthy and cancer patients.

Previous studies suggest both underestimation in MSP (methylation-specific PCR) and overestimation of methylation levels in NGS or EPIC array, with all these studies using PCR-based tools for validation[54,55,107–109]. Here we present in BSM-SiMREPS an independent amplification-free approach that suggests a significant underestimation of BCAT1 promoter methylation in whole blood from healthy humans. We also call for caution in data validation of PCR-based measurements in bisulfite-converted samples due to the significant PCR bias, which we posit has been underappreciated in the past.

One severe challenge we encountered in bisulfite conversion was potentially biased yield of the Me-BCAT1 Forward and Reverse DNA strands (**Figure 2.12**). By contract, we did not observe such a bias when bisulfite converting WGA, as indicated in **Figure 2.11**B. To circumvent this problem, we bisulfite-converted Me-BCAT1 Forward and Reverse individually and then mixed them in equimolar amounts to compensate for biased yield during bisulfite treatment. This bias might be due to differential binding efficiency of the two target strands to the resin used in the purification column when DNA is relatively short (close to 100 nt), which may explain why the yield bias disappeared when treating WGA that contains larger DNA fragments. It is likely that different batches of resin material have a differential binding

efficiency for short oligodeoxynucleotides since we did not observe yield bias in experiments for **Figure 2.9** in our early experiments (see **Figure 2.9**C).

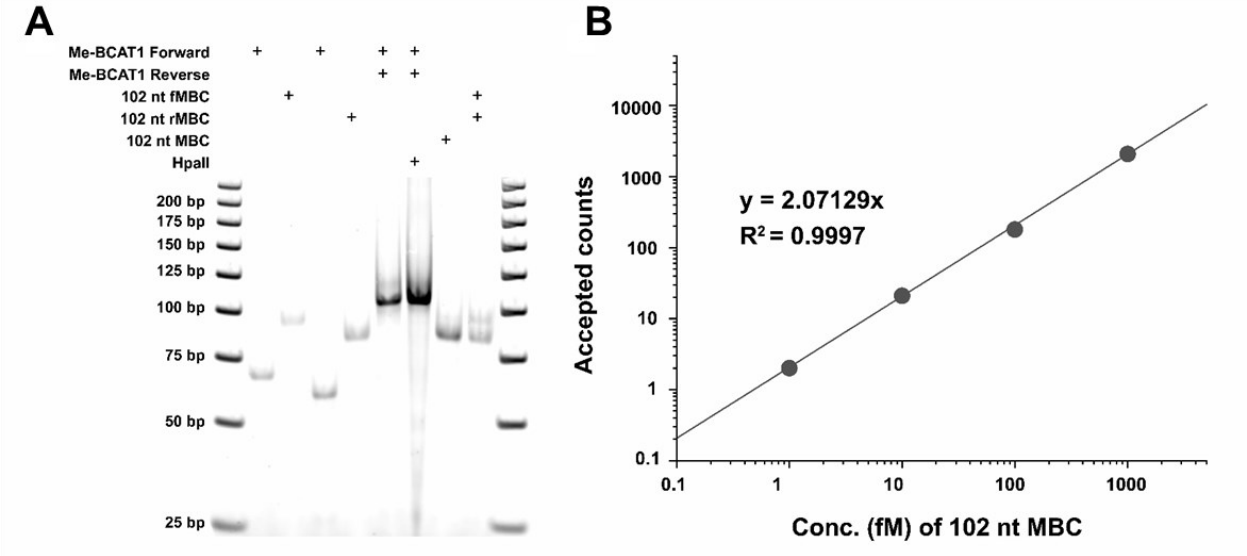

Figure 2.12. Bias yield in bisulfite conversion. (A) 5% native PAGE to assess amount of individual strands after bisulfite conversion. For descriptions of each strand, see **Table 2.1** for details. HpaII is a methylation-sensitive restriction enzyme. Methyl group will inhibit digestion activity of HpaII. The third lane to the left clearly shows disproportional amount of forward and reverse strand after bisulfite treatment and downstream column cleanup. (B) quantification of different concentrations of 102 nt MBC using the same batch of samples in panel a. We can clearly see a drop in accepted counts compared standard curves in **Figure 2.9**. The input concentration of fMBC is significantly overestimated by Qubit since the majority of ssDNA is rMBC.

Our current LOD is still limited by capture and detection efficiency[52,74]. Capture efficiency is defined as the percentage of target molecules that are immobilized on the surface compared to total molecules in solution. Detection efficiency is defined as the percentage of molecules that are detected by our assay compared to total molecules immobilized. With a 60x objective and our camera chip size, the size of one FOV is 136.53 μm by 136.53 μm. Our sample well has an inner diameter of 5.842 mm – the size of total surface for capturing targets. Therefore, roughly only 0.7% of target molecules that are captured are detectable in the currently 10 FOVs we are combining. Furthermore, diffusion-limited mass transport to the surface results in a rather low capture efficiency, around 0.5% to 1.5%, as reported previously for SiMREPS[52].

The total analytical efficiency is only 3.5 to 10.5 x 10^{-5}. Therefore, pre-enrichment methods like aqueous-two-phase system or digitalization of SiMREPS, as well as reading out more FOVs, could further lower LODs[74]. Bisulfite treatment is another limiting factor in our assay performance due to loss of material (see **Figure 2.12**), DNA damage, and reaction selectivity. Both incomplete conversion of unmethylated cytosines and over-conversion of methylated cytosines will compromise capture efficiency and assay specificity.

A promising feature of BMS-SiMREPS is its integration potential since the FP sequences are independent of the target sequences. That is, the same FP pair can be used for detection of at multiple methylation loci simultaneously, which would not only increase sensitivity but also read out the total level of multiple methylation-related biomarkers in cancer patient blood. The results presented here establish the foundation for SiMREPS-based biosensors that directly measure DNA methylation by kinetic fingerprinting and have the potential to be used for clinical applications in cancer diagnostics upon further improvements in LOD.

Chapter 3 Engineering and Characterization of a Direct Single Molecule Imager for DNA Methylation Detection

3.1 Introduction

Current techniques in DNA methylation detection, with or without amplification, heavily rely on pretreatment of methylated and unmethylated DNA, either using chemical treatment like bisulfite conversion, enzymatic treatment like methylation-sensitive restriction enzyme, or selective enrichment using anti-methyl antibodies or methyl-binding domain (MBD) protein family members[86]. In either case, distinct products of methylated DNA and unmethylated DNA after pretreatment are key determinants of detection specificity. The "gold standard" detection approaches, bisulfite sequencing or digital Methylation-specific PCR (MSP), involve bisulfite conversion as the key step for distinguishing methylated DNA versus unmethylated DNA. However, several technical concerns arise in the process of bisulfite conversion. First, a significant portion of DNA material is lost after purification. Second, the harsh chemical environment during bisulfite conversion causes fragmentation of long DNA as well as depurination or depyrimidination. Third, an incomplete conversion of unmethylated cytosines and aberrant conversion of methylated cytosines increases false positives and false negatives, respectively, compromising overall assay performance. While lots of current efforts are focusing on optimizing existing pretreatment methods or finding alternatives, pretreatment-free direct detection of DNA methylation has been a long-missing tool[110].

3.1.1 Seeking a methyl CpG binder

Guided by the principle of SMFKF, we were seeking to develop an amplification-free pretreatment-free single molecule detection method for DNA methylation.

To directly measure DNA methylation using SMFKF, the very first question coming up is where to find a weakly interacting partner for methylated DNA, specifically methyl CpG. In fact, nature has evolved a large variety of proteins that can recognize methyl CpG in both prokaryotes and eukaryotes[111–116]. In prokaryotes, methyl-cytosine is most involved in the naïve restriction-modification immune system, where the unmethylated genome DNA of an invasive pathogen is digested by methylation-sensitive restriction enzyme, while the bacterial genome is protected by DNA methylation[117,118]. However, DNA methylation in prokaryotes and viruses rarely occurs in CpG dinucleotides[117,119]. In contrast, methyl CpG is the dominant form of DNA methylation in mammalian genomes and extensively studied over the past half-century. We therefore restricted our search to mammalian binders of methyl CpGs.

Regardless of the diverse roles of DNA methylation in mammalian, proteins binding methyl CpG can be generally classified into three categories: writers, readers and erasers, just like for other epigenetic modifications[24]. In principle, either writers, readers or erasers may be chosen as a sensor candidate for methylation detection. However, writers and erasers catalyze enzymatic reactions right at CpG dinucleotides and often have binding affinity towards both the substrate and product of their reaction. Screening for mutations that not only inactivate their catalytic activity but also alter them to become a weak binder suitable for SMFKF is a laborious process, not to mention the requirement of avoiding unwanted interactions with the unmethylated substrates or products of their enzymatic reactions. Thus, readers for methyl CpGs are a better candidates for SMFKF sensors.

All canonical reader proteins for DNA methylation in mammals belong to a single family, the MBD superfamily[112,120,121]. As shown in **Figure 3.1**, the superfamily members discovered so far are part of one of three classes: (histone acetyltransferase MBD) HAT_MBD, (histone methyltransferase MBD) HMT_MBD and MeCP2_MBD, all of which share a conserved MBD domain consisting of approximately 70-85 amino acids, responsible for recognizing methyl CpG motifs[120]. Apart from the MBD superfamily, Zhu *et al.* suggested that many transcription factors can also bind methyl CpG motifs[122]. However, this is beyond our scope due to their huge heterogeneity in sequences and structures and still ill-defined binding patterns.

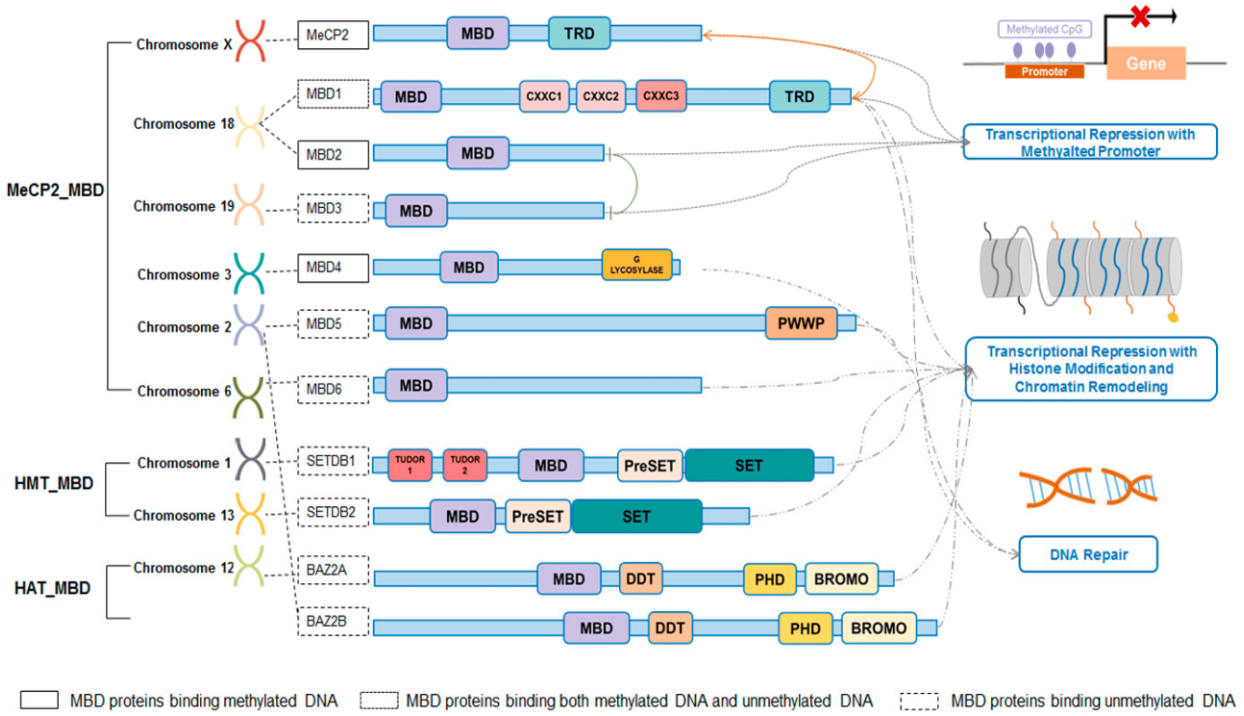

Figure 3.1. Genome locations, sequence homologies and functions of the three classes of the MBD superfamily. MBD: methyl-CpG-binding protein, TRD: transcription repression domain, CxxC: unmethylated-CpG-binding zinc finger, G/R: glycine/arginine rich domain, CC: coiled-coil domain, Glycosylase: DNA glycosylase function, P-rich: proline rich domain, PWWP: Pro-Trp-Trp-Pro motif domain, Tudor: Tudor domain, SET: suvar3-9, enhancer-of-zeste, trithorax domain, PreSET: the domain N-terminal to SET, DDT: DNA binding protein, Bromo: bromodomain, PHD: PHD (plant homeodomain) zinc finger motif. Figure taken from Li et al[120].

3.1.2 The MBD domain as an ideal candidate for SMFKF

Based on the above discussion, we therefore focused our search on the MBD superfamily and specifically on MBD domains. There are several advantages of building a direct fluorescent reader for SMFKF based on a simple MBD domain. First, its unique and small size of around 75 amino acids allows for relatively fast diffusion (potentially with a desirable high k_{on}) and makes it less likely to have unwanted interactions. Second, the MBD domain is solely responsible for recognizing methylated CpG and evolutionarily conserved to distinguish unmethylated from methylated CpGs well, thus ensuring detection specificity. Third, the MBD domain alone only needs a single methyl CpG dinucleotide motif for binding, minimizing the size of the detection region for a simplest possible sensor assembly. Fourth, the binding affinity and kinetics of the MBD domain has been well characterized in different sequence contexts[123–126] and has been extensively used in MBD-Seq, where magnetic beads coated with MBD domains enrich methylated genomic DNA fragments, whereas unmethylated DNAs are washed away[86,127–129]. Lastly, both crystal structures and NMR solution structures of MBD-methyl-DNA complexes have been solved, making it convenient for choosing a proper fluorophore labeling site[130,131].

For an SMFKF imager, the most important question is whether its binding kinetics fall into the desired regime where both τ_{on} and τ_{off} are fast within our observation window given an imager concentration of lower than 100 nM (as an imager at higher concentration will generate too much background). Specifically, we would need a k_{on} between approximately $0.06\ \mu M^{-1}s^{-1}\ and\ 20\ \mu M^{-1}s^{-1}$ and a k_{off} between approximately $0.01\ s^{-1}\ and\ 1\ s^{-1}$. That is, the K_d of a SMFKF imager should be between approximately $100\ nM\ and\ 10\ \mu M$. **Figure 3.2** describes the binding affinity of GFP-MBD where a single MBD domain is labeled by a green fluorescent protein (GFP)[125]. With different methylation patterns and sequence contexts,

prior work has shown that GFP-MBD exhibits a wide range of K_ds from approximately 0.1 μM up to 100 μM, with the highest affinity towards symmetrically methylated dsDNA, well matching the thermodynamic regime of a suitable SMFKF imager. **Figure 3.2** also suggests that we can easily tune the interaction of an MBD sensor with methylated CpG motifs by changing the methylation pattern in the auxiliary probes of our sensor construct for optimal binding kinetics.

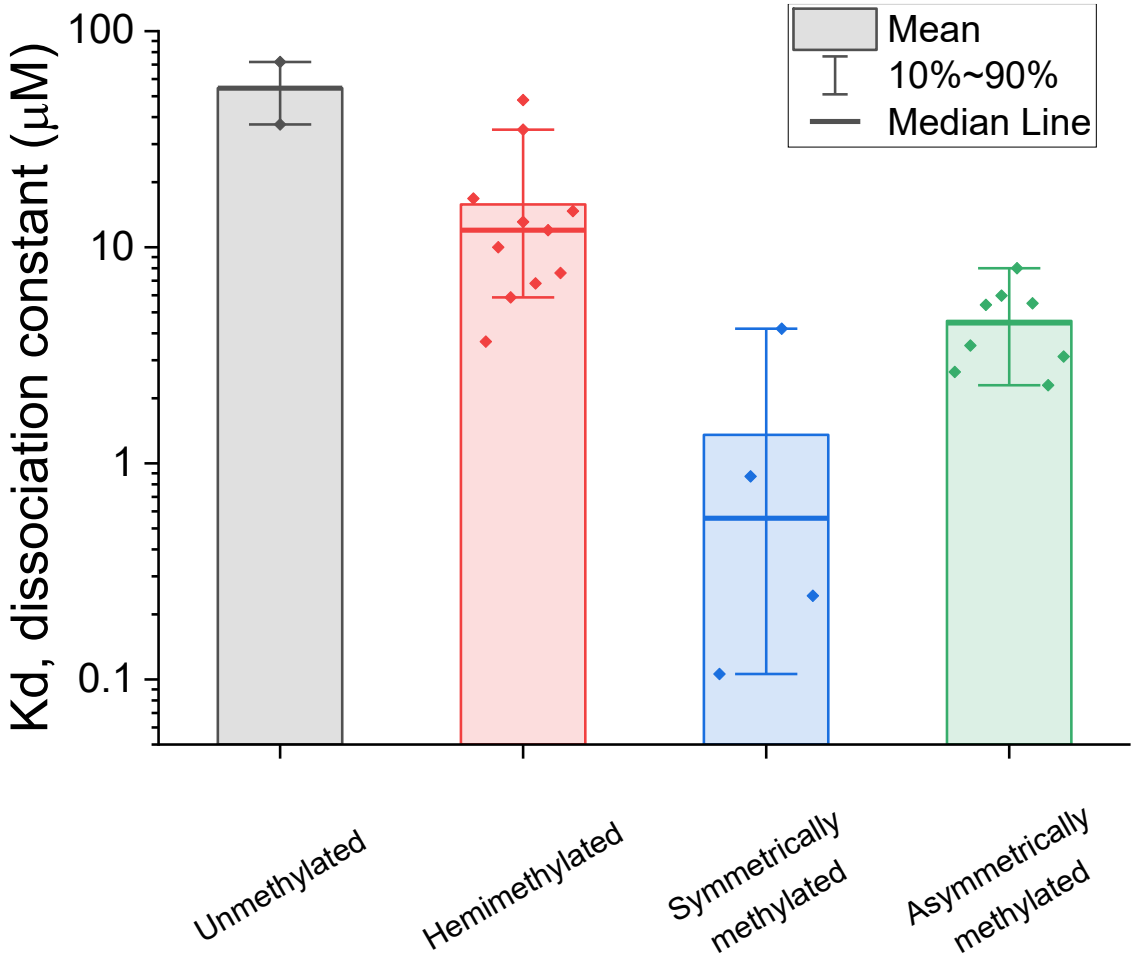

Figure 3.2. Distribution of binding affinity of GFP-MBD under different methylation patterns. Data replotted from Heimer et al[125].

3.1.3 Outline

Based on the above discussion, we decided to use a well-known MBD domain for engineering and characterization of a direct single molecule imager for detection of DNA methylation. To fluorescently label the MBD domain, either ybbR-tag or HaloTag tags were fused with hMBD1 MBD (aa 1-77) for site-specific covalent modification with a fluorophore. Ensemble-level functional assays were used for testing the methyl-CpG binding activity for both unlabeled and labeled methyl-CpG targets. Subsequently, based on the fluorescently labeled Halo-tagged MBD, we developed a pretreatment-free amplification-free single molecule detection assay using SMKFK, termed MBD-SiMREPS. Finally, the effects of methylation pattern and existence of a DNA branch motif on the m5C-binding kinetics were studied systematically using MBD-SiMREPS.

3.2 Materials and methods

Table 3.1. Lists of DNA strands, their code names, sequences and descriptions.

Code name	Sequences	Description
	Target	
BCATa_+MeC_55	GTCTTCCTGCTGATGCAATC/iMe-dC/GCTAGGT/iMe-dC/G/iMe-dC/GAGTCTC/iMe-dC/GC/iMe-dC/G/iMe-dC/GAGAGGGC/iMe-dC/GG	Fully Methylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
BCATa _55	GTCTTCCTGCTGATGCAATCCGCTAGGTC GCGAGTCTCCGCCGCGAGAGGGCCGG	Unmethylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
BCATa_+MeC_comple m_55	C/iMe-dC/GGCCCTCT/iMe-dC/G/iMe-dC/GG/iMe-dC/GGAGACT/iMe-dC/G/iMe-dC/GACCTAG/iMe-dC/GGATTGCATCAGCAGGAAGAC	Fully Methylated 55 nt BCAT1 promoter reverse strand, directly purchased from IDT

BCATa_omplem_55	CCGGCCCTCTCGCGGCGGAGACTCGCGAC CTAGCGGATTGCATCAGCAGGAAGAC	Unmethylated 55 nt BCAT1 promoter reverse strand, directly purchased from IDT
BCATa_+MeC_55_v1a	GTCTTCCTGCTGATGCAATC/iMe-dC/GCTAGGTCGCGAGTCTCCGCCGCGAG AGGGCCGG	Single Methylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
BCATa_+MeC_55_v1b	GTCTTCCTGCTGATGCAATCCGCTAGGTC GCGAGTCTCCGCCGCGAGAGGGC/iMe-dC/GG	Single Methylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
BCATa_+MeC_55_v2	GTCTTCCTGCTGATGCAATCCGCTAGGT/ iMe-dC/G/iMe-dC/GAGTCTCCGCCGCGAGAGGGCCGG	Double Methylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
BCATa_+MeC_55_v3	GTCTTCCTGCTGATGCAATCCGCTAGGTC GCGAGTCTC/iMe-dC/GC/iMe-dC/G/iMe-dC/GAGAGGGCCGG	Triple Methylated 55 nt BCAT1 promoter forward strand, directly purchased from IDT
dsMe-BCAT_55	NA	Prepared by mixing equal amount of BCATa_+MeC_55 and BCATa_+MeC_complem_55nt in PBS buffer
	Sensor construct	
CP_Br	/5Biosg/ATAATTAATAGCATCAGCAGG AAGAC	Biotinylated capture probe, partially complementary to target sequence, with a "branch" motif at 5' end, directly purchased from IDT
CP	GCATCAGCAGGAAGAC/3BioTEG/	Biotinylated capture probe with no "branch" motif at 5' end, directly purchased from IDT
Aux_+MeC	C/iMe-dC/GGCCCTCT/iMe-dC/G/iMe-dC/GG/iMe-dC/GGAGACT/iMe-dC/G/iMe-dC/GACCTAG/iMe-dC/GGATT	Fully methylated auxiliary probe, fully complementary to target sequence, with no

		"branch" motif, directly purchased from IDT
Aux	CCGGCCCTCTCGCGGCGGAGACTCGCGACCTAGCGGATT	Unmethylated auxiliary probe, fully complementary to target sequence, with no "branch" motif, directly purchased from IDT
Aux_Br	CTTATCTGTTTCGCGACCTAGCGGATT	Unmethylated auxiliary probe, partially complementary to target sequence, with a "branch" motif at 5' end, directly purchased from IDT
	Cloning	
Gy-hMBD-FWD	tgctagtaagcttgcgATGGCTGAGGACTGGCTGGAC	Forward primer for insertion of ybbR tag upstream of hMBD1 MBD in addgene plasmid #119966, directly purchased from IDT
Gy-hMBD-REV	ataaattcaagagaatcACTACCACGCGGAACCAGGCC	Reverse primer for insertion of ybbR tag upstream of hMBD1 MBD in addgene plasmid #119966, directly purchased from IDT
Halo-hMBD-BF	GGAGGTGGAAGCGGTGAA	Forward primer for linearization of backbone plasmid pBD003_mut_VCP (R155H) for constructing Halo-hMBD, directly purchased from IDT
Halo-hMBD-BR	ATGTATATCTCCTTCTTAAAGTTAA	Reverse primer for linearization of backbone plasmid pBD003_mut_VCP (R155H) for constructing Halo-hMBD, directly purchased from IDT
Halo-hMBD-IF	GAAGGAGATATACATATGGCTGAGGACTGGCTGG	Forward primer for PCR amplification of insert hMBD1 MBD from addgene plasmid #119966 for constructing

		Halo-hMBD, directly purchased from IDT
Halo-hMBD-IR	ACCGCTTCCACCTCCATGGGCCTTGGGGGCTGG	Reverse primer for PCR amplification of insert hMBD1 MBD from addgene plasmid #119966 for constructing Halo-hMBD, directly purchased from IDT

3.2.1 Assay pipeline and working principle

As shown in **Figure 3.3**, we applied the principle of SiMREPS to methylation detection by using a directly fluorescence labeled MBD imager. We chose the 55 nt BCAT1 promoter as the gene of interest to demonstrate MBD-SiMREPS. The assay starts with mixing methylated or unmethylated target with an excess of auxiliary probes that are either methylated or unmethylated (shown in **Table 3.1**). We first heat-denature the double-stranded target and slowly cool down the system to ensure complete incorporation of targets into sensor construct. After sensor assembly, the sensor mixture is then added to sample wells, where a biotinylated capture probe (shown in **Table 3.1**) is attached to a streptavidin-coated PEG surface through the biotin-streptavidin interaction. Capture probes were used to immobilize sensors with either methylated or unmethylated targets. Following surface capture, an imaging buffer containing fluorescently labeled MBD as well as oxygen scavenger system is added and the sample well is imaged under TIRF (total internal reflection fluorescence) illumination using an oil objective. Transient interaction of MBD with methyl-CpGs generates kinetic fingerprints upon laser excitation. The resulting fluorescence emission signal is collected for downstream data analysis. Repeated binding and dissociation of MBD protein with methyl-CpG cluster is visualized by fluorescence-time traces, which are used for distinguishing methylation-negative and background signals.

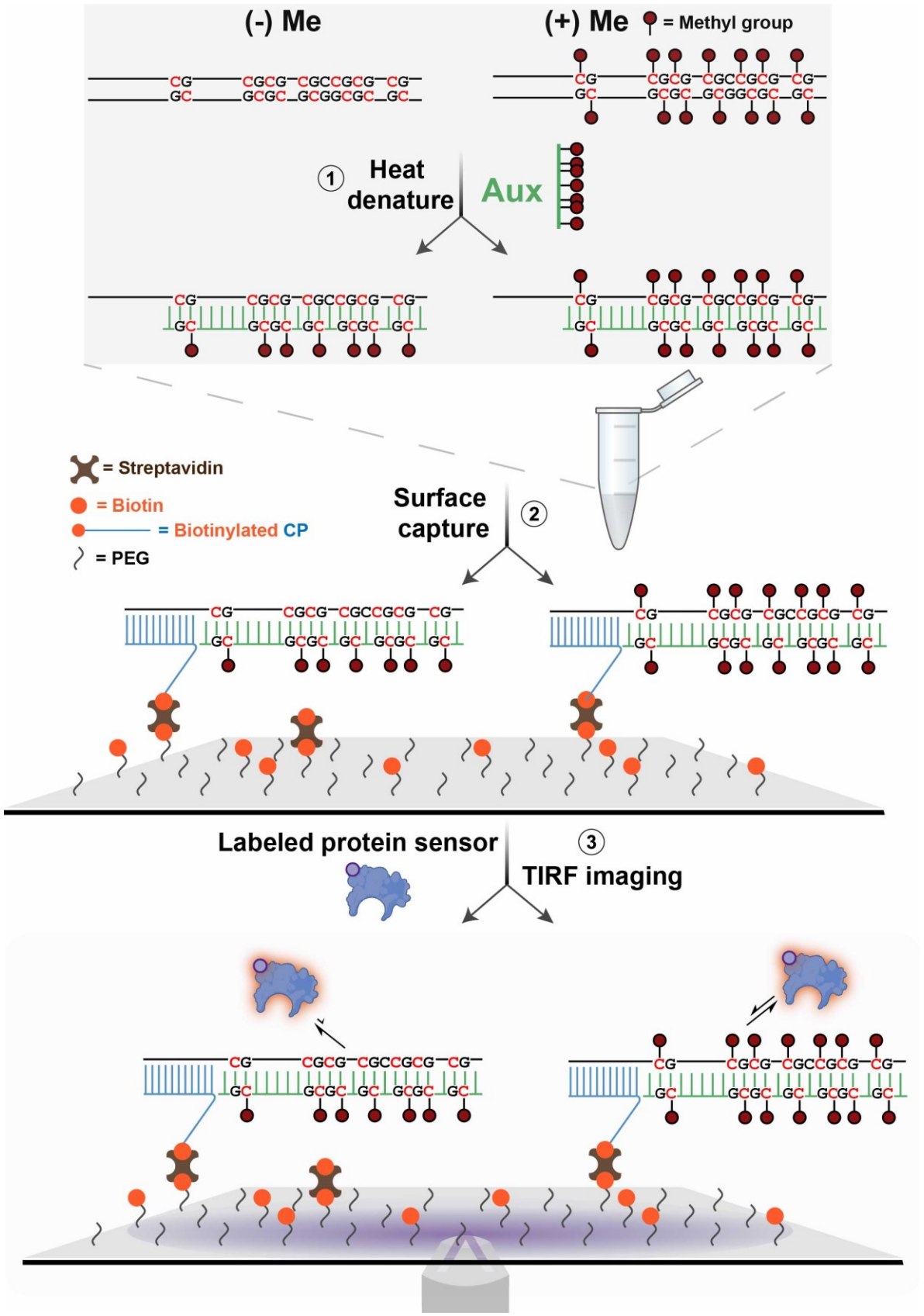
(-) Me
(+) Me
= Methyl group
① Heat denature
Aux
= Streptavidin
= Biotin
= Biotinylated CP
= PEG
Surface capture
②
Labeled protein sensor
③
TIRF imaging

Figure 3.3. Schematic of the MBD-SiMREPS pipeline. (I) The BCAT1 promoter (methylated or unmethylated) of 55 nt in sample matrix is bound by an excess of auxiliary probe (Aux) under high temperature followed by slow cooling down. (II) After sensor assembly, the mixture is incubated with a capture probe coated streptavidin-pegylated coverslip. (III) And finally, an imaging buffer containing two fluorescently labeled MBD protein is added to the sample well and the surface is excited by a laser under TIRF illumination. Fluorescence is collected as readout. Repeated binding and dissociation of MBD protein with methyl-CpG cluster will be visualized by fluorescence-time traces, which are used for distinguishing methylation-negative and background signals.

3.2.2 Oligonucleotides

All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT, www.idtdna.com) with standard desalting purification, unless otherwise noted. The 55 nt promoter sequence of branched-chain amino acid transaminase 1, BCAT1 was chosen as our detection target — its genomic coordinates were Chr12: 24,949,105 - 24,949,159 (genome build: UCSC Genome browser GRCh38/hg38 version)[85,91]. Probe CP is PAGE-purified. See **Table 3.1** for descriptions of each target and their acronyms.

3.2.3 Cloning, expression and purification of Gy-hMBD

To generate the construct for expressing Gy-hMBD, we started with the vector pET28GST-6xHis-MBD (addgene plasmid #119966) where human MBD1 MBD (aa 1-77) is fused with a N-terminal GST tag followed by 6xHis tag. Downstream of 6xHis is a thrombin recognition and cleavage site. Insertion of ybbR tag (aa sequence: DSLEFIASKLA, coding DNA sequence: gattctcttgaatttattgctagtaagcttgcg) was achieved by PCR linearization of pET28GST-6xHis-MBD using two phosphorylated primers: Gy-hMBD-FWD and Gy-hMBD-REV, each of which contains one half of ybbR coding sequence. After DpnI digestion (NEB, Cat. # R0176S) followed by PCR purification (QIAGEN, Cat. # 28106), purified linearized products were ligated using T4 ligase (NEB, Cat. # M0202S) following the vendor-provided protocol for blunt end ligation. Finally, ligation product was used for transformation of NEB, 5-alpha competent cells

(NEB, Cat. # C2987H) and colonies were selected using 50 µg/ml Kanamycin LB agar plate and their plasmid sequences were validated by Sanger sequencing.

For overexpression and purification of Gy-hMBD, its construct was transformed into BL21(DE3) competent E. coli (NEB, Catalog # C2527H) and transformed cells were spread on a 50 µg/ml Kanamycin LB agar plate. Single colonies were inoculated into a 100 ml LB culture containing 50 µg/ml Kanamycin and grown overnight at 250 rpm, 37°C. The OD600 of this overnight culture was measured and a certain amount of it was further inoculated into a 1 L TB culture with 50 µg/ml Kanamycin for large-scale expression such that the starting OD600 was exactly at 0.01. After incubating at 250 rpm, 37°C for approximately 3-4 h, its OD600 reached 0.6 and overexpression of Gy-hMBD was induced by addition of 0.05 M IPTG right after we cooled it down in ice bath. Large expression culture was further incubated at 250 rpm 20-22°C for another 16 h. After expression, E. coli culture was spun down at 5,000 g, 4°C for 20 min and cell pellets were pooled and resuspended in a lysis buffer containing 1X PBS, 1X protease inhibitor cocktail, 1 mg/ml lysozyme (Millipore Sigma Cat. # L6876-10G) and 1 mM freshly thawed DTT. Otherwise, cell pellets would be flash-frozen and be stored at -80°C until purification. Note that 1X protease inhibitor cocktail was prepared by dissolving 2 tablets of cOmplete™, EDTA-free Protease Inhibitor Cocktail ((Millipore Sigma Cat. # 11873580001) in 50 ml buffer. In general, 50 ml lysis buffer was used per liter of TB cell culture. Cell lysis was achieved by sonication in ice water with 5 s on and 15 s off at 70% amplitude for 30 min (total time) until cell suspension became semitransparent. Subsequently, lysate was spun down at 20,000 g, 4°C for 60 min. Supernatants were further clarified through a 0.2 µm syringe filter (Fisher Scientific, Cat. # 09-719C) before sample application using FPLC (Fast Protein Liquid Chromatography, Bio-Rad NGC 10 Medium-Pressure Chromatography System). All buffers

were filtered through 0.2 µm filter and degassed prior to running FPLC purification. A prepacked glutathione affinity column (GSTrap™ HP Columns, Cytiva Cat. # 17528201) was attached to the FPLC system and equilibrated with 1X PBS, pH 7.4 and 1 mM DTT at 1 ml/min for 4-5 column volumes (CVs). Following that, filtered supernatants were loaded at 1 ml/min and washed with 1X PBS, pH 7.4 and 1 mM DTT at 2 ml/min for 10 CVs. Resin-bound Gy-hMBD was collected at 1 ml/min for 10-15 CVs using a gradient elution by mixing a buffer of 50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 mM reduced glutathione and 1 mM DTT with another buffer containing everything the same except 10 mM reduced glutathione. Fractions were loaded on denaturing PAGE to examine protein purity before combining them together. To remove nucleic acid contamination, combined fractions were further loaded onto a prepacked Hitrap heparin column (Cytiva, Cat. # 29051324) after equilibrating with 10 CVs of 1X PBS pH 7.4 and 1 mM DTT at 1 ml/min. After washing for 10 CVs of 1X PBS pH 7.4 and 1 mM DTT at 1 ml/min, nucleic-acid-free Gy-hMBD was eluted with a continuous gradient from 0% to 100% elution buffer containing 2 M NaCl, 1X PBS pH 7.4 and 1 mM DTT at 1 ml/min for 10 CVs. Fractions of pure Gy-hMBD were combined and dialyzed in 1X PBS, pH 7.4 and 1 mM DTT, concentrated and stored in 50% (v/v) glycerol, 1X PBS, pH 7.4 and 1 mM DTT at -20°C.

3.2.4 Synthesis and HPLC purification of CoA-Cy5

To prepare and purify CoA-Cy5, a solution of 8 mM Cy5 maleimide dissolved in 40 µl DMF (0.32 mmol) was diluted in 10 µl 50 mM Tris HCl pH 7.4 and 90 µl DMF. The solution mixture was added to 0.9 mg (1.26 mmol) coenzyme A (CoA) disodium salt powder. This reaction mixture was stirred for 4 h at room temperature in dark. The product was isolated by preparative reverse-phase HPLC (detection at 260 nm) with linear gradients from 50 mM ammonium acetate, pH 7 to acetonitrile on a SunFire C18 column (100 Å, 5 µm, 4.6x250 mm

column, Waters Cat. # 186002560) at a flow rate of 1 ml/min. Fractions containing the desired product were combined and concentrated in vacuum and analyzed by positive electrospray ionization mass spectrometry (ESI-MS) for purity. Combined CoA-Cy5 was stored in 50% glycerol (v/v) at -20°C to avoid repeated freezing and thawing. The concentration of CoA-Cy5 was determined using the extinction coefficient of Cy5 (ε (648 nm) = 250,000 $M^{-1}cm^{-1}$). Positive ESI-MS (m/z) calculated for CoA-Cy5: 1546.3931 [M(+1)] and 773.7002 [M(+2)], found 1546.3926 [M(+1)] and 773.6984 [M(+2)] (**Figure 3.7**).

3.2.5 Labeling of Gy-hMBD

To prepare Cy5-labeled Gy-hMBD, a 500 µl reaction mixture of 12 µM CoA-Cy5 and 5 µM Gy-hMBD in 50 mM HEPES pH 7.4 and 10 mM $MgCl_2$ was catalyzed by 1 µM Sfp synthase (NEB, discontinued) for 2 h at 37°C in dark. A labeling efficiency was estimated to be around 24% by running a denaturing PAGE (data not shown here). However, any attempts to purify Cy5-Gy-hMBD failed due to aggregation as discussed in "Results" section.

3.2.6 Cloning, expression and purification of Halo-hMBD

To generate the construct for expressing Halo-hMBD, a HaloTag-containing vector pBD003_mut_VCP (R155H) was gifted to us from Stephanie Moon's lab in the Human Genetics Department at the University of Michigan. To fuse hMBD1 MBD (aa 1-77) with a C-terminal HaloTag using Gibson assembly, we first PCR linearized pBD003_mut_VCP (R155H) with a pair of backbone primers: Halo-hMBD-BF and Halo-hMBD-BR, followed by gel purification to remove any template plasmids. Insert containing hMBD1 MBD (aa 1-77) was then generated by PCR amplification of addgene plasmid # 119966 with a pair of insert primers: Halo-hMBD-IF and Halo-hMBD-IR, followed by gel purification to remove any template plasmids and

unwanted DNA fragments. Finally, the linearized backbone and insert were mixed in a 1:2 molar ratio and ligated using Gibson assembly (NEB, Cat. # E5510S) at 50°C for 15 min. The ligation mix was then transformed into NEB 5-alpha competent cells (NEB, Cat. # C2987H) and colonies were selected using 50 μg/ml Kanamycin LB agar plate. Their plasmid sequences were validated by Sanger sequencing.

For overexpression and purification of Halo-hMBD, its construct was transformed into BL21(DE3) competent E. coli (NEB, Catalog # C2527H) and transformed cells were spread on a 50 μg/ml Kanamycin LB agar plate. Single colonies were inoculated into a 10 ml LB culture containing 50 μg/ml Kanamycin and grown overnight at 250 rpm, 37°C. The OD600 of this overnight culture was measured and a certain amount of it was further inoculated into a 100 ml TB culture with 50 μg/ml Kanamycin for large-scale expression such that the starting OD600 was exactly at 0.01. After incubating at 250 rpm, 37°C for approximately 3-4 h, its OD600 reached 0.6 and overexpression of Halo-hMBD was induced by addition of 0.05 M IPTG right after we cooled it down in ice bath. Large expression culture was further incubated at 250 rpm 20-22°C for another 16 h. After expression, E. coli culture was spun down at 5,000 g, 4°C for 20 min and cell pellets were pooled and resuspended in a lysis buffer containing 1X Base buffer, 1X protease inhibitor cocktail, 1 mg/ml lysozyme (Millipore Sigma, Cat. # L6876-10G) and 5 mM freshly thawed β-mercaptoethanol. Otherwise, cell pellets would be flash-frozen and be stored at -80°C until purification. Note that 1X protease inhibitor cocktail was prepared by dissolving 2 tablets of cOmplete™, EDTA-free Protease Inhibitor Cocktail ((Millipore Sigma, Cat. # 11873580001) in 50 ml buffer. 2X Base buffer contained 40 mM Tris-HCl pH 8.0, 0.2% (v/v) Tween 20, 1200 mM NaCl and 20 mM imidazole and was premixed since it was a common component for all buffers. In general, 50 ml lysis buffer was used per 100 ml of TB cell culture.

Cell lysis was achieved by sonication in ice water with 5 s on and 15 s off at 70% amplitude for 20 min (total time) until cell suspension became semitransparent. Subsequently, lysate was spun down at 20,000 g, 4°C for 60 min. Supernatants were further clarified through a 0.45 μm syringe filter (Millipore Sigma, Cat. # SLHV033RS) before sample application using a gravity column (Bio-Rad, Cat. # 7372512). A 2 ml Ni-NTA resin (Qiagen, Cat. # 30210) was equilibrated with 10 ml of 1X Base buffer and 5 mM β-mercaptoethanol for 10 min by constantly rotating. Following that, filtered supernatants were incubated with resin for 60 min by constantly rotating. Halo-hMBD-bound resin was then slowly depositing into the gravity column and washed by 20 ml of 1X Base buffer and 5 mM β-mercaptoethanol. A series of step gradient of elution buffers were used containing: 20 mM imidazole, 50 mM imidazole, 80 mM imidazole, 100 mM imidazole, 150 mM imidazole and 200 mM imidazole in 1X Base buffer and 5 mM β-mercaptoethanol. Each gradient step was 10 ml and each fraction of eluate was around 5 ml. Fractions were loaded on denaturing PAGE to examine protein purity before combining them together. Combined pure Halo-hMBD were concentrated and buffer-exchanged in a storage buffer containing 20 mM Tris-HCl pH 8.0, 0.1% (v/v) Tween 20, 10% (v/v) glycerol, 300 mM NaCl and 5 mM β-mercaptoethanol. Finally, Halo-hMBD was aliquoted, flash-frozen in liquid nitrogen and stored at -80°C. All steps following cell harvest were at 4°C.

3.2.7 Labeling of Halo-hMBD

To label Halo-hMBD with HaloTag ligand Alexa Fluor 660 (AF660, Promega, Cat. # G8471), 1 μM of Halo-hMBD was mixed with 5 μM AF660 in a reaction buffer of 20 mM Tris-HCl pH 8.0, 0.1% (v/v) Tween 20, 10% (v/v) glycerol, 300 mM NaCl and 5 mM β-mercaptoethanol for 1 h in dark at 4°C. Free AF660 was removed using 10K MWCO centrifugal filter (Millipore Sigma, Cat. # UFC501024) at 12,500 g, 4°C by 4 rounds each of which was 15

min until all free dyes were removed (examined by denaturing PAGE). AF660-Halo-hMBD was estimated to have close to 100% labeling efficiency using denaturing PAGE (data not shown here). Finally, 50% (v/v) glycerol stock was prepared, aliquoted and stored at -20°C.

3.2.8 Electrophoresis mobility shift assay (EMSA)

Both polyacrylamide and agarose gel were used for EMSA to examine methyl-CpG binding activity of Gy-hMBD and Halo-hMBD. A common 10X running buffer containing 50 mM $Mg(OAc)_2$ and 400 mM Tris-HOAc pH 7.5 were premixed and used for both agarose and polyacrylamide EMSA. Both 2% agarose and 5% native PAGE were prepared in 1X running buffer. In both cases, DNA substrates and proteins were incubated in the binding buffer of 5 mM $MgCl_2$, 10% glycerol, 1 mM DTT and 1X PBS pH 7.4 at room temperature for 2 h. A 5X gel loading buffer was prepared by dissolving 0.0625% (w/v) Bromophenol Blue and 0.0625% (w/v) Xylene cynaol FF in 5 ml 10X running buffer and 5 ml glycerol. Following 2-h incubation, samples were mixed with 5X loading buffer in a volume ratio of 4:1. Gel running buffer was prepared by diluting 10X running buffer to 1X and precooled to 4°C along with casted gels. The entire run of electrophoresis was kept in 4°C to avoid overheating. 2% agarose was running at approximately 7 V/cm for 3 h and 5% native PAGE was running at approximately 15 V/cm for 3 h.

3.2.9 MBD-SiMREPS assay protocol

Sample cells made of cut P20 pipette barrier tips were attached to glass coverslips passivated with a 1:100 mixture of biotin-PEG and mPEG. A detailed protocol of slide preparations is discussed in previously published papers[81]. Sample cells were first washed with T50 buffer (10 mM Tris-HCl [pH 8.0 at 25°C], 50 mM NaCl) and then incubated with 40 µl 0.25

mg/ml streptavidin in T50 buffer for 10 min. Following a wash with 1X PBS for 3 times, 100 nM capture probe in 1X PBS that was preheated at 90°C for 5 min in a metal bath, annealed at 37°C for 5 min in a water bath, and cooled down to room temperature, was then added to the sample well. The sample well was incubated for 10 min and washed with 4X PBS for 3 times waiting for the target strand. A mixture of sensor components was prepared in a PCR tube that contained 10 nM auxiliary probe and 10 pM targets in 4X PBS / 2 μM poly-T oligodeoxyribonucleotide (dT10) carriers. All dilutions of targets were performed in the presence of 2 μM dT10 in GeneMate low-adhesion 1.7 mL microcentrifuge tubes (VWR, Cat No. 490003-230). PCR tubes that contained sensor components including targets were then heated at 80°C for 3 min, annealed at 64°C for 5 min, subsequently 57°C for 5 min and cooled down at 38°C for another 5 min and finally held at 22°C. This sensor assembly process was performed in a thermocycler. The sensor construct that was properly assembled was added to the sample cell and then incubated for 1 h at room temperature. After target capture, sample cells were washed 3 times with 4X PBS followed by one-time wash of 50 mM Tris-HCl pH 8.0. 50 μl imaging buffer containing the desired concentration of AF660-Halo-hMBD in the presence of an oxygen scavenger system (OSS) — 1 mM Trolox, 5 mM 3,4-dihydroxybenzoate (PCA), 50 nM protocatechuate dioxygenase (PCD) — was added and then imaged by objective-TIRF microscopy. 1 μM PCD stock was prepared in 100 mM Tris-HCl pH 8.0, 50 mM KCl, 1 mM EDTA, 50% glycerol; 100 mM PCA was dissolved in water and titrated with 5 M KOH to a pH of 8.3; Trolox was dissolved in water and titrated with 5 M KOH to a pH around 10-11. All three components are stored in -20°C prior to use.

3.2.10 Single-molecule fluorescence microscopy

All single-molecule experiments were performed on the Oxford Nanoimager (ONI), a compact benchtop microscope capable of objective-type TIRF (See https://oni.bio/nanoimager/ for spec sheet regarding camera, illumination and objective). A 100X 1.4NA oil-immersion objective was installed on ONI together with a built-in Z-lock control module for autofocus. Since the built-in temperature control system on ONI could not keep imaging temperature below 25°C, to avoid overheating by turning laser on for too long, we attached the outer box of ONI to a metal clamp where circulating cold water coming from a water bath could run through. To maintain an imaging temperature of 22°C, water batch was kept at 16°C given a room temperature of 22°C. For recording AF660 fluorescence emission with optimal signal-to-noise ratio (S/N), samples were excited at 640 nm with 20% laser power (approximately 30 mW) at an illumination angle of around 54° (note that this "illumination angle" shown on ONI was not actually the incident angle. The relationship between illumination angle and incident angle was not clear to us.). The signal integration time (exposure time) per frame was 100 ms unless otherwise noted, movies of 5 min were collected per field of view (FOV).

3.2.11 Processing and analysis of objective-TIRF type data

A set of custom MATLAB codes were used to identify spots with significant intensity fluctuation within each FOV, generate intensity-versus-time traces at each spot, fit these traces with two-state hidden Markov modeling (HMM) algorithm to generate idealized traces, and eventually identify and characterize transitions with idealized traces. A set of filtering criteria were generated to distinguish methylation-specific signal and non-specific signal by feeding traces from no target control experiments as negative dataset and traces from methylated target experiments as positive dataset into a SiMREPS optimizer (see **Table 3.2**). A detailed discussion of data analysis pipeline can be found in papers previously published in our group[81].

Table 3.2. Optimized parameter sets for trace generation and analysis. See **Table 1.1** for detailed description of each parameter.

Trace Generation Parameters	
use fluctuation map?	0
Stdfactor	3.5
start frame	1
end frame	3000
edgePx	20
Percentilecut	0.95
ROI size (pixels)	5

Trace Analysis Parameters (KFC)	
start frame	1
end frame	3000
exposure time (s)	0.1
Smoothframes	1
remove_single_frame_events	FALSE
Ithresh	1000
SNthresh	2
SNthresh_trace	1.4
min_Nbd	2
max_Nbd	Inf
min_tau_on_median (s)	0.1
min_tau_off_median (s)	0
max_tau_on_median (s)	0.5
max_tau_off_median (s)	44
max_tau_on_cv	Inf
max_tau_off_cv	Inf
max_tau_on_event (s)	Inf
max_tau_off_event (s)	106
max_I_low_state	Inf
vary_I_vals	FALSE
num_intensity_states	2
ignore_post_bleaching	FALSE
bleaching_wait_time (s)	Inf
use_FRET_threshold	FALSE
FRET_threshold	0

3.3 Results

3.3.1 Engineering and characterization of Gy-hMBD

Ohki *et al.* first solved the solution NMR structure of a hMBD1 MBD domain (aa 1-75) in complex with methylated dsDNA in 2003 and Liu *et al.* solved the crystal structure of hMBD1 MBD domain (aa 1-77) in complex with methylated dsDNA in 2018[130,131]. They both characterized the MBD's methyl-binding affinity. We decided to choose the GST-tagged MBD used by Liu *et al.* since its plasmid is readily available on Addgene and the optimized purification protocol of a similar construct, MBD-GFP (mMBD, aa 1-75) was well documented by Boyd *et al.*[132] Note that the MBD we chose originated from hMBD1, with the highly conserved MBD sequence of the MBD superfamily and a few mutations and small length

variations.

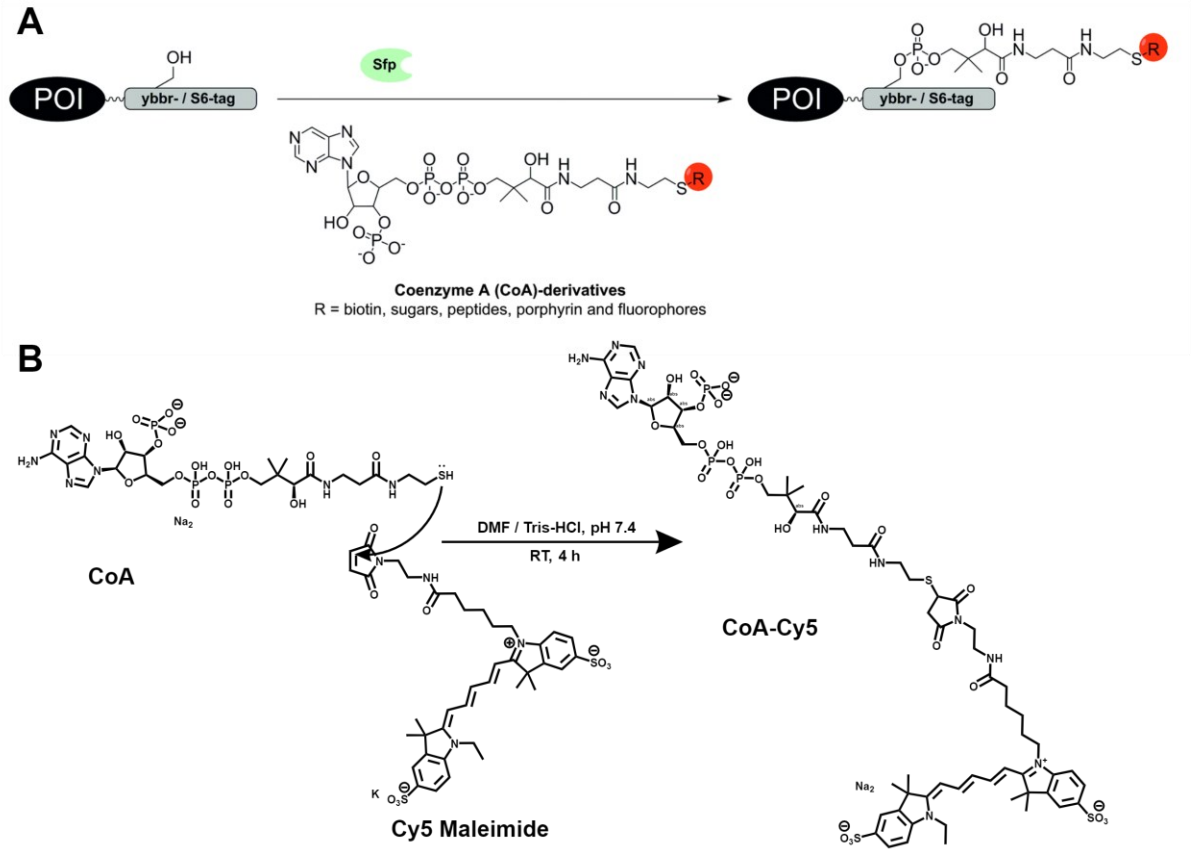

Figure 3.4. Site-specific labeling using ybbR tag. (A) Schematic of ybbR labeling reaction catalyzed by a phosphopantheteinyl transferase (PPTase) Sfp. POI: protein of interest. Figure taken from Lotze *et al.*[133] (B) Synthesis of CoA-Cy5.

For site-specific labeling, a 11-residue peptide tag, DSLEFIASKLA (ybbR tag), was inserted in between 6xHis and hMBD as demonstrated in **Figure 3.5**. The second hydroxyl group within the ybbR tag attacks a CoA derivative where a functional group is linked by a sulfide bond as illustrated in **Figure 3.4**. This functional group is an organic fluorophore in our case for the purpose of labeling. **Figure 3.4**B shows the diagram of synthesizing CoA-Cy5, serving as the substrate for the designed enzymatic labeling reaction. ybbR tag labeling strategy offers multiple benefits. First, due to its compact size and simple single-step reaction, ybbR tag labeling

minimizes the risk of interfering methyl-binding activity. Second, the labeling ratio is guaranteed to be 1:1, avoiding a multi-level fluorescence intensity. Third, this approach results in covalent attachment, thus making this reaction irreversible and its resulting product chemically stable. Last but not least, unlike fluorescent proteins, organic dyes incorporated by ybbR labeling are photostable and exhibit superior brightness, well suited for single molecule observation.

Figure 3.5B also shows the structure of the resulting protein, termed Gy-hMBD, as predicted by AlphaFold2 using colabfold. During expression and purification of Gy-hMBD, we also observed severe nucleic acid contamination since we did not introduce nuclease in the process of purification (**Figure 3.5**D). Subsequent heparin affinity chromatography resolved this issue completely and also completely removed any impurities during the last step (**Figure 3.5**E&F). Nucleic acid contamination is problematic for downstream application for two reasons. First, the additional absorbance by nucleic acids at 280 nm skewed the protein concentration measurement. Second, the presence of nucleic acids prevented us from observing methyl-binding activity since an electrophoretic gel shift was observed regardless of addition of methylated DNA target (data not shown here).

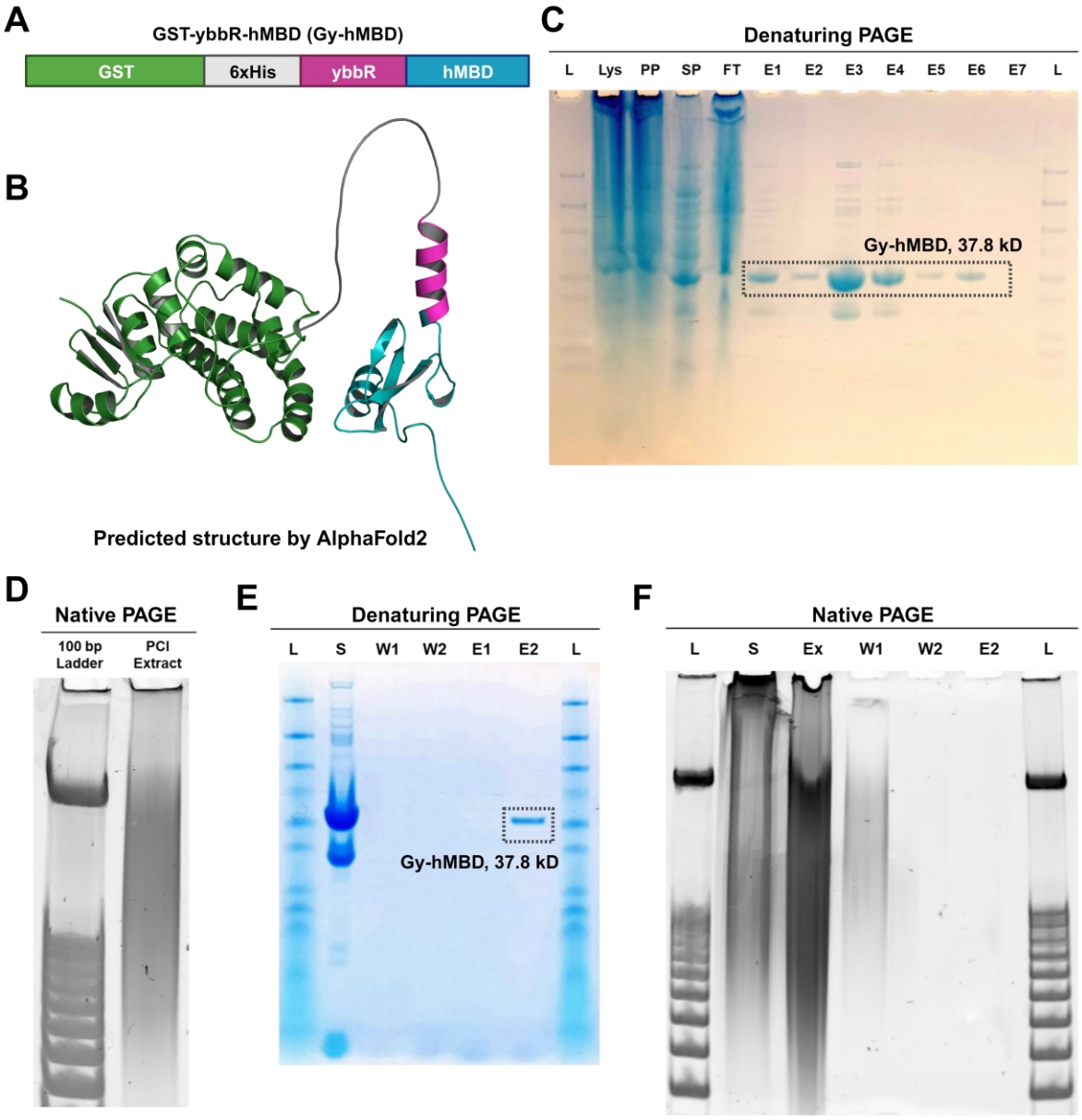

Figure 3.5. Engineering, purification and characterization of Gy-hMBD. (A) Construct of GST-ybbR-hMBD (Gy-hMBD). A thrombin protease cleavage site is inserted in between ybbR tag and 6xHis tag. Thus, all purification tags can be removed in case they interfere with methyl-CpG binding. (B) Predicted structure of Gy-hMBD by AlphaFold2 using colabfold. All settings are default. (C) 4%-12% gradient denaturing PAGE, running profile of purification of Gy-hMBD using glutathione affinity chromatography. An impure protein band right below Gy-hMBD was coeluted. All fractions E1-7 are combined and concentrated. L: ladder, Lys: lysate, PP: precipitate, SP: supernatant, FT: flowthrough, E1-E7: eluate #1 to #7. (D) Native 5% PAGE to examine nucleic acid contamination after purification in panel C using phenol-chloroform-isopropanol extract (PCI Extract) of combined fractions in panel C. Stained by SYBR gold and visualized using Cy2 illumination. (E) 4%-12% gradient denaturing PAGE, running profile of heparin affinity chromatography purification of combined fractions in panel C to remove nucleic acid contamination. A pure band of Gy-hMBD is shown in lane E2. L: ladder, S: samples of combined fractions, W1-2: wash #1 to #2, E1-2: eluate #1-2. (F) Native 5% PAGE to examine nucleic acid contamination in heparin

affinity chromatography purification in panel E. Stained by SYBR gold and visualized using Cy2 illumination. L: 100 bp DNA ladder, S: same as S in panel E, Ex: PCI extract of S, W1-2: same as in panel E, E2: same as in panel E. No nucleic acid existed after elution as shown in lane E2.

3.3.2 Aggregation caused by ybbR labeling of Gy-hMBD

After successful purification of Gy-hMBD without significant nucleic acid contamination, we examined its methyl-CpG binding activity by mixing with one of three different dsDNA substrates and looking for a gel shift. As expected as in **Figure 3.6**B, a gel shift primarily occurred in the presence of symmetrically methylated dsDNA (SM). And the amount of additional band representing Gy-hMBD/dsDNA complex depends on the molar ratio of Gy-hMBD and dsDNA substrate. Note that we also observed a weak band in the presence of hemimethylated dsDNA (HM), suggesting a weaker affinity towards HM compared to SM. We also observed a tendency of dissociation in the unmethylated dsDNA substrate (UM) and HM in the presence of Gy-hMBD but not in the case of SM, suggesting a destabilizing effect of Gy-hMBD binding to the unmethylated dsDNA, likely due to transiently twisting or bending of the helix structure.

Once obtaining functional Gy-hMBD, CoA-Cy5 was synthesized, HPLC-purified and characterized using mass spectrometry (**Figure 3.7**). The expected distribution of mass/charge ratio suggested a correct structure of CoA-Cy5. Subsequently, we labeled Gy-hMBD with CoA-Cy5 and examined its methyl-CpG binding activity without removing free CoA-Cy5. Surprisingly, we observed a gel shift in the presence of all substrates, UM, HM and SM (**Figure 3.6**C). The middle red lane in the gel is the loading dye. Many attempts were made to remove free dyes including both gravity column and prepacked column with either glutathione affinity chromatography or heparin affinity chromatography. However, right upon sample application, labeled species became stuck to the resin and we were no longer able to elute them out. Only 0.1

M NaOH was able to wash them off. Eventually, we discovered an immediate formation of blue precipitates after mixing Gy-hMBD and CoA-Cy5. These blue aggregates were Cy5-labeled Gy-hMBD (Cy5-Gy-hMBD) after examining on a denaturing gel (data not shown here) and we were not able to redissolve them back in a native solution except when using 0.1 M NaOH (data not shown here). One hypothesis for explaining aggregation is that the labeling site on ybbR tag might be too close to the hydrophobic residue on MBD that recognizes methyl-CpG. Hydrophobic interactions between Cy5 moiety and residues on MBD might misfold MBD and cause aggregation. This misfolded MBD then would no longer exhibit methyl-CpG binding activity but just non-specific nucleic acid binding as shown in **Figure 3.6**C.

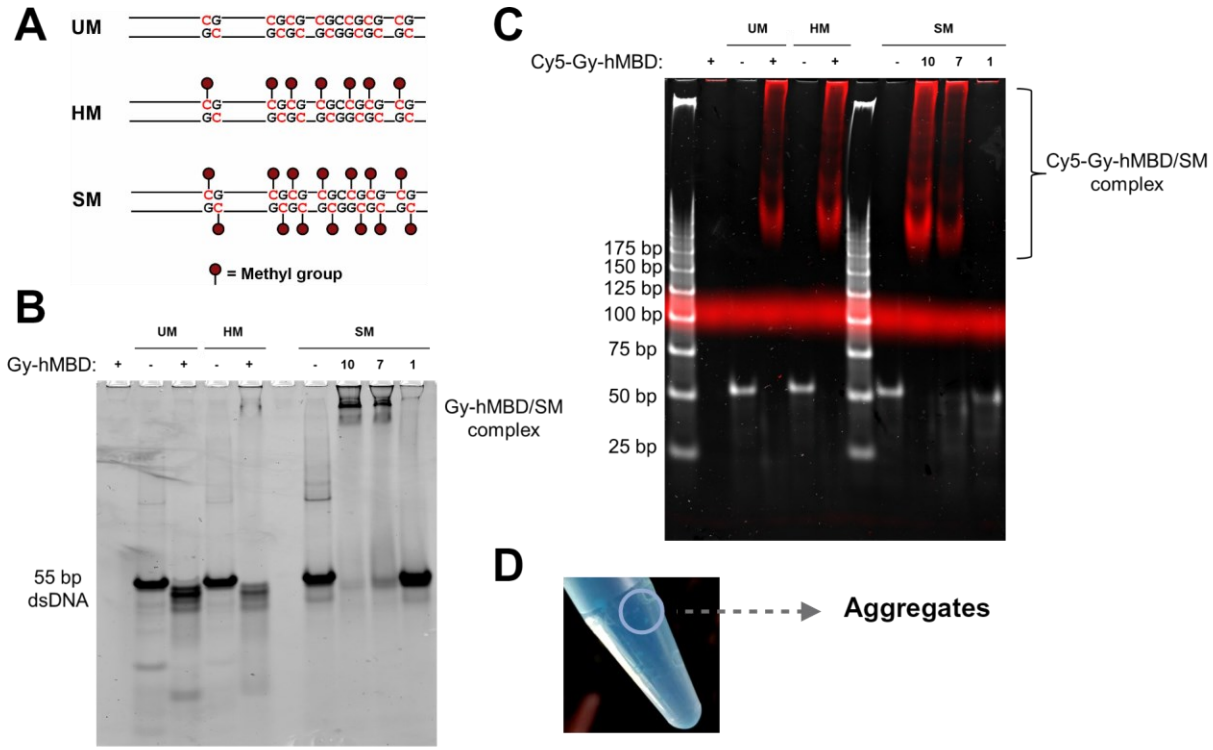

Figure 3.6. Functional assay of Gy-hMBD by EMSA before and after labeling. (A) Three substrates used for EMSA. UM: unmethylated, HM: hemimethylated, SM: symmetrically methylated. All sequences are 55 bp BCAT1 promoter. (B) EMSA in 2% agarose gel for unlabeled Gy-hMBD, stained by SYBR-Au and visualized by Cy2 illumination. Gy-hMBD is always at 10-fold concentration of dsDNA substrates except in lanes containing SM. Molar ratio of Gy-hMBD to SM is 10, 7 and 1:1, respectively. (C) EMSA in 5% native PAGE for Cy5-labeled Gy-hMBD (Cy5-Gy-hMBD), with DNA stained by SYBR-Au and visualized by Cy2 illumination (for the DNA stain) and Cy5 (for the protein stain) illumination. Molar ratio follows the same loading order as in panel B. (D) Picture of

labeling reaction mixture after reacting for 30 min. Blue foliated precipitates can be visually observed in the deep violet circle.

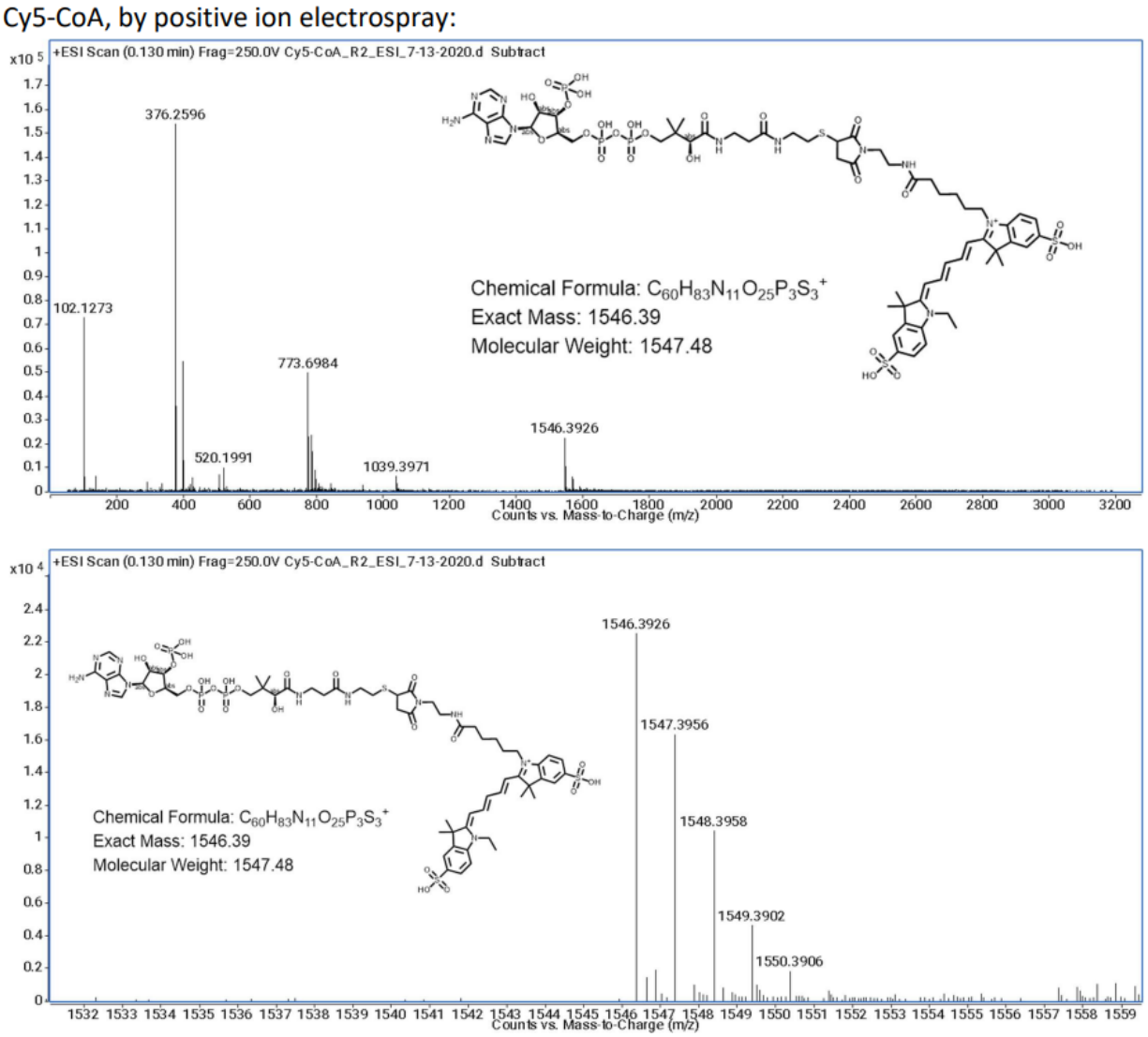

Figure 3.7. Electrospray ionization mass spectrum (ESI-MS) of HPLC-purified CoA-Cy5.

3.3.3 Engineering and characterization of Halo-hMBD

Due to the aggregation issue encountered during labeling Gy-hMBD, we decided to switch to a more established and widely used labeling approach, HaloTag labeling. The HaloTag is a modified haloalkane dehalogenase designed to covalently couple to haloalkane-derivative

ligands and was first reported by Los *et al.* in 2008[134]. This one-step covalent labeling approach soon democratized due to its fast and irreversible reactivity as well as great flexibility of incorporating a diverse pool of ligands. Nowadays, HaloTag labeling has become a common approach for in vitro, in situ and in vivo imaging[135].

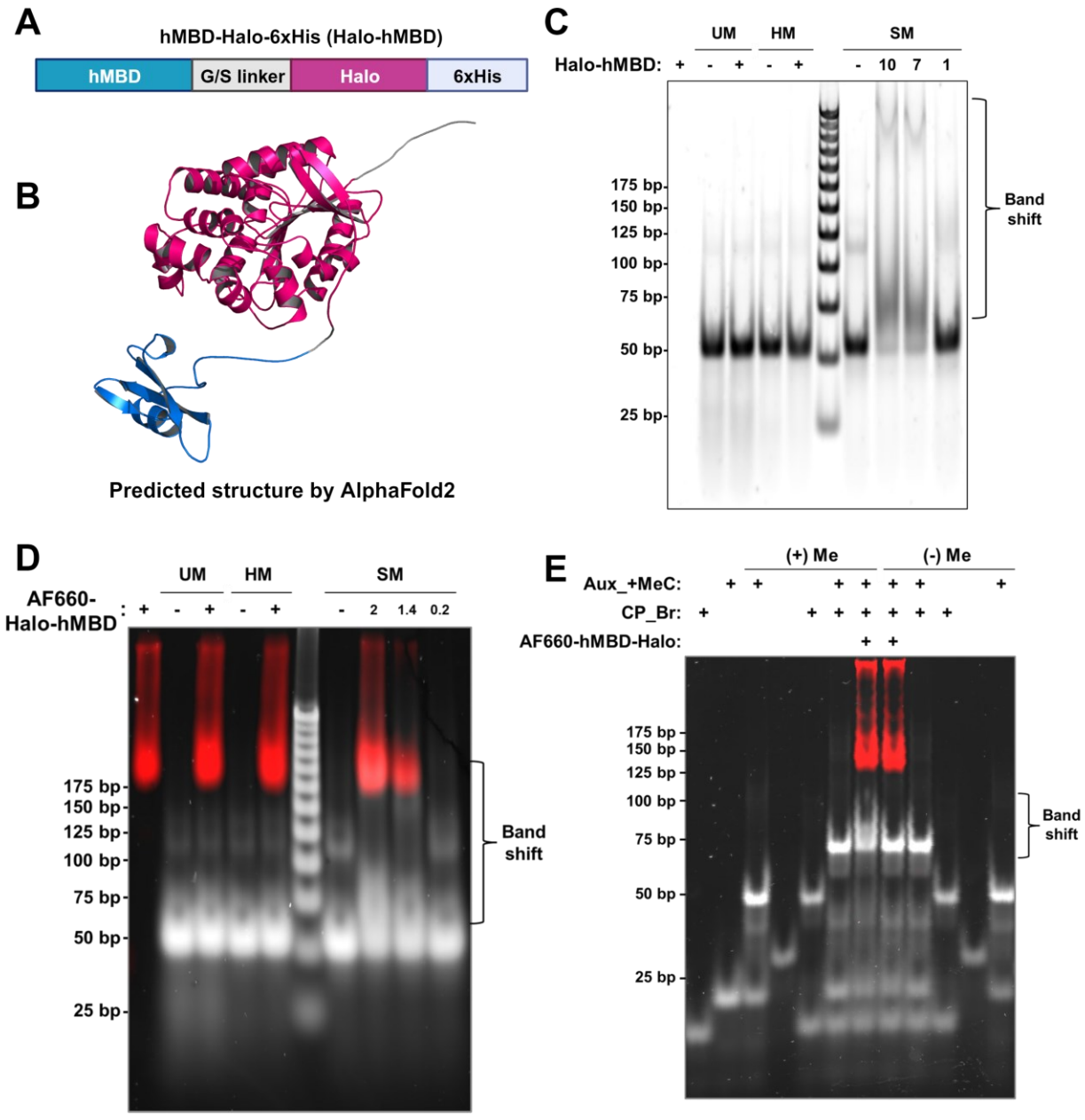

Figure 3.8. Construct and functional assay of Halo-hMBD. The same dsDNA substrates, UM, HM and SM, are used for panel C-E as in **Figure 3.6**A. (A) Construct of hMBD-Halo-6xHis (Halo-hMBD). (B) Structure of Halo-hMBD as predicted by AlphaFold2 using colabfold. (C) EMSA in 5% native PAGE for unlabeled purified Halo-hMBD,

stained by SYBR-Au and visualized by Cy2 illumination. Halo-hMBD is always at 10-fold concentration of dsDNA substrates except in lanes containing SM. Molar ratio of Halo-hMBD to SM is 10, 7 and 1 respectively. (D) EMSA in 5% native PAGE for Alexa Fluor 660 (AF660) labeled purified Halo-hMBD (AF660-Halo-hMBD), stained by SYBR-Au and visualized by Cy2 illumination and Cy5 illumination. Free AF660 was removed and Halo-hMBD is always at 2-fold concentration of dsDNA substrates except in lanes containing SM. Molar ratio of Halo-hMBD to SM is 2, 1.4 and 0.2 respectively. (E) EMSA in 5% native PAGE for AF660-Halo-hMBD, stained by SYBR-Au and visualized by Cy2 illumination and Cy5 illumination. Free AF660 was removed and AF660-Halo-hMBD is at 2-fold concentration of assembled sensor construct. (+) Me: BCATa_+MeC_55, (-) Me: BCATa _55, See **Table 3.1** for description of Aux_+MeC and CP_Br.

Figure 3.8A&B shows the construct of hMBD-Halo-6xHis (Halo-hMBD) as well as the structure predicted by AlphaFold2 using colabfold. Note that a G/S linker was introduced in between hMBD and HaloTag to avoid potential interference caused by HaloTag on hMBD's methyl binding activity. Purification of Halo-hMBD was performed using immobilized metal affinity chromatography (IMAT) and nucleic acids were completely removed in the presence of high NaCl concentrations throughout the purification process, validated by SYBR-Au-stained native gel (data not shown here). Methyl-CpG binding activity of pure unlabeled Halo-hMBD was examined using an EMSA as shown in **Figure 3.8**C where a band shift was only observed in the presence of SM. However, compared to Gy-hMBD in **Figure 3.6**, a smearing band occurred on a gel instead of a distinct band of stable Halo-hMBD/SM complex. And the degree of smearing increased as more Halo-hMBD was added. This suggested dissociation of Halo-hMBD/SM complex once it entered the gel. In other words, this suggests that the dissociation rate constant k_{off} of Halo-hMBD/SM was relatively fast compared to Gy-hMBD/SM, which significantly benefits single-molecule observation. **Figure 3.8**D demonstrates the methyl-binding activity of AF660-Halo-hMBD. As expected, we only observed two bands in the presence of SM at different molar ratios. No distinct band appeared with a migration slower than both AF660-Halo-hMBD and SM. Once again, the smearing effect was the outcome of methyl-CpG binding activity of AF660-Halo-hMBD. We further tested whether our assembled sensor construct also

showed methyl-CpG binding activity (**Figure 3.8**E). Only when both target sequence and the auxiliary probe were methylated, we were able to observe a smearing band, proving methyl-CpG binding specificity towards our assembled construct using CP_Br and Aux_+MeC (**Table 3.1**).

3.3.4 Methyl-CpG binding activity observed at the single-molecule level

Ensemble-level methyl-CpG binding activity of labeled and unlabeled Halo-hMBD were support by our EMSA results. The exact sensor construct to be used in MBD-SiMREPS assay also showed methyl-binding activity only when both DNA strands were methylated. Therefore, we decided to use the same sensor and test whether we could observe binding and dissociation of AF660-Halo-hMBD to methylated CpG clusters at the single-molecule level. Initially, we used a similar imaging condition to the one used in EMSA where 5 mM Mg^{2+} and 1X PBS were present. However, we observed transient interactions for all three dsDNAs: SM (symmetrically methylated sensor construct), HM (hemimethylated sensor construct) and NTC (no-target control) (data not shown here) (**Figure 3.9**A). NDC (no-DNA control) was the only case where close to 0 traces behaved with repeated binding and dissociation (data not shown here). After a thorough investigation and optimization of imaging conditions (data not shown here), we concluded that both 5 mM Mg^{2+} and high salt would introduce significant non-methyl-CpG interaction with dsDNA itself, superposing methyl-CpG binding events. Therefore, we decided to switch to a simple, low-ionic strength 50 mM Tris-HCl imaging buffer without any Mg^{2+} or additional alkali metal cations. One hypothesis is that both Mg^{2+} and alkali cations may facilitate folding of single-stranded capture probe that was saturating the entire surface, making them readily available binding sites for AF660-Halo-hMBD.

Eventually, we were able to observe methyl-CpG-specific interactions under a very simple imaging condition containing just 50 mM Tris-HCl, 10% (v/v) glycerol and OSS buffer

containing a low amount of Na^+ and K^+. Under these conditions, as demonstrated in **Figure 3.9** only when both auxiliary probe and target sequence were methylated rapid repeated binding and dissociation were observable at the single-molecule level. Neither HM, NTC nor NDC exhibited such transition behavior. A distinct population appeared in both N_{b+d} distribution and dwell time distribution, representing methyl-CpG interaction with the hMBD probe. These results demonstrate the implementation of a pretreatment-free and amplification-free SMFKF biosensor for direct detection of methyl-CpG cluster.

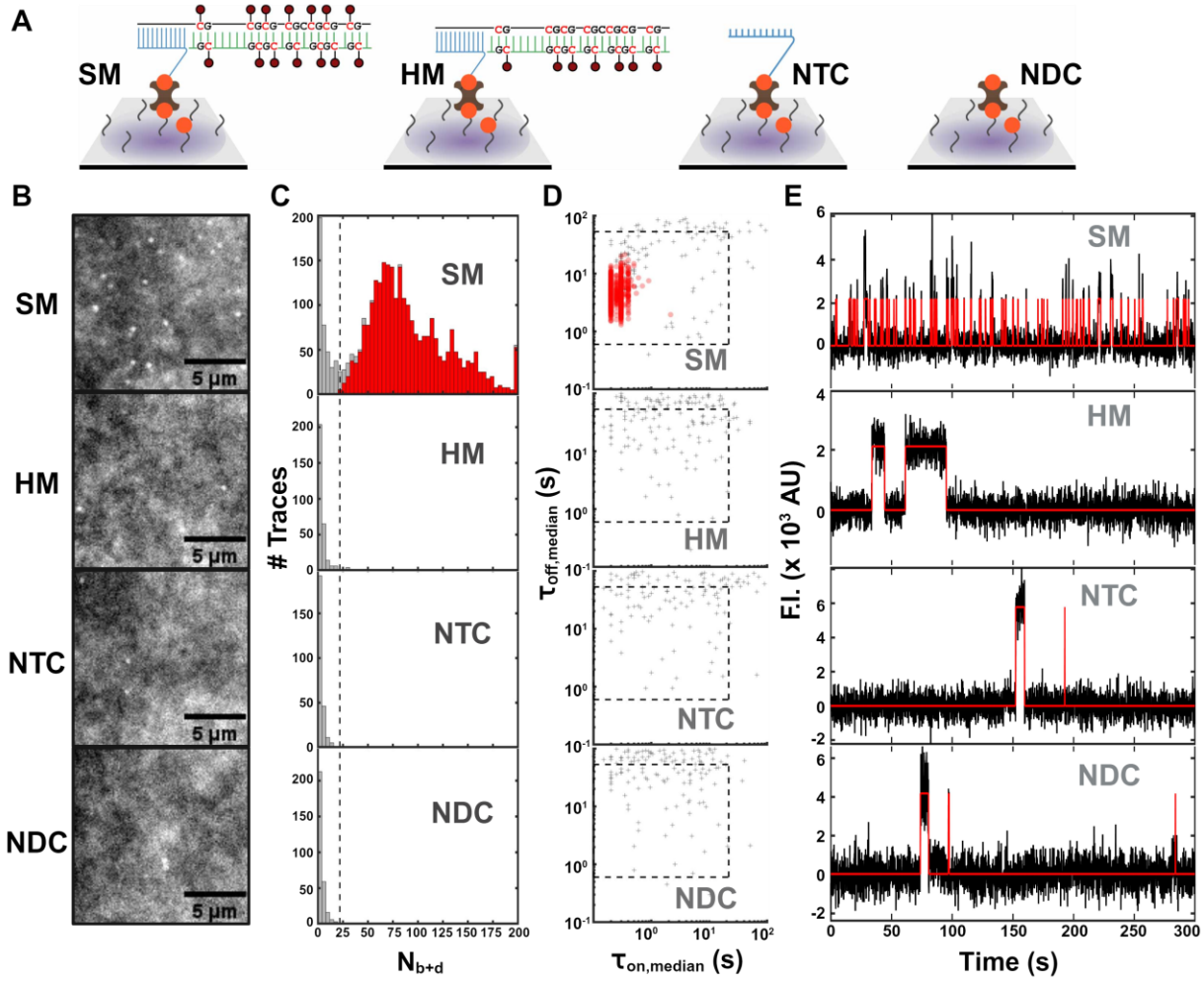

Figure 3.9. Single-molecule observation of methyl-CpG binding activity. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were optimized by selecting SM dataset against NDC and NTC datasets. (A) Assembled sensor construct under different conditions. SM: symmetrically methylated sensor where both target and auxiliary probe are methylated, HM: hemimethylated

sensor where target is unmethylated, NTC: no-target control, NDC: no DNA control. (B) Screenshots of a cropped square from the entire FOV under different conditions. (C) Distribution of N_{b+d} under different conditions. (D) Distribution of $\tau_{on,median}$ and $\tau_{off,median}$ under different conditions. (E) Representative fluorescence-intensity traces under different conditions. F.I.: fluorescence intensity, AU: arbitrary units. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

3.3.5 Effects of methyl-CpG number and "branch" motif

The very first sensor construct has two structural features: 1) all 7 pairs of CpG dinucleotides on both strands are methylated; 2) CP_Br is partially complementary to target sequence, leaving a "branch" structure motif. We hypothesized that the number of methyl-CpG and the "branch" structure motif may have effects on methyl-binding kinetics. Therefore, we designed a total of 36 constructs by changing number of methyl-CpGs as well as altering existence of "branch" motif as shown in **Figure 3.10**.

Figure 3.10. Table list of all 36 sensor constructs used for investigating effects of methyl-CpG number and "branch" motif on single-molecule methylation-specific binding kinetics.

Figure 3.11 shows the results of MBD-SiMREPS using 6 different single-branch sensor constructs with different numbers of methyl-CpGs. **Figure 3.11**B shows a significant left shift in the N_{b+d} distribution and a huge decrease in the number of accepted traces for [construct]. **Figure 3.11**C&D suggests that an increase in τ_{off} with fewer methylation sites caused a smaller N_{b+d}. A seemingly worse signal-to-noise ratio contributed to fewer accepted counts but could not fully explain it. It seemed that the overall "accessibility" of methylated DNA became worse with the decreased number of methyl-CpGs. Another important observation is that in the case of the fully methylated sensor, we always saw a high-intensity level state as shown in **Figure 3.11**D, suggesting that multiple AF660-Halo-hMBD may bind to the same target molecule

within a single frame. This high-intensity state was observed much less in triple-methylated and double-methylated sensor constructs.

Figure 3.12 shows results of MBD-SiMREPS against 6 different hemimethylated single-branch target constructs with different numbers of methylation sites. This time, with unmethylated auxiliary probe, no distinct N_{b+d} population appeared that would represent methyl-CpG interaction with AF660-Halo-hMBD. A population of traces with $\tau_{off,meidan}$ around 40 s to 50 s might be the methylation-specific behavior. However, these traces did not pass through the filtering criteria. Therefore, we concluded that in a single-branch sensor, both strands require methylation to observe methylation-specific signals that are well separated from background signals or off-methylation signals.

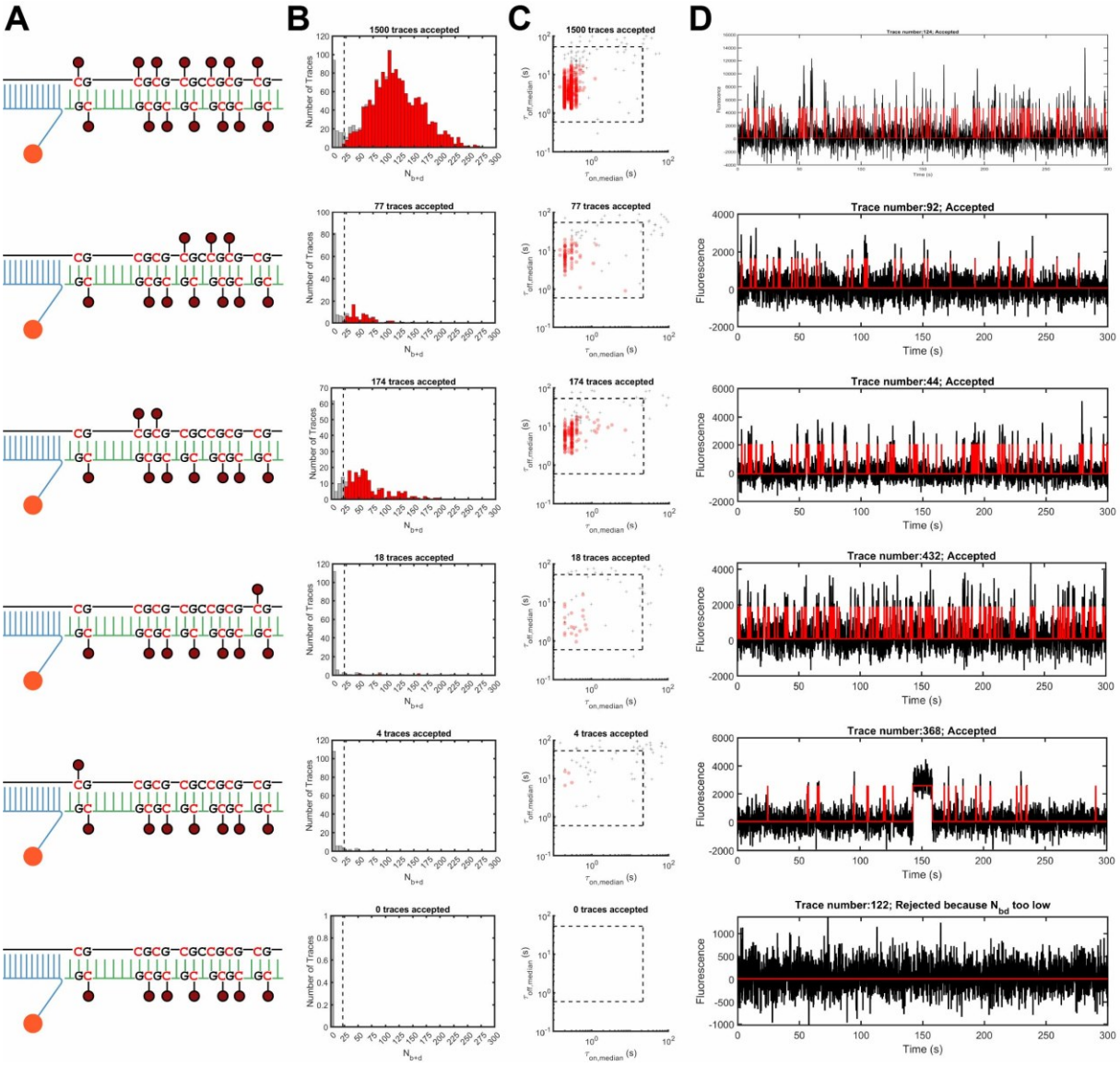

Figure 3.11. Effect of methyl-CpG number on methylation binding kinetics in a single-branch sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were optimized by selecting fully methylated SM dataset against NDC datasets. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

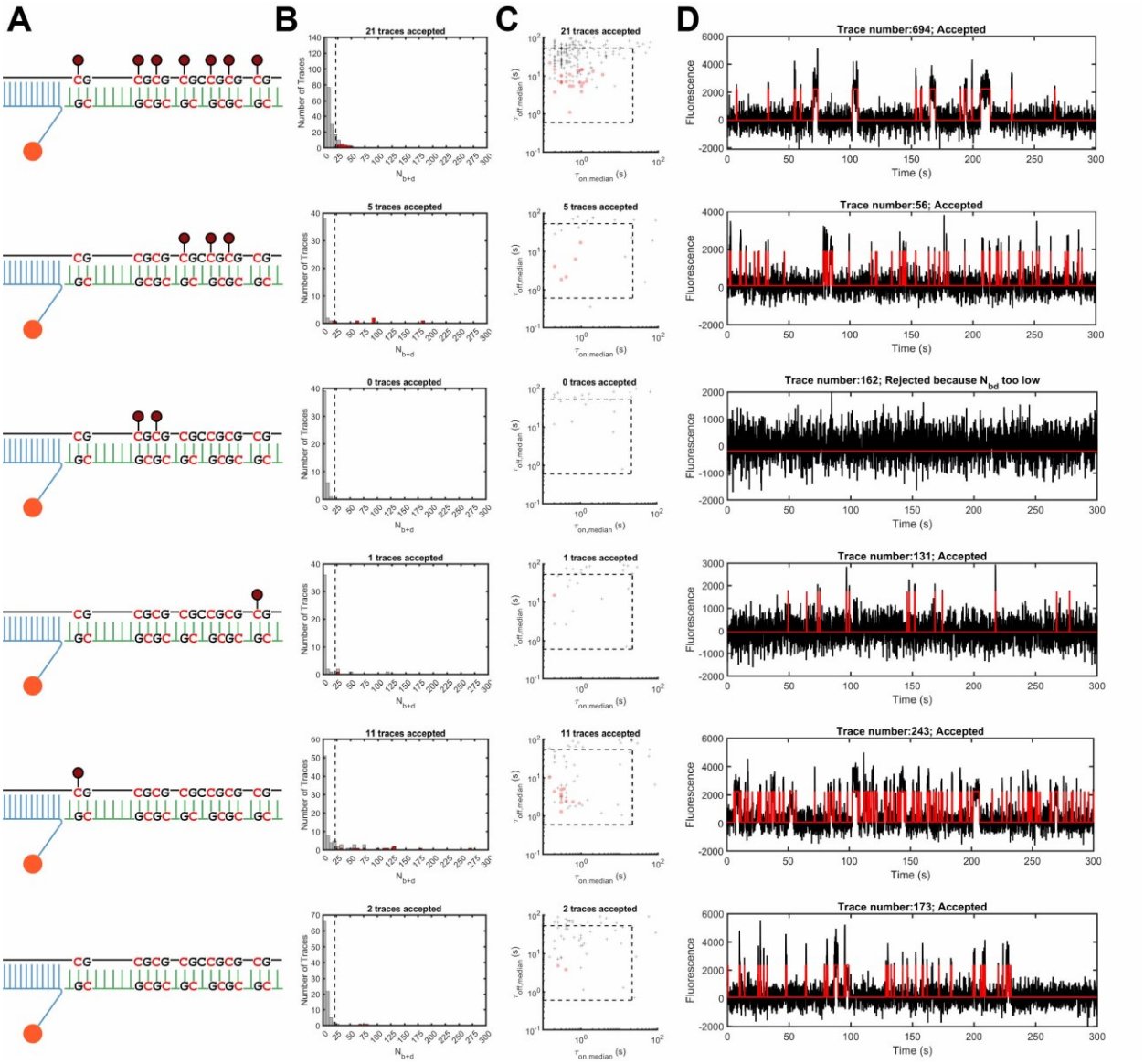

Figure 3.12. Effect of methyl-CpG number on methylation binding kinetics in a single-branch hemimethylated sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were the same as in **Figure 3.11**. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

After testing sensor constructs with a single branch, we decided to incorporate an additional "branch" motif in the auxiliary probe and tested if there would be any effects on methyl-binding kinetics. **Figure 3.13** shows the results of MBD-SiMREPS with 6 different

double-branch hemimethylated sensor constructs. Surprisingly, compared to the single-branch targets in **Figure 3.12** where no methylation-specific signals were observed, a distinct population in both N_{b+d} distribution and dwell time distribution appeared representing methylation-specific interactions. However, this population clearly had a left shift compared to the behavior with fully methylated sensor in **Figure 3.9** and **Figure 3.11**, caused by a significant increase in τ_{off}. No apparent difference in τ_{on} was observed. Therefore, we can conclude that this additional "branch" motif facilitates binding of AF660-Halo-hMBD to hemimethylated dsDNA but to a lesser extent than full methylation of the reverse strand. It is also worth noting that the forward strand still required full methylation to be accessible to AF660-Halo-hMBD, suggesting that methylation is necessary for MBD binding regardless of existence of this additional "branch" motif.

What remained to be answered is whether it is the overhang on the auxiliary probe or the overhang on the forward target sequence or both contributed to the accessibility of this sensor construct. And what is the effect of length of single-stranded sequences? Also, what is the effect of methylation site on the single-stranded region? How exactly does an individual AF660-Halo-hMBD approach and stay bound to this sensor molecule? A remotely relevant but biologically important question is whether these effects influence DNA-methylation associated gene regulation.

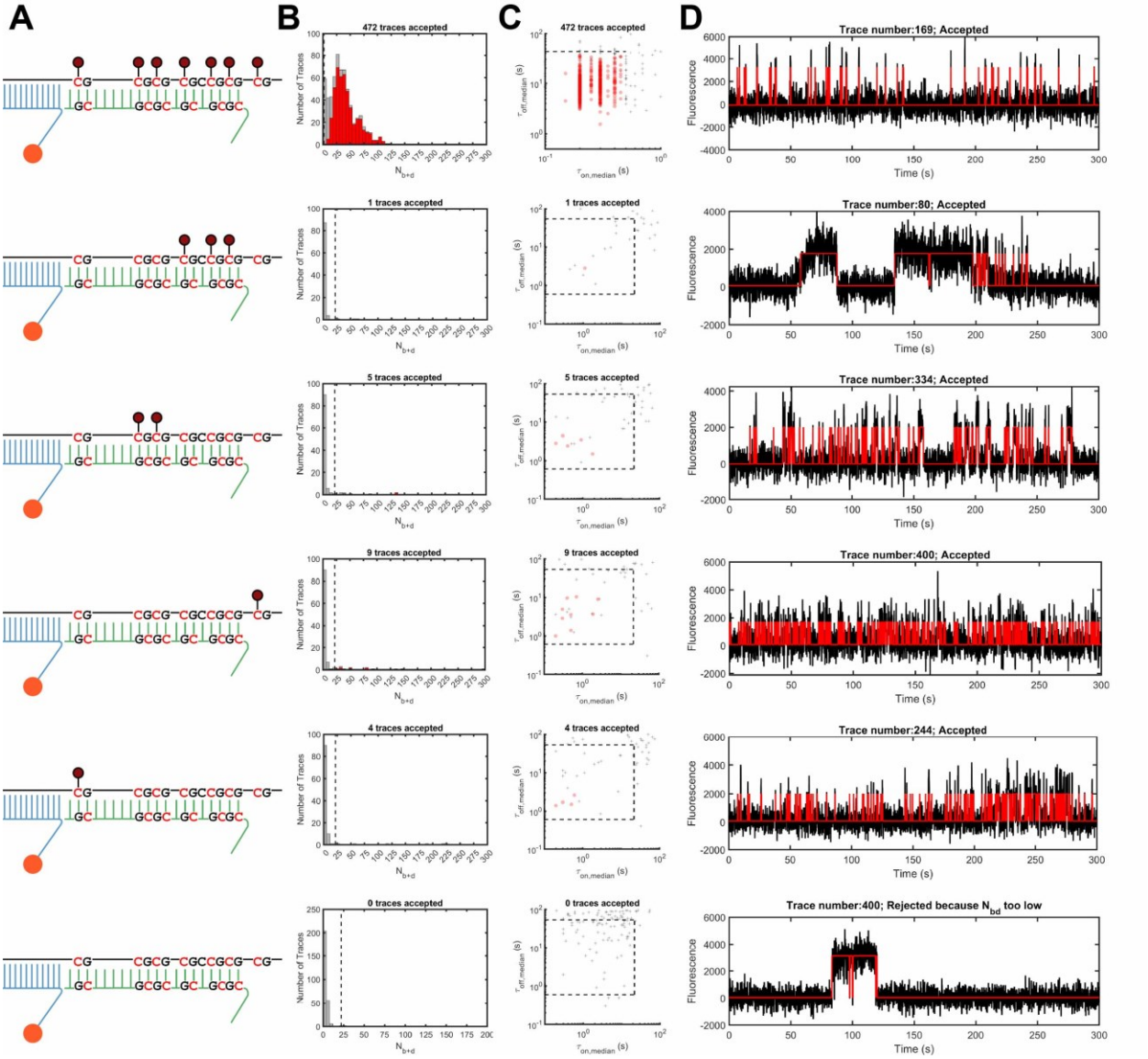

Figure 3.13. Effect of methyl-CpG number on methylation binding kinetics in a double-branch hemimethylated sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were the same as in **Figure 3.11**. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

After studying effects of the "branch" motif on methyl-binding activity, we decided to focus on number of methylation sites by switching back to a simpler design with no branch at all. **Figure 3.14** and **Figure 3.15** show the results of MBD-SiMREPS using 12 different branch-free sensor constructs. In **Figure 3.14**, auxiliary probe is also methylated. As expected, we observed

distinct populations in both N_{b+d} distribution and dwell time distribution, representing methylation-specific interactions with fully methylated, triple-methylated and double-methylated targets. All these three constructs exhibited similar binding kinetics to those with the same three targets in **Figure 3.11** but with higher number of accepted traces. It seems that the "branch" motif on capture probe did not modulate binding kinetics but changed overall accessibility. The minimum number of methylation sites on target sites for observing methylation-specific binding kinetics is 2 as illustrated in **Figure 3.14**. However, it is still not clear whether the positioning of methylation sites on the forward strand plays a role given the reverse strand is fully methylated. On the other hand, on the reverse strand, is it sufficient to keep just the forward CpGs that are symmetrical for observing methyl-binding activity? **Figure 3.15** keeps the auxiliary probe completely unmethylated and in all constructs, we did not observe any methyl-binding activity. Combining **Figure 3.14** and **Figure 3.15**, we can conclude that in a branch-free sensor construct, it is necessary to have both strands methylated for observing methyl-binding activity. However, it would be interesting to see what would happen if we further increased the number of methyl-CpGs on one strand and keep the complementary strand unmethylated. Would it be possible that, once the methyl-CpG number bypasses some value, we eventually observe methyl-binding activity?

Finally, we kept the "branch" motif in the auxiliary probe and removed the overhang in the capture probe (**Figure 3.16**). Doing this is not only because we can change the positioning of "branch" motif, but also because these two branches are fundamentally different since there are essentially two unpaired regions when keeping the "branch" motif on the auxiliary probe. Interestingly, the sensor construct with fully methylated target exhibited a distinct population in both N_{b+d} distribution and dwell time distribution, representing methylation-specific

interactions. Its distribution seems a combination of two overlapping populations where one behaved similarly to the kinetics in **Figure 3.13** (top row) and the other one featured long τ_{on} and thus small N_{b+d} although this population was not separable from background or off-methylation signals by filtering criteria and thus was rejected upon filtering. We do not have a good explanation for the newly emerging slow transitions. However, it is clear that for a hemimethylated sensor construct, a "branch" motif containing two overhangs is necessary for exhibiting methylation-specific interaction, in which case a fully methylated strand is required.

Combining all the results together, "branch" motifs with one overhang versus two overhangs seem to work in opposite ways where the single-overhang "branch" motif does not sufficiently introduce methyl-binding activity in a hemimethylated construct (**Figure 3.12**) and double-overhang "branch" motif alone is able to increase accessibility of a hemimethylated construct. Regarding effects of methyl-CpG number and pattern, a symmetrically fully methylated construct always exhibit methyl-binding activity and at least two methyl-CpGs on both strands in a symmetrically partially methylated construct are required for observing methyl-binding activity.

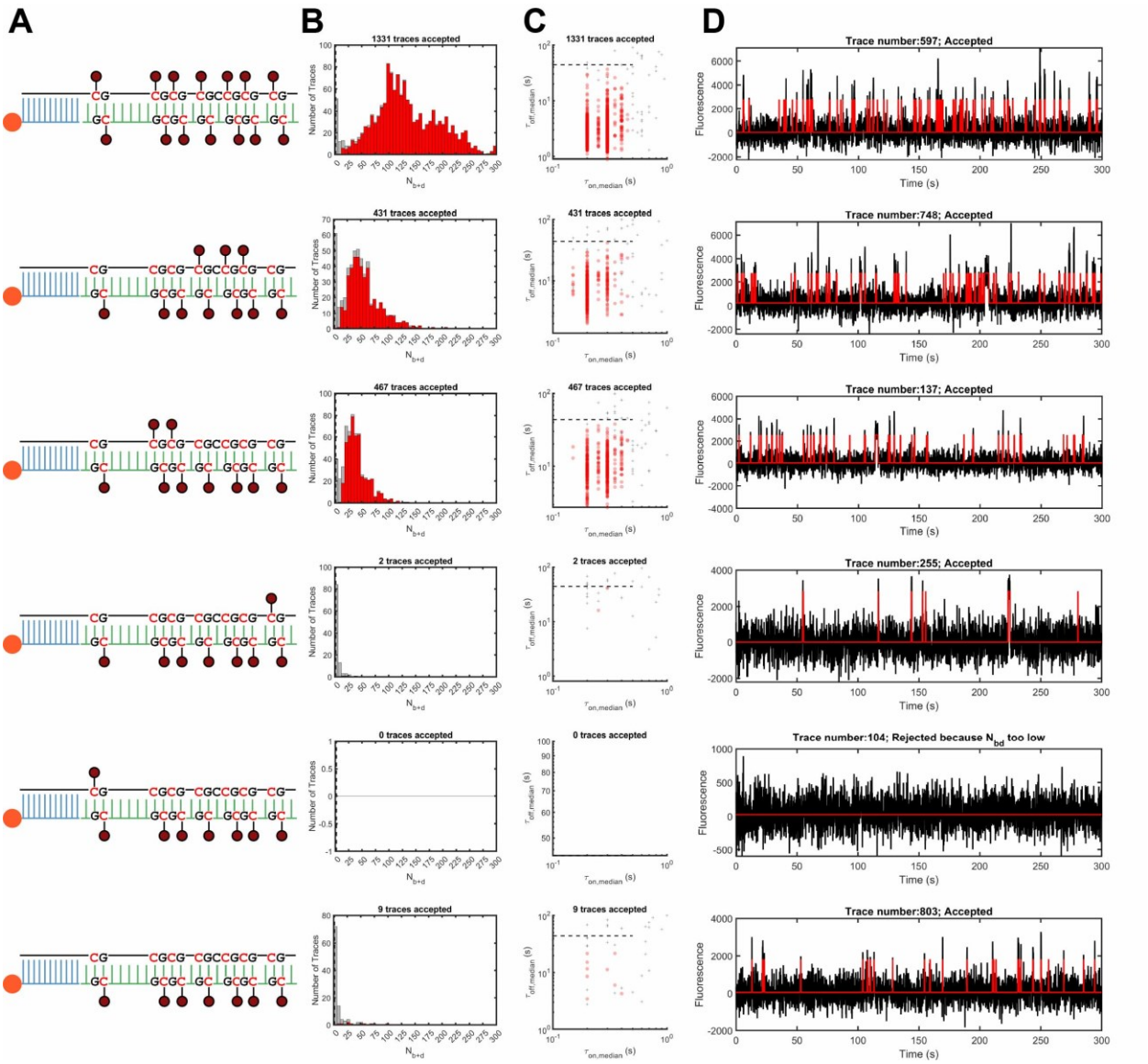

Figure 3.14. Effect of methyl-CpG number on methylation binding kinetics in a branch-free sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were the same as in **Figure 3.11**. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

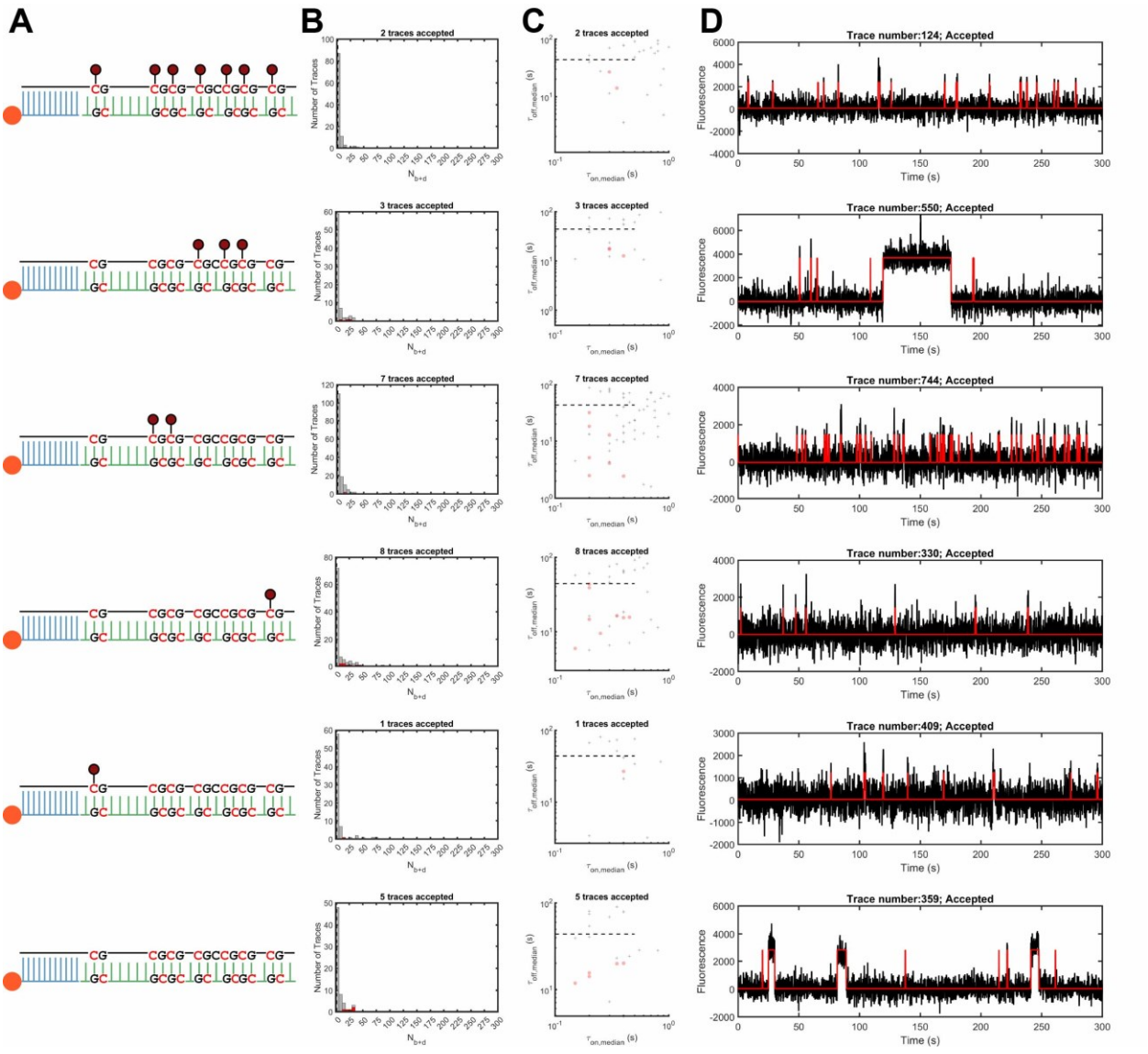

Figure 3.15. Effect of methyl-CpG number on methylation binding kinetics in a branch-free hemimethylated sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were the same as in **Figure 3.11**. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

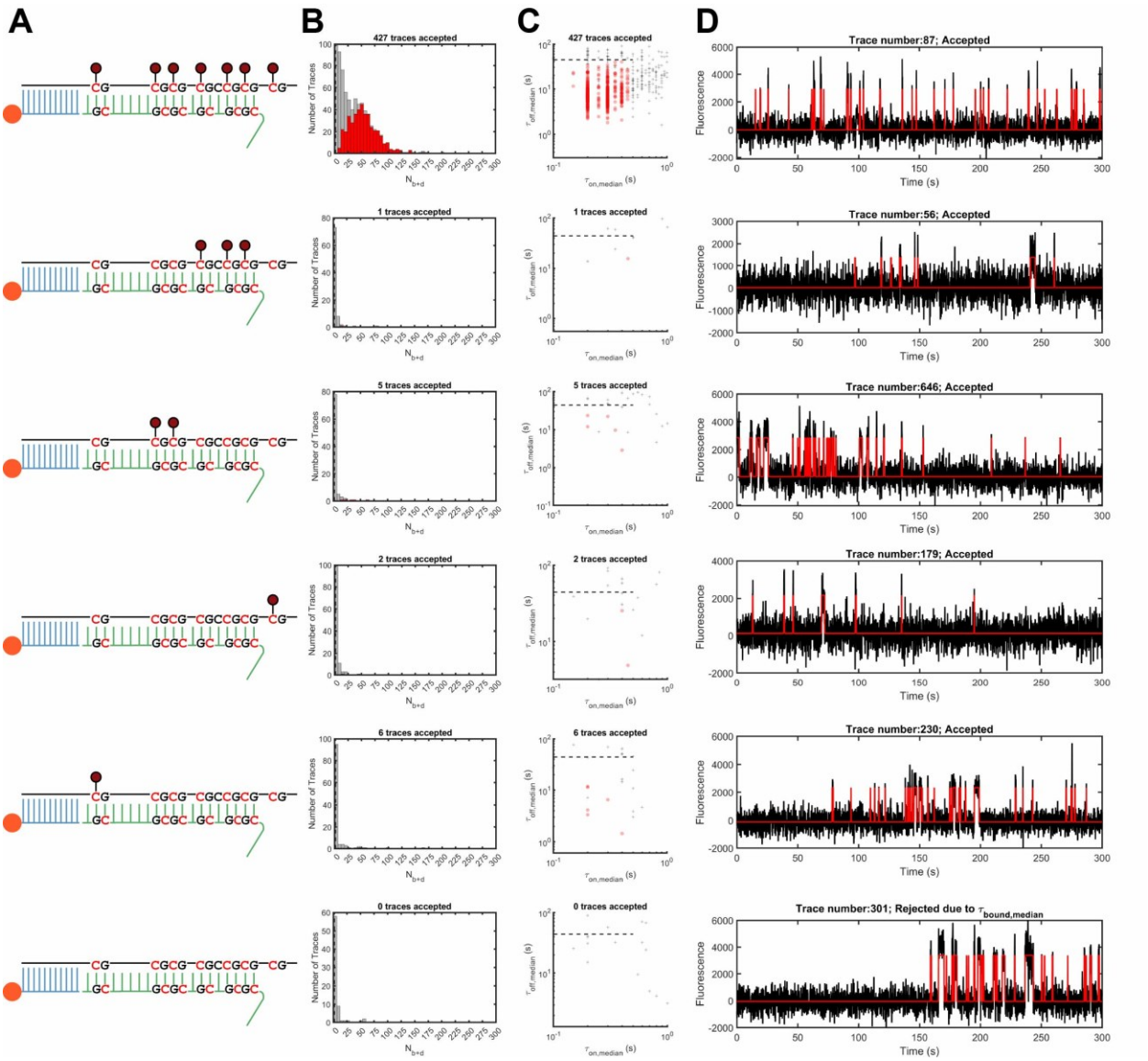

Figure 3.16. Effect of methyl-CpG number on methylation binding kinetics in a single-branch hemimethylated sensor construct. Imaging condition: 50 mM Tris-HCl pH 8.0, 10% glycerol (v/v), 50 nM AF660-Halo-hMBD, exposure time: 100 ms. Filtering criteria were the same as in **Figure 3.11**. (A) Sensor construct illustration. (B) N_{b+d} distribution using different sensor constructs. (C) Dwell time distribution using different sensor constructs. (D) Representative traces of different sensor constructs. Dark lines represent raw traces and read lines are idealized traces by HMM fitting.

3.4 Discussion

In Chapter 3, we cloned, expressed and purified two functional protein probes for methyl-CpG clusters: Gy-hMBD and Halo-hMBD, by incorporating two different labeling tags – ybbR tag and HaloTag, respectively. However, the Cy5 labeling caused irreversible aggregation of Cy5-Gy-hMBD and made it impossible to recover any functional labeled protein imager. In contrast, AF660-Halo-hMBD exhibited methyl-binding activity on both ensemble level and single-molecule level, demonstrating the first implementation of a pretreatment-free amplification-free SMFKF imager for direct detection of DNA methylation. In fact, we also tried the cell-permeable Janelia Fluor 549 (JF549) for labeling Halo-hMBD. But JF549-Halo-hMBD stuck to our streptavidin coating on the surface and the presence of non-ionic detergent like Tween20 also formed an absorbent layer for JF549-Halo-hMBD, making this labeling unsuitable for MBD-SiMREPS. This stickiness of JF549-Halo-hMBD was a property of JF549 itself since a free JF549 HaloTag ligand stuck to both streptavidin and Tween20 on the surface (data not shown here).

Through a thorough investigation and optimization of imaging conditions, AF660-Halo-hMBD became the ultimate imaging probe for MBD-SiMREPS under Mg^{2+}-free and low-salt imaging conditions. We further investigated effects of the methyl-CpG number and patterns as well as "branch" motif on methyl-CpG binding activity and gained fruitful insights on the unique binding behavior of MBD on methyl-CpG clusters. To our knowledge, this is one of several pioneering reports studying methyl-binding kinetics of MBD at the single-molecule level[136–139].

However, this research was just the initial attempt of systematically studying effects of DNA structure and methylation profile on MBD-binding kinetics. Our results of 36 different constructs only opened more intriguing questions including but not limited to positioning of

"branch" motif, positioning of methyl-CpGs and binding mode of MBD at different methylation and DNA structural contexts. Furthermore, how do these subtle changes on methyl-binding activity influence proteins of the MBD superfamily or specifically the methyl-CpG readers, the MeCP2_MBD group in regulation of gene expression? Such questions are to be answered both in vitro and in vivo and our single-molecule imaging approach promises to be a powerful tool that will contribute significantly to this field.

Chapter 4 Ultrafast Disease Biomarker Detection through Fluorogenic Single Molecule Recognition

4.1 Introduction

SMFKF is a powerful amplification-free analytical approach with remarkable specificity and outstanding sensitivity. However, for direct implementation of SMFKF-type biosensors in clinical diagnostics and prognostics, further improvements on the limit of detection must be made due to the scarcity of DNA biomarkers. For example, circulating cell-free tumor DNAs are generally below 10 copies per μl blood plasma[53], less than 16.6 aM in molar concentration. Current SiMREPS assays do not typically reach such sensitivity without the aid of pre-enrichment[74]. The physical limitation on sensitivity of SMFKF detection is essentially the product of insufficient detection and capture efficiency[52]. First, roughly only 1% of target molecules in solution will be immobilized on the surface due to diffusion-limited mass transport[52]. At low concentration below 10 pM, target molecules form a concentration gradient close to surface due to limited passive diffusion rate. Consequently, near-surface concentration is much lower than picomolar range and the apparent capture rate will decrease over time and eventually reach a plateau at which extra capturing time does not increase immobilized molecules significantly at all. Also, active mixing, e.g. pipetting or circulating, doesn't change near-surface concentration profile significantly due to limited turbulent flow. Second, among all the immobilized target molecules, less than 1% molecules will be detected due to the size limitation of the FOV. In Chapter 2, BSM-SiMREPS acquired total accepted counts from 10 FOVs as a single readout such that the signal response was "amplified" by 10-fold without

significant increase of false positives from non-target molecules, which were completely filtered out by kinetic fingerprinting. This demonstrates that an easy way of linearly improving sensitivity is to image as many FOVs as possible.

However, observing many FOVs also linearly increases the acquisition time, compromising the overall analytical performance by sacrificing time efficiency. To reach attomolar detection limits without sacrificing detection time, one solution is to shorten the acquisition required per FOV by accelerating the binding kinetics of the imager itself. Khanna *et al.* in 2021 reported an intramolecular SMFKF sensor for detecting cancer biomarkers of mutant DNA and microRNAs. They successfully demonstrated acquisition within 10 seconds. However, their sensor underwent rapid photobleaching since these intramolecular imagers were immobilized on the surface. Another simple way of doing this is to increase the imager concentration due to equation (1.3) and (1.4) to obtain a τ_{off} as small as possible. Unfortunately, an imager concentration higher than 100 nM compromises the overall assay performance due to unacceptable signal-to-noise ratio as shown in the following relationship[140]:

$$b = \beta \cdot c \cdot \xi \tag{4. 1}$$

Where b is the average background level per unit surface area, β is the molecular brightness of an imager in solution (number of detected photons per imager molecule), c is the concentration of imagers and ξ is the thickness of the observed volume. Although TIRF illumination effectively reduces ξ, in practice the maximum concentration of the imager in an SMFKF assay can be no more than 100 nM. Otherwise, true signals start to be overwhelmed by random fluctuation of the background noise.

The question arises whether there is any way for us to decrease β such that a higher c and thus shorter τ_{on} can be applied. Chung *et al.* in 2022 reported a super-fast DNA-PAINT

approach that demonstrated a 26-fold increase in imaging speed over regular DNA-PAINT by using a fluorogenic imager instead of a regular fluorescent probe[140]. A fluorogenic imager is simply a probe dual-labeled with a quencher and a fluorophore at distal ends. When freely diffusing in solution, fluorescence is mostly suppressed by the quencher moiety due to its proximity to the fluorophore moiety on a compacted, random coil probe, reducing the free imager background and allowing for higher concentrations. In contrast, whenever the fluorogenic imager is bound to the target site, the probe molecule is stretched and separates fluorophore and quencher, allowing strong fluorescence emission. Inspired by Chung *et al.*, Chapter 4 describes an ultrafast SMFKF sensor approach, "fluorogenic single molecule recognition by equilibrium through Poisson sampling" (FG-SiMREPS), which utilizes a fluorogenic DNA imaging probe to allow identification of single cancer mutant molecules in liquid biopsies with ultrahigh speed through kinetic fingerprinting and digital counting. We successfully achieved detection within just 2 seconds. This high-speed data acquisition rate are facilitated by combining two advances: 1) rational design of a probe sequence with mismatches and minimal self-structure; 2) utilization of an improved fluorophore-quencher pair with high fluorogenic ratio to utilize a high imager concentration of 5 μM. Consequently, we were able to image more than 100 FOVs within a few minutes, demonstrating detection for three cancer-related DNA targets, HPV, T790M, and L858R.

4.2 Materials and methods

Table 4.1. Lists of DNA strands, their code names, sequences and descriptions.

Code name	Sequences	Description
	T790M	
Exon20 T790M 25nt	5' CTCATCATGCAGCTCATGCCCTTCG 3′	Target DNA sequence of 25 nt T790M, directly purchased from IDT
Exon20 T790 WT 25nt	5' CTCATCACGCAGCTCATGCCCTTCG 3'	Target sequence of wildtype 25 nt T790, directly purchased from IDT
Cy5 F-gen V4	5' /Cy5/AGCTAAATAATGAG 3'	Cy5-labeled imager for T790M detection, 14 nt, containing 3 mismatches with Exon20 T790M 25nt and 4 mismatches with Exon20 T790 WT 25nt, directly purchased from IDT
BHQ2+Cy5 F-gen V4	5' /Cy5/AGCTAAATAATGAG/3BHQ2/ 3'	Dual-labeled imager for T790M detection by Black Hole Quencher 2 (BHQ2) and Cy5, 14 nt, containing 3 mismatches with Exon20 T790M 25nt and 4 mismatches with Exon20 T790 WT 25nt, directly purchased from IDT
BHQ2+Cy3B F-gen V4	5' /Cy3B/AGCTAAATAATGAG/3BHQ2 / 3'	Dual-labeled imager for T790M detection by Black Hole Quencher 2 (BHQ2) and Cy3B, 14 nt, containing 3 mismatches with Exon20 T790M 25nt and 4 mismatches with Exon20 T790 WT 25nt, directly purchased from Jena Bioscience
BHQ2+Cy3 F-gen V4	5' /Cy3/AGCTAAATAATGAG/3BHQ2/ 3'	Dual-labeled imager for T790M detection by Black Hole Quencher 2 (BHQ2) and Cy3, 14 nt, containing 3 mismatches with Exon20 T790M 25nt and 4 mismatches with Exon20 T790 WT 25nt, directly purchased from Jena Bioscience

T790M LNA CP	5' /Biotin TEG/C+GAA+GGGCA+TG 3'	Locked nucleic acid (LNA) capture probe for detecting T790M, containing 3 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
T790M LNA CP Blocker	5' ATGCCCTTCG 3'	Capture probe blocker (CP Blocker) complementary to T790M LNA CP to prevent non-specific interaction of imager sequence with unoccupied capture probes, directly purchased from IDT
T790M F-gen V4_Comple	5' CTCATTATTTAGCT 3'	A DNA sequence completely complementary to imagers for detecting T790M, directly purchased from IDT
Exon20 T790M Cy5 25nt	5' /5Cy5/CTCATCATGCAGCTCATGCC CTTCG 3'	Cy5-labeled Exon20 T790M 25nt, directly labeled from IDT
T790M LNA CP_v2_0	5' /Biotin TEG/CG+A+A+G+G+GC+A+T+G 3'	The second version of locked nucleic acid (LNA) capture probe for detecting T790M, containing 8 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
	L858R	
EGFR L858R MUT 26	5' GTCAAGATCACAGATTTTGGGCGGGC 3'	Target DNA sequence of 26 nt EGFR L858R, directly purchased from IDT
EGFR L858 WT 26	5' GTCAAGATCACAGATTTTGGGCTGGC 3'	Target sequence of wildtype 26 nt EGFR L858, directly purchased from IDT
L858R FG Imager_v1_0_Cy3B	5' /5Cy3B/GCTCGCTCTATATCT/3BH Q2/ 3'	Dual-labeled imager for L858R detection by Black Hole Quencher 2 (BHQ2) and Cy3B, 15 nt, containing 4 mismatches with EGFR L858R MUT 26 and 5 mismatches with EGFR L858 WT 26, directly purchased from Jena Bioscience

L858R FG Imager_v1_0_Cy3	5' /5Cy3/GCTCGCTCTATATCT/3BHQ 2/ 3'	Dual-labeled imager for L858R detection by Black Hole Quencher 2 (BHQ2) and Cy3, 15 nt, containing 4 mismatches with EGFR L858R MUT 26 and 5 mismatches with EGFR L858 WT 26, directly purchased from Jena Bioscience
L858R FG Imager_v1_0_Comple	5' AGATATAGAGCGAGC 3'	A DNA sequence completely complementary to imagers for detecting L858R, directly purchased from IDT
L858R LNA CP_v1_0	5' G+TGAT+CT+T+GAC 3'	Locked nucleic acid (LNA) capture probe for detecting L858R, containing 4 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
L858R LNA CP_v2_0	5' +G+T+G+AT+C+T+T+G+A+C/3Bio TEG 3'	The second version of locked nucleic acid (LNA) capture probe for detecting T790M, containing 10 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
L858R LNA CPBlocker_v2_0	5' GTCAAGATCA 3'	Capture probe blocker (CP Blocker) complementary to L858R LNA CP to prevent non-specific interaction of imager sequence with unoccupied capture probes, directly purchased from IDT
	HPV	
HPV16 26nt_v0	5' TAGTATAAAAGCAGACATTTTATGCA 3'	Target DNA sequence of 26 nt HPV, directly purchased from IDT
HPV16 FG Imager_v1_0_Cy3B	5' /5Cy3B/TCTGCTCCTACCCTA/3BH Q2/ 3'	Dual-labeled imager for HPV detection by Black Hole Quencher 2 (BHQ2) and Cy3B, 15 nt, containing 4 mismatches with HPV16 26nt_v0, directly purchased from Jena Bioscience

HPV16 FG Imager_v1_0_Cy3	5' /5Cy3/TCTGCTCCTACCCTA/3BHQ 2/ 3'	Dual-labeled imager for HPV detection by Black Hole Quencher 2 (BHQ2) and Cy3, 15 nt, containing 4 mismatches with HPV16 26nt_v0, directly purchased from Jena Bioscience
HPV16 FG Imager_v1_0_Comple	5' TAGGGTAGGAGCAGA 3'	A DNA sequence completely complementary to imagers for detecting HPV, directly purchased from IDT
HPV16 LNA CP_v1_0	5' /Biotin TEG/T+GCAT+A+AAA+TG 3'	Locked nucleic acid (LNA) capture probe for detecting HPV, containing 4 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
HPV16 LNA CP_v2_0	5'/Biotin TEG/+T+G+C+A+T+A+A+A+A+T+G 3'	The second version of locked nucleic acid (LNA) capture probe for detecting HPV, containing 11 locked nucleotides, "+N" locked nucleotide, directly purchased from IDT
HPV16 LNA CPBlocker_v2_0	5' ATTTTATGCA 3'	Capture probe blocker (CP Blocker) complementary to HPV16 LNA CP to prevent non-specific interaction of imager sequence with unoccupied

4.2.1 Assay pipeline and working principle of FG-SiMREPS

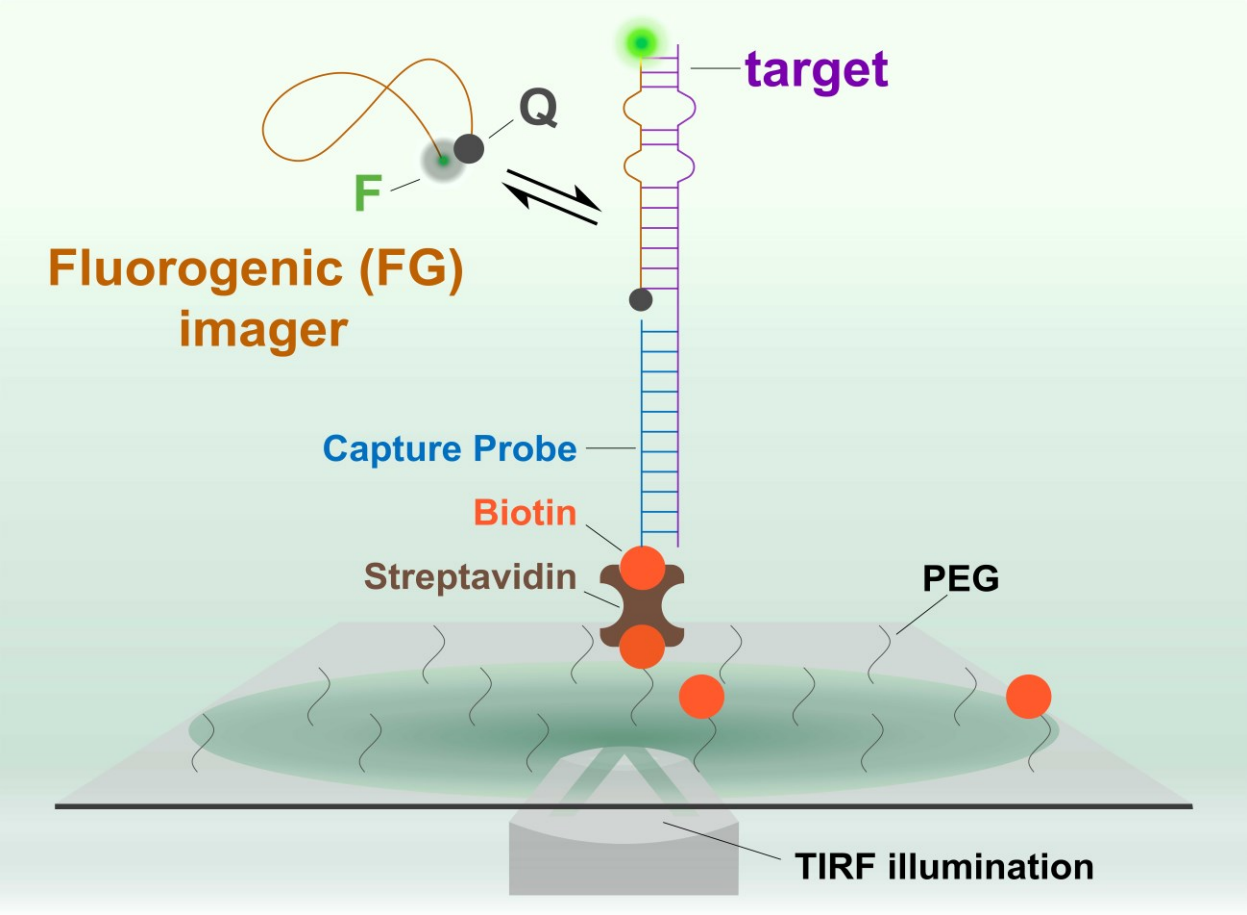

Figure 4.1. Schematic of FG-SiMREPS. A fluorogenic DNA probe labeled with a fluorophore (F) and a quencher (Q) is designed to recognize a target DNA immobilized on the surface. PEG: polyethyleneglycol. TIRF: total internal reflection fluorescence.

Figure 4.1 shows the assay pipeline of FG-SiMREPS. This simple assay starts with a target solution added to sample wells, where a biotinylated capture probe (shown in **Table 4.1**) is precoated on a streptavidin-coated PEG surface through biotin-streptavidin interaction. Capture probes immobilized both MUT and WT molecules at 37°C with or without formamide for 1 h. This relatively harsh condition was necessary for surface capture to prevent formation of secondary structure of LNA capture probes. Following surface capture, a 1 μM capture probe blocker was added and incubated at 37°C for 20 min to saturate any vacant capture probes. After that, an imaging buffer containing fluorogenic imager as well as oxygen scavenger system was

added, and the sample well was imaged under TIRF (total internal reflection fluorescence) illumination using an oil objective. Transient interaction of FG imager with immobilized MUT molecules generated kinetic fingerprinting upon laser excitation. Fluorescence emission signal is collected for downstream data analysis. Repeated binding and dissociation of FG imager with will be visualized by fluorescence-time traces, which are used for distinguishing WT and Blk signals.

4.2.2 Oligonucleotides

All DNA oligonucleotides except BHQ2+Cy3B F-gen V4, BHQ2+Cy3 F-gen V4, L858R FG Imager_v1_0_Cy3B, L858R FG Imager_v1_0_Cy3, HPV16 FG Imager_v1_0_Cy3B and HPV16 FG Imager_v1_0_Cy3 were purchased from Integrated DNA Technologies (IDT, www.idtdna.com) with standard desalting purification, unless otherwise noted. BHQ2+Cy3B F-gen V4, BHQ2+Cy3 F-gen V4, L858R FG Imager_v1_0_Cy3B, L858R FG Imager_v1_0_Cy3, HPV16 FG Imager_v1_0_Cy3B and HPV16 FG Imager_v1_0_Cy3 were purchased from Jena Bioscicence (https://www.jenabioscience.com/) with HPLC purification and quality-checked by LC-MS. All fluorophore labeled DNA oligonucleotides were HPLC-purified when purchased.

4.2.3 Rational design of sensor constructs

An in-house matlab code was written by our previous lab member Aaron Blanchard to implement the above selection process. In the case of detecting HPV, we first carefully chose the imager-binding region with the lowest number of AT pairs and GC pairs as well as with the lowest probability of forming secondary structure predicted by Nupack ((https://www.nupack.org/). For detecting T790M or L858R, target sequence was fixed and imager-binding region was chosen at the end with the with the lower number of AT pairs and GC

pairs as well as with the lowest probability of forming secondary structure predicted by Nupack. First of all, length, number of internal mismatches and number of base pairs on each end are given for numerating all possible sequences. Secondly, a score was calculated for each sequence based on 1) how close the binding energy of a sequence is to the regular T790M imager[72]; 2) how big a difference is in the binding energy between a imager sequence to the MUT and to the WT, maximizing detection specificity. As a result, clusters of sequences ranked by their scores were spit out. Finally, within the top-ranked cluster, we estimated probability of forming secondary structure for the sequences with lowest number of AT pairs and GC pairs using Nupack. The candidate sequences with lowest probability of forming secondary structure at each base were then fed into a Nupack prediction together with the imager-binding region of a target to make sure that a desired duplex formed with mismatches.

4.2.4 Fluorogenicity measurement

Fluorogenic ratio measurement was conducted on a SpectraMax ID3 plate reader (Molecular Devices, https://www.moleculardevices.com/) with a clear bottom-flat 96 plate. Before fluorescence measurement, 1 μM fluorophore-labeled imagers with or without quencher were mixed with their 10 μM fully complementary unmodified counterparts in 4X PBS. Using medium PMT for excitation and 100 ms for signal integration, samples were excited at 530 nm and fluorescence emission was measured at 570 nm for Cy3B-labeled or Cy3-labeled imagers and for Cy5-labeled imagers, samples were excited at 630 nm and fluorescence emission was measured at 670 nm. Readings of a blank 4X PBS buffer under the same imaging condition were subtracted from sample signals for fluorogenicity calculation.

4.2.5 3D-printed sample wells

Design of sample wells for 3D printing was drew with Autodesk Fusion 360 (https://www.autodesk.com/products/fusion-360/). Sample wells came with a wide bottom base in the middle of which was the capture region. Detailed dimensions are illustrated in **Figure 4.4**A. Sample wells were printed on Stratasys J750 using a clear material at the Duderstadt Center's Fabrication Studio at the University of Michigan. Before use, 3D-printed sample wells were cleaned of support resin by sonicating in milliQ water for 20 min twice, rinsing with 200 proof ethanol twice and milliQ water twice and finally drying out under vacuum overnight.

4.2.6 FG-SiMREPS assay protocol

Sample cells either made of cut P20 pipette barrier tips or 3D printing were attached to glass coverslips passivated with a 1:10 mixture of biotin-PEG and mPEG. A detailed protocol of slide preparations is discussed in previously published papers[81]. Sample cells were first washed with T50 buffer (10 mM Tris-HCl [pH 8.0 at 25°C], 50 mM NaCl) and then incubated with 40 µl 1 mg/ml streptavidin in T50 buffer for 10 min. Following a wash with 1X PBS for 3 times, 100 nM capture probe in 1X PBS that was preheated at 90°C for 5 min in a metal bath, annealed at 37°C for 5 min in a water bath, and cooled down to room temperature, was then added to the sample well. The sample well was incubated for 10 min and washed with 4X PBS for 3 times waiting for the target strand. A target-containing solution was prepared in a PCR tube that contained 1 pM or 10 pM targets in 4X PBS / 2 µM poly-T oligodeoxyribonucleotide (dT10 or dT30) carriers. All dilutions of targets were performed in the presence of 2 µM dT10 or dT30 in GeneMate low-adhesion 1.7 mL microcentrifuge tubes (VWR, Cat No. 490003-230). PCR tubes were then heated at 80°C for 2 min, cooled down at 30°C for another 30 s and finally held at 22°C in a thermocycler. Subsequently, the target-containing solution was added to the sample cell and then incubated for 1 h at 37°C. For L858R capturing, samples incubated in 15%

formamide, 4X PBS, 2 µM dT10 or dT30 at 37°C. After target capture, sample cells were washed 3 times with 4X PBS and incubated with 1 µM capture probe blocker in 4X PBS. 50 µl imaging buffer containing the desired concentration of FG imagers in the presence of an oxygen scavenger system (OSS) — 1 mM Trolox, 5 mM 3,4-dihydroxybenzoate (PCA), 50 nM protocatechuate dioxygenase (PCD) — was added and then imaged by objective-TIRF microscopy. 1 µM PCD stock was prepared in 100 mM Tris-HCl pH 8.0, 50 mM KCl, 1 mM EDTA, 50% glycerol; 100 mM PCA was dissolved in water and titrated with 5 M KOH to a pH of 8.3; Trolox was dissolved in water and titrated with 5 M KOH to a pH around 10-11. All three components are stored in -20°C prior to use.—For specific buffer conditions, please refer to **Figure 4.3**.

4.2.7 Single-molecule fluorescence microscopy

All single-molecule experiments were performed on the Oxford Nanoimager (ONI), a compact benchtop microscope capable of objective-type TIRF (See https://oni.bio/nanoimager/ for spec sheet regarding camera, illumination and objective). A 100X 1.4NA oil-immersion objective was installed on ONI together with a built-in Z-lock control module for autofocus. Since the built-in temperature control system on ONI could not keep imaging temperature below 25°C, to avoid overheating by turning laser on for too long, we attached the outer box of ONI to a metal clamp where circulating cold water coming from a water bath could run through. Before imaging, imaging temperature of ONI was pre-equilibrated. And after turning on laser, the temperature of water bath was adjusted accordingly in real-time for maintaining a constant imaging temperature. An illumination angle of around 54° was used for TIRF. Detailed acquisition settings please refer to **Figure 4.3**.

4.2.8 Processing and analysis of objective-TIRF type data

A set of custom MATLAB codes were used to identify spots with significant intensity fluctuation within each FOV, generate intensity-versus-time traces at each spot, fit these traces with two-state hidden Markov modeling (HMM) algorithm to generate idealized traces, and eventually identify and characterize transitions with idealized traces. A set of filtering criteria were generated to distinguish methylation-specific signal and non-specific signal by feeding traces from no target control experiments as negative dataset and traces from methylated target experiments as positive dataset into a SiMREPS optimizer (see). A detailed discussion of data analysis pipeline can be found in papers previously published in our group[81].

Table 4.2. Optimized parameter sets for trace generation and analysis in detection of T790M with 4-second and 2-second acquisition. See **Table 1.1** for detailed description of each parameter.

Trace Generation Parameters		
Parameters	**4 s**	**2 s**
use fluctuation map?	1	1
Stdfactor	3.5	4
start frame	1	1
end frame	200	200
edgePx	20	20
Percentilecut	0.95	0.95
ROI size (pixels)	5	5

Trace Analysis Parameters (KFC)		
Parameters	**4 s**	**2 s**
start frame	1	1
end frame	200	100
exposure time (s)	0.02	0.02
Smoothframes	1	1
remove_single_frame_events	FALSE	FALSE
Ithresh	2000	300
SNthresh	2	2
SNthresh_trace	1.7	3
min_Nbd	11	1

max_Nbd	40	Inf
min_tau_on_median (s)	0.06	0.06
min_tau_off_median (s)	0.08	0.02
max_tau_on_median (s)	0.2	Inf
max_tau_off_median (s)	Inf	1.22
max_tau_on_cv	Inf	Inf
max_tau_off_cv	Inf	Inf
max_tau_on_event (s)	Inf	Inf
max_tau_off_event (s)	Inf	Inf
max_I_low_state	19400	2400
vary_I_vals	FALSE	FALSE
num_intensity_states	2	2
ignore_post_bleaching	FALSE	FALSE
bleaching_wait_time (s)	Inf	Inf
use_FRET_threshold	FALSE	FALSE
FRET_threshold	0	0

Table 4.3. Optimized parameter sets for trace generation and analysis in detection of L858R with 4-second and 2-second acquisition. See **Table 1.1** for detailed description of each parameter.

Trace Generation Parameters		
Parameters	**4 s**	**2 s**
use fluctuation map?	1	1
Stdfactor	3.5	3
start frame	1	1
end frame	200	200
edgePx	20	20
Percentilecut	0.95	0.95
ROI size (pixels)	5	5

Trace Analysis Parameters (KFC)		
Parameters	**4 s**	**2 s**
start frame	1	1
end frame	200	100
exposure time (s)	0.02	0.02
Smoothframes	1	1
remove_single_frame_events	FALSE	FALSE
Ithresh	2000	1000
SNthresh	2	2

SNthresh_trace	2	3
min_Nbd	4	2
max_Nbd	Inf	8
min_tau_on_median (s)	0.02	0.02
min_tau_off_median (s)	0.06	0.02
max_tau_on_median (s)	1.24	Inf
max_tau_off_median (s)	1.6	Inf
max_tau_on_cv	Inf	Inf
max_tau_off_cv	Inf	Inf
max_tau_on_event (s)	Inf	Inf
max_tau_off_event (s)	Inf	Inf
max_I_low_state	400	Inf
vary_I_vals	FALSE	FALSE
num_intensity_states	2	2
ignore_post_bleaching	FALSE	FALSE
bleaching_wait_time (s)	Inf	Inf
use_FRET_threshold	FALSE	FALSE
FRET_threshold	0	0

Table 4.4. Optimized parameter sets for trace generation and analysis in detection of HPV with 4-second and 2-second acquisition. See **Table 1.1** for detailed description of each parameter.

Trace Generation Parameters		
Parameters	**4 s**	**2 s**
use fluctuation map?	1	1
Stdfactor	4	4
start frame	1	1
end frame	200	200
edgePx	20	20
Percentilecut	0.95	0.95
ROI size (pixels)	3	3

Trace Analysis Parameters (KFC)		
Parameters	**4 s**	**2 s**
start frame	1	1
end frame	200	100
exposure time (s)	0.02	0.02

Smoothframes	1	1
remove_single_frame_events	FALSE	FALSE
Ithresh	1000	1000
SNthresh	2	2
SNthresh_trace	1	0
min_Nbd	1	2
max_Nbd	Inf	Inf
min_tau_on_median (s)	0.06	0.02
min_tau_off_median (s)	0.1	0.02
max_tau_on_median (s)	0.18	2.14
max_tau_off_median (s)	0.62	2.62
max_tau_on_cv	Inf	Inf
max_tau_off_cv	Inf	Inf
max_tau_on_event (s)	Inf	Inf
max_tau_off_event (s)	1.88	51.82
max_I_low_state	1000	Inf
vary_I_vals	FALSE	FALSE
num_intensity_states	2	2
ignore_post_bleaching	FALSE	FALSE
bleaching_wait_time (s)	Inf	Inf
use_FRET_threshold	FALSE	FALSE
FRET_threshold	0	0

4.3 Results

4.3.1 Sensor constructs for detecting three cancer DNA biomarkers

The key consideration for designing a FG-SiMREPS imager is ensuing a binding kinetics similar to a regular SiMREPS probe. All downstream optimizations on imaging conditions rely on the fact that the dissociation rate constant k_{off} is high enough for observing sufficient N_{b+d} and the apparent association rate constant k'_{on} can be easily modulated by imager concentrations. In fact, as long as the binding kinetics of a fluorogenic imager falls within the SiMREPS regime, e.g., a k_{on} between approximately 0.06 $\mu M^{-1}s^{-1}$ and 20 $\mu M^{-1}s^{-1}$ and a k_{off} between approximately 0.01 s^{-1} and 1 s^{-1}, it is then possible to adjust parameters like temperature,

chaotropic reagents and salt concentration, etc. In other words, the K_d of a SMFKF imager should be approximately between $100\ nM \sim 10\ \mu M$ at a standard condition (25°C, 1X PBS pH 7.4).

The very first feature is the length of FG-SiMREPS imager. To ensure a complete separation between fluorophore and quencher in bound state, a 14-15 nt imager (approximately 50 Å, a distance close to most Forster Distances) is preferred. The actual distance in the bound state between the fluorophore and quencher will be longer than just the length of the dsDNA backbone due to the existence of linkers in fluorophore moiety and quencher moiety. Any length shorter than that may introduce partial quenching even when bound. In contrast, any length higher than that may not result in optimal quenching when unbound simply due to the gyration radius of the random coil of a ssDNA then becomes too big. Finally, a fully complementary 14-15 nt imager also irreversibly binds its partner sequence. Therefore, we introduce several internal mismatches to make its binding affinity close to that of a regular SiMREPS (**Figure 4.2**). Usually 3-4 mismatches will be sufficient.

Although the binding affinity can be easily calculated given a specific imager sequence containing several mismatches, it is difficult to predict individual k_{on} and k_{off}. Therefore, in our design, we minimize any potential risks of competing interactions against imager-target hybridization. One major competitor is the secondary structure formation of the imager itself and target sequence itself. Any formation of secondary structures by imager or target prevents the nucleation step during their hybridization and thus decreases k_{on} dramatically. Therefore, after screening and picking out candidate sequences with similar binding affinity as previously used regular SiMREPS imager[72], we estimate the probability of forming any transient secondary structure and ensure a relatively large k_{on}. Coming together, the ultimate k_{off} will naturally fall

into the range of an optimal SiMREPS imager. Once imager sequences were generated, the capture probe sequences were designed and locked nucleotides were introduced to ensure a stable capture, tolerating chaotropic reagents like formamide and high temperature (**Figure 4.2**). Of note, we observed significant dissociation of surface-captured target with capture probes having no LNA modification when using formamide to shorten τ_{on} by destabilizing imager-target duplex (data not shown here).

Ultimately, the final sensor constructs for detecting three targets, T790M, L858R and HPV were shown in **Figure 4.2**A. After determining the putatively optimal sequences for fluorogenic imagers, we measured their fluorogenicity by hybridization with a fully complementary unmodified DNA sequence (**Table 4.1**, **Figure 4.2**B). Previous lab members, Zi Li and Aaron Blanchard also tested a different quencher, Iowa Black® RQ. They discovered a strong non-specific interaction of Iowa Black® RQ-labeled imager with streptavidin-coated surface itself (data not shown here) and thus we chose black hole quencher 2 (BHQ2). Different fluorophores of Cy5, Cy3 and Cy3B were also tested for T790M imager and Cy3B gave the best fluorogenic ratio in all three imager sequences since its emission spectrum overlapped the most with the absorbance spectrum of BHQ2. This was also consistent with Chung *et al.*'s result. However, Cy3 was less quenched by BHQ2 than Cy5 even with more overlap with BHQ2. Careful design of imager sequence and capture probe sequence is essential for downstream optimizations on imaging conditions. In the end, BHQ2-Cy3B dual-labeled imager gave 32-fold fluorescence increase in bound state for T790M and over 60-fold fluorescence increase for L858R and HPV. This difference in fluorogenic ratio may be due to a longer imager in the case of HPV and L858R, 15 nt versus 14 nt in the case of T790M, causing better physical separation between Cy3B and BHQ2 when forming duplex.

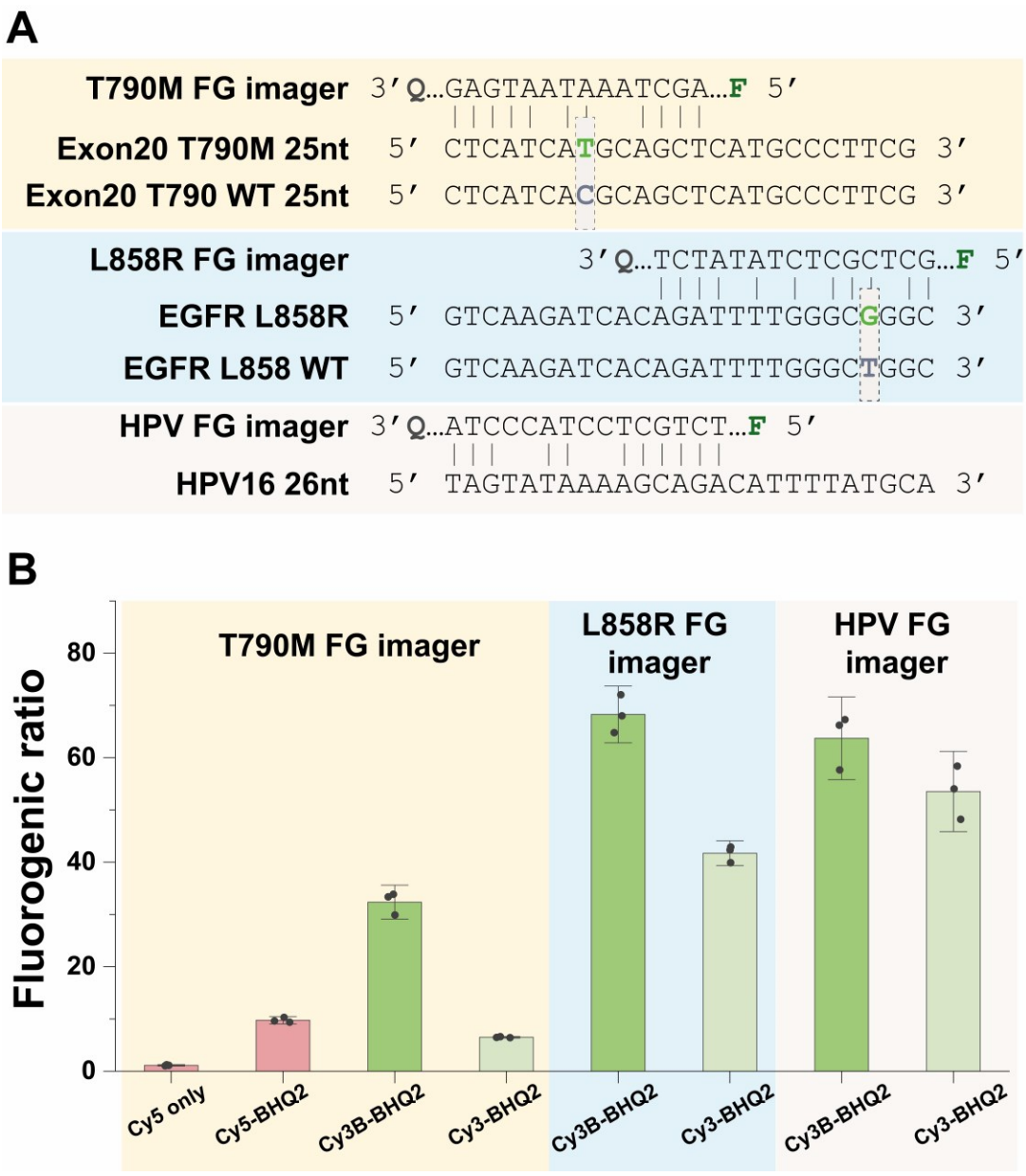

Figure 4.2. Sensor constructs designed for detecting three targets: T790M, L858R and HPV. (A) Sequences of fluorogenic probes designed for detection of three cancer biomarkers - Exon20 T790M mutation, EGFR L858R mutation and HPV viral sequence. We incorporate 3-4 mismatches in our sequence design to allow fast binding and dissociation. (B) Fluorogenicity measurement of imagers labeled with different fluorophores. $Fluorogenic\ ratio = \frac{Intensity_{bound}}{Intensity_{unbound}}$. Datapoints are presented as mean ± s.d., where n = 3 independent experiments.

4.3.2 Two-second detection for three cancer DNA biomarkers

After proving a good fluorogenicity of imagers for all three target on the ensemble-level, we tested detection of all three targets using FG-SiMREPS. Numerous optimizations on imaging conditions were conducted for each target. Briefly speaking, we started with an imaging condition similar to the regular SiMREPS: 50 nM imager and room temperature with acquisition time of 5 min under three conditions: MUT, WT and Blk. After identifying MUT-specific or target-specific kinetic fingerprinting, we estimated the τ_{on} and τ_{off} under this regular imaging condition. Dwell time measured at 50 nM imager and room temperature inferred us the next step of optimizations. For example, if τ_{on} was larger than 10 s, then chaotropic reagents like formamide and higher temperature could be applied to decrease τ_{on} dramatically. To achieve 2 s acquisition for each FOV, a general optimization scheme was starting with shortening τ_{on} by changing one parameter (increasing formamide percentage or temperature) at a time until we achieved around 0.1 s for median τ_{on}, then increasing concentration of imager until median τ_{off} was also smaller than 1 s since imager concentration does not change τ_{on}. However, τ_{off} did not inversely decrease with concentration higher than approximately 1 μM. Instead, τ_{off} reached a lower plateau independent of concentration at micromolar range. This phenomenon was observed for detection of all three targets. One hypothesis is that at micromolar concentrations of imager, the local concentration of imager on the detection surface was saturated such that any further increase in the ensemble-level concentration no longer had an effect. In other words, the PEG-Streptavidin-Capture probe surface had an upper limit on the local imager concentration. PEG is a dominant surface "solvent" on the detection surface and since PEG is a great passivating reagent for nucleic acids, it doesn't "dissolve" imager very well, thus restricting how many imager molecules can approach a PEG surface.

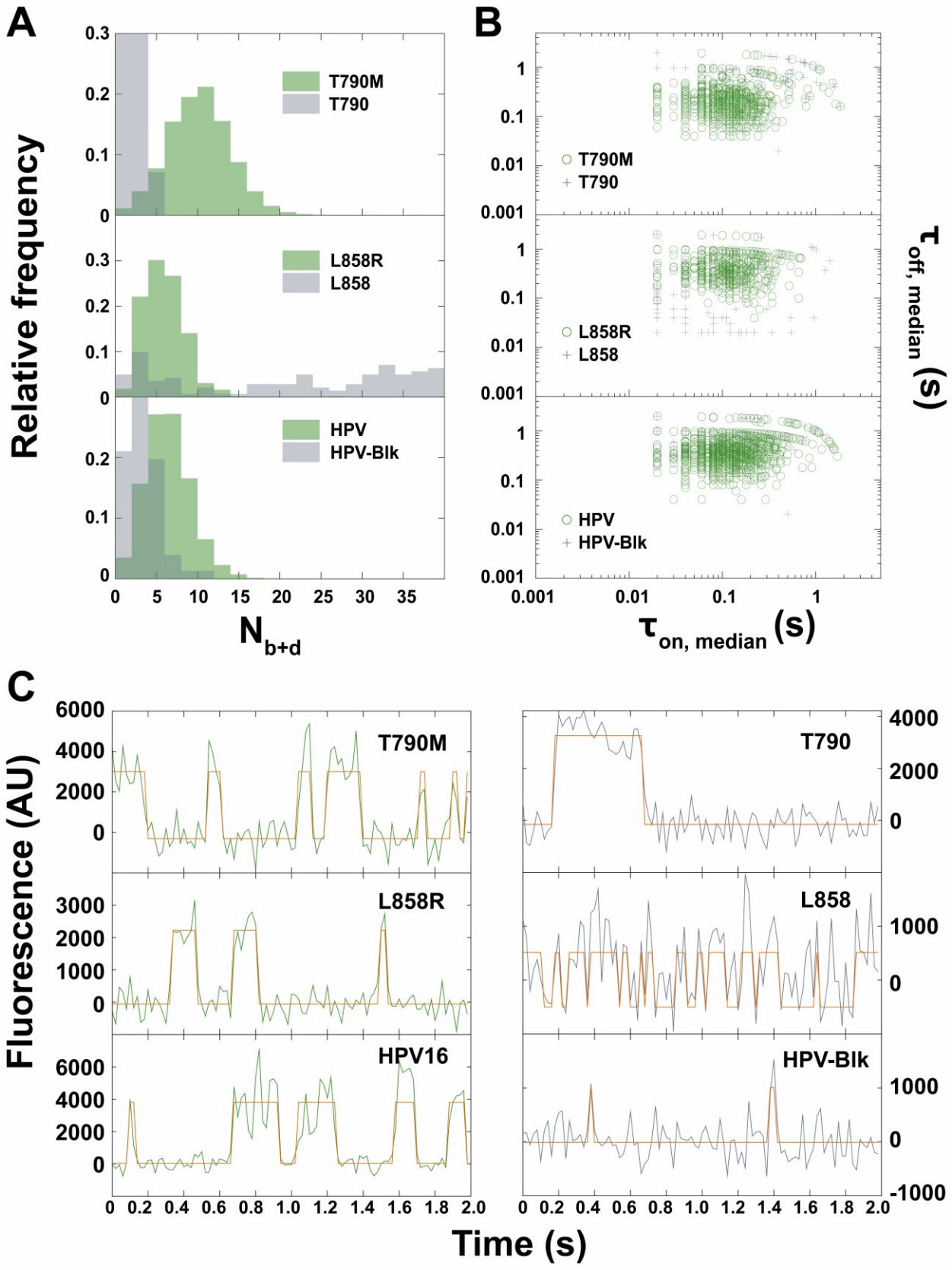
A
Relative frequency
T790M
T790
L858R
L858
HPV
HPV-Blk
N b+d
B
T790M
T790
L858R
L858
HPV
HPV-Blk
τ on, median (s)
τ off, median (s)
C
Fluorescence (AU)
T790M
T790
L858R
L858
HPV16
HPV-Blk
Time (s)

Figure 4.3. FG-SiMREPS detection of three targets with just two-second acquisition. Imaging conditions: for T790M detection, 5 μM imager 10% formamide 28°C; for L858R detection, 1 μM imager 15% formamide 34°C; for HPV detection, 5 μM imager 0% formamide 25°C; acquisition setting: 20 ms exposure time, 100 frames, around 155 mW 532 nm excitation. (A) Left column: distributions of N_{b+d}, number of binding and dissociations in each trace; right column: distributions of median bound time, $\tau_{on,median}$ and median unbound time, $\tau_{off,median}$ in each trace. Each spot represents a single trace. (B) Comparison of kinetic fingerprinting between mutant (left column) and wildtype/blank (right column). Green and gray lines are raw intensity traces and orange lines are idealized traces by HMM fitting. AU stands for arbitrary units.

Figure 4.3 shows the results of FG-SiMREPS for detection three targets with just two-second acquisition. **Figure 4.3**C compares representative traces between MUT and WT/Blk and based on their distinct kinetics features, we were able to separate MUT completely from WT and Blk with just two-second acquisition by combining thresholds of N_{b+d}, τ_{on} and τ_{off}. The optimal imaging conditions were unique to each individual target although their binding affinity was all similar to a regular SiMREPS imager at a standard condition (25°C, 1X PBS pH 7.4). This suggested that hybridization kinetics were strongly influenced by sequence variations alone.

4.3.3 Multiple-FOV detection

Following demonstration of two-second detection of a single FOV, we next tested multiple-FOV acquisition to achieve maximum detection efficiency by scanning the entire bottom surface within a sample well. **Figure 4.4**A shows the illustration of the custom design of 3D-printed sample wells. The size of the bottom hole is slightly larger than the total size of 104 FOVs. **Figure 4.4**B illustrates the serpent-line movement pattern of the objective starting from the bottom right corner as the first FOV.

Figure 4.5 shows the distribution of accepted counts across 104 FOVs for three different targets using the optimized imaging conditions shown in **Figure 4.3**. In **Figure 4.5**, the acquisition time for each FOV was 4 seconds to sufficiently separate MUT signals from WT and Blk signals by allowing higher number of repeated binding and dissociation. Out of three targets,

only the distribution of accepted counts in detecting T790M exhibited a distinct Poisson peak, well separatable from WT and Blk. Both L858R and HPV exhibited a uniform-like distribution, suggesting strong capture heterogeneity on detection surface. In fact, the spatial distribution of accepted counts was never isotropic after switching to the 3D-printed sample wells. The uneven capture might arise from leakage of epoxy when gluing sample wells on the surface or damaged detection surface by scratch of pipette tips since the bottom surface left no space for pipetting when touching the bottom.

We also observed that in detection of all three targets, significant false positives exist in both WT and Blk even with four-second acquisition for each FOV. Combined population of WT or Blk in N_{b+d}, τ_{on} and τ_{off} across all 104 FOVs overlapped significantly with the combined population of MUT. The SiMREPS optimizer could not generate an optimal set of filtering criteria completely separating off-target signals. In other words, combining readouts from 104 FOVs not only linearly amplified true signals but also linearly amplified false positive signals. As a result, no analytical improvement in sensitivity was obtained.

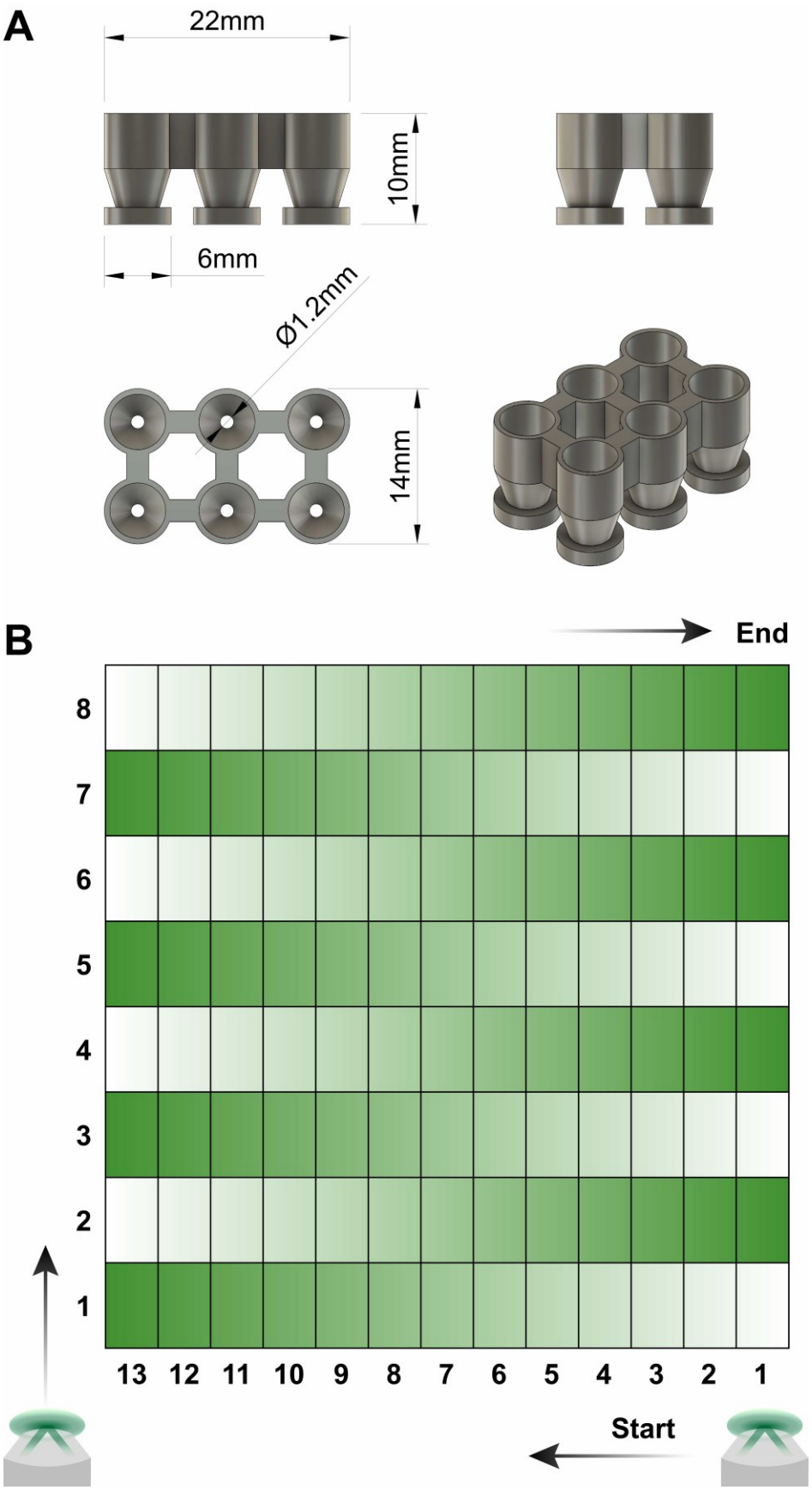
A
22mm
10mm
6mm
Ø1.2mm
14mm
B
End
Start
8
7
6
5
4
3
2
1
13 12 11 10 9 8 7 6 5 4 3 2 1

Figure 4.4. Design of 3D-printed sample wells and acquisition scheme for 104-FOV detection. (A) Dimensions and drawings of 3D-printed sample wells. (B) Serpent line acquisition of 104 FOVs. Objective moves from the bottom right to the left on the first row and then moves up to the right, following the darkening color gradient shown in this panel. Eace cell represents a single FOV.

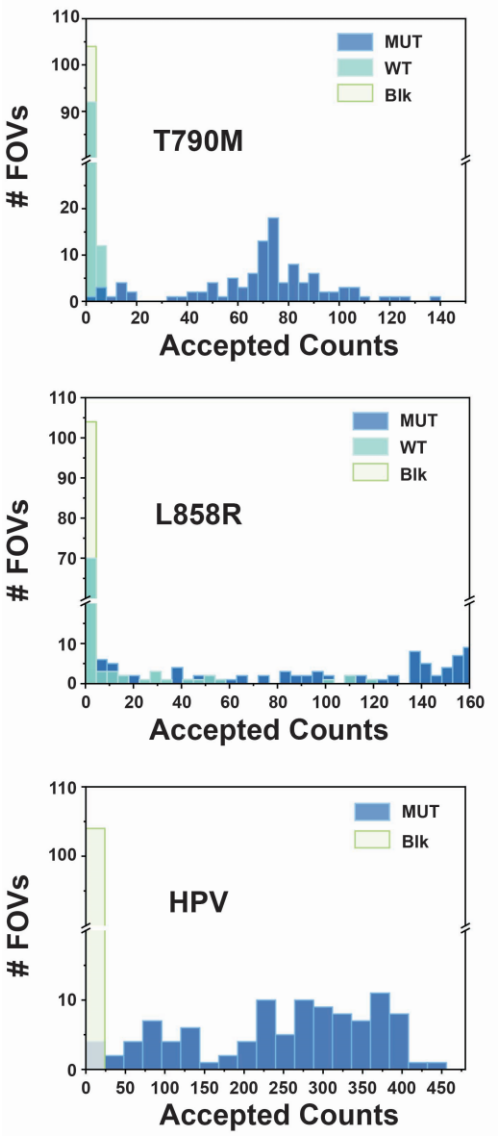

Figure 4.5. Distribution of accepted counts across 104 FOVs for detection of three targets: 1 pM T790M versus 10 pM T790, 10 pM L858R versus 10 pM L858 and 10 pM HPV. Acquisition for each FOV is 4 s. Kinetic Filtering criteria were trained using MUT as positive training datasets and WT/Blk as negative training datasets. Imaging conditions follow as in **Figure 4.3**.

4.4 Discussion

In Chapter 4, we developed an ultrafast, amplification-free single molecule kinetic fingerprinting detection approach using a fluorogenic imager. We successfully demonstrated detection of three different cancer DNA biomarkers: T790M, L858R and HPV within just 2 seconds, after thorough optimizations of imaging conditions following rational design of sensor constructs. Key design considerations were discussed and the optimal fluorogenicity of all three imagers were achieved by using Cy3B-BHQ2. In the end, an over 60-fold increase in fluorescence was achieved at the ensemble level and the acquisition speed was increased by 150-fold (from 5 min to 2 s) compared to regular SiMREPS approaches and a 5-fold increase in acquisition speed was achieved compared to intramolecular SiMREPS sensor without issues of photobleaching[104,141]. The FG-SiMREPS allowed us to use an imager concentration as much as 5 μM, opening a gate to high-throughput detection platform using SMFKF.

However, our attempts to improve sensitivity using FG-SiMREPS by combining readouts across 104 FOVs failed due to significant false positives in WT and Blk that were not separable from MUT populations. Another issue encountered was uneven spatial distribution on the detection surface using 3D-printed sample wells where a small hole restricted the total area for immobilizing targets. The motivation of using the 3D-printed sample wells was to maximize the detection efficiency by scanning the entire capture surface. However, this would not improve sensitivity since there was no increase in surface density of captured target. In the end, the total counts remained the same if the same number of FOVs was detected. Also, leakage of epoxy during attachment of sample wells to PEGylated coverslips was difficult to avoid and caused unpredictable variations of the effective detection surface, compromising reproducibility and

assay feasibility. To conclude, there was really no benefit for using the 3D-printed sample wells other than causing more inconvenience.

Therefore, the question remains how to avoid false positives arising from WT and Blk samples. These false positives were either artifact of fluorescent variations due to unstable stage position or off-target interactions between imager and capture probe, streptavidin or non-passivated spot on the surface. The interaction between imager and capture probe was mostly blocked by high concentrations of dT10 or dT30 as carriers and preincubation of capture blockers (**Table 4.1**). Thus, the main source of off-target came from imagers interacting with streptavidin or non-passivated spot. Using a coverslip with lower biotin ratio, e.g., 1:100 PEG/BioPEG instead of 1:10 PEG/BioPEG could significantly prevent the interaction with streptavidin by decreasing its surface density. In the end, it became a question of how to prevent imagers from sticking to non-passivated spots. Assuming that it was the moieties of fluorophore and quencher that stuck to these spots, then adding an imager with the sequence but labeled with a different fluorophore, e.g., Cy5 should produce a different spatial profile of off-target binding. Colocalization between Cy3B emission channel and Cy5 emission channel should completely remove those interaction with non-passivated spots. Also, three additional sets of filtering criteria could be applied to each individual channel as well as to combined two channels of colocalized spots. Ultimately, 4 sets of filtering thresholds can be used to eliminate false positives in WT and Blk, ensuring linear amplification of only true signals when combining readouts from as many FOVs as possible.

Chapter 5 Summary and Outlook

5.1 BSM-SiMREPS revealing underestimation of DNA methylation by PCR-based approaches

In Chapter 2, we developed a quantitative amplification-free SMKFK assay, termed BSM-SiMREPS, for detecting clusters of DNA methylation in a cancer DNA biomarker, the BCAT1 promoter sequence. Analytical performance was extensively characterized by standard curves of synthetic DNAs, bisulfite-convert DNAs and spike-in DNA in a background of unmethylated DNAs as well as genomic DNA. BSM-SiMREPS exhibited extraordinary specificity of 99.9999% in most cases and a limit of detection at the sub-femtomolar level, one of the highest sensitivities among all amplification-free approaches described to date. We further demonstrated BSM-SiMREPS measurement of BCAT1 promoter methylation in extracted genomic DNA from whole blood and discovered over 30% methylation, significantly higher than detected by two mainstream PCR-based approaches (whole-genome bisulfite sequencing and methylation EPIC array). In fact, both existing whole-genome bisulfite sequencing and methylation EPIC array gave around 2-5% methylation beta values on average, suggesting close to 0% methylation in the input samples[109]. We further validated the quality of samples using bisulfite-pyrosequencing (data not shown here), which also gave around 2% methylation at BCAT1 promoter using the same tube of whole-blood DNA. Since the major technical difference in sample preparation between BSM-SiMREPS and gold-standard approaches was PCR amplification, we posit that there may be a risk of underestimation by PCR-based approaches for

DNA methylation detection due to "PCR bias". In fact, PCR bias is a known measurement error in regular sequencing assays but has been ignored in bisulfite-sequencing. Warnecke *et al.* in 1997 first reported both overestimation and underestimation in methylation-specific PCR (MSP) due to "PCR bias"[54]. However, only a few follow-up papers paid attention to this bias and most of them mainly focused on overestimation[55,107,108]. BSM-SiMREPS independently supported underestimation of DNA methylation using PCR-based approaches by using an orthogonal, amplification-free approach, applying instead single-molecule fluorescence kinetic fingerprinting.

However, BSM-SiMREPS only provided evidence for an underestimation of DNA methylation in the BCAT1 promoter, a particular genomic locus. The naturally following question is whether there are any other genomic loci where methylation is underestimated by current gold-standard approaches (bisulfite sequencing and methylation EPIC array). To answer this question, more BSM-SiMREPS assays have to be developed and used for screening a panel of DNA disease biomarkers. In fact, the sensor constructs of BSM-SiMREPS were designed to easily adapt for detecting different targets. The imager-binding sequences on two overhangs of auxiliary probes are independent of the target sequence. Therefore, in principle, we can easily change sequences of capture probes and auxiliary probes for detecting different targets and use the same imager since the most expensive probes are the fluorophore-labeled imagers. This approach would allow for spatial multiplexing (in distinct sample wells).

5.2 Direct quantification of DNA methylation using MBD-SiMREPS

Bisulfite-coupled PCR-based detection approaches are the gold standard for DNA methylation detection. Bisulfite treatment exhibits more than 99% conversion efficiency for unmethylated cytosines and is widely used due to its low cost, compatibility with PCR and

sequencing and numerous readily available commercial kits. However, bisulfite together with high temperature is essentially a harsh redox condition for DNA. Three technical issues arise from bisulfite conversion: (i) significant loss of input DNA after purification due to degradation; (ii) biased PCR amplification due to partial depurination or depyrimidination during bisulfite conversion; (iii) false positives due to incomplete conversion of unmethylated cytosines or false negatives due to aberrant conversion of methylated cytosines. Although there have been great efforts in extensive optimizations of bisulfite conversion and alternative pretreatment approaches, a pretreatment-free amplification-free detection approach for DNA methylation has been a long-standing missing tool.

Therefore, in Chapter 3, we developed MBD-SiMREPS for direct quantification of DNA methylation using a fluorophore-labeled MBD imager. Two constructs with different labeling strategies were tested to establish a functional protein construct as a SMFKF sensor binding methyl CpG. Gy-hMBD fuses a 11-peptide ybbR tag, DSLEFIASKLA, with the hMBD1 domain (aa 1-77). Following expression and purification of Gy-hMBD, an enzymatic site-specific labeling reaction installed Cy5-CoA on the serine side chain of the ybbR tag. However, Cy5-Gy-hMBD formed aggregates during the labeling process and no soluble labeled protein could be recovered. We therefore tested Halo-hMBD that fuses a HaloTag (34 kD) with hMBD1 (aa 1-77) for site-specific covalent labeling with AF660. Halo-hMBD was expressed, purified and functionally characterized by EMSA, demonstrating methyl-CpG specific binding activity. Surprisingly, EMSA showed just weak binding affinity of Halo-hMBD or AF660-Halo-hMBD compared to the case of Gy-hMBD where a distinct band of protein-DNA complex was formed, suggesting a weaker K_d potentially suited for SMFKF measurement. The ensemble-level gel

results laid the groundwork for single-molecule observation of AF660-Halo-hMBD interacting with methyl-CpG.

Through thorough investigation and optimization of the imaging conditions during SMFKF measurement, an optimal low-salt Mg^{2+}-free condition was identified for observing methyl-CpG-specific interactions with AF660-Halo-hMBD. Sensor constructs with different positioning and number of methyl-CpGs were tested for studying their effects on methyl-CpG binding kinetics. We further discovered that a hemimethylated sensor construct was interacting with AF660-Halo-hMBD only when a "branch" motif was present at the 5' end of the auxiliary probe. Neither a "branch" motif at the 5' end of the capture probe nor a branch-free construct showed methyl-CpG binding activity for a hemimethylated sensor. This unique response of AF660-Halo-hMBD to architectural changes in the dsDNA substrate was first reported here by SMFKF measurement.

The above single-molecule characterization laid the groundwork for developing MBD-SiMREPS as a direct, amplification-free tool for quantifying methyl-CpG. In fact, either a "branch"-containing sensor construct, or a "branch"-free sensor construct may be used for detecting strand-specific DNA methylation. In the former case, a fully complementary biotinylated capture probe and an auxiliary probe with 5' overhang are required for observing fully methylated BCAT1 promoter, specifically the forward sequence. And neither the capture probe nor the auxiliary probe need to be methylated. In the latter case of a "branch"-free sensor construct, a fully complementary biotinylated capture probe and a fully complementary auxiliary probe with full methylation were found to be necessary for detecting methylated forward BCAT1 promoter. This type of sensor cannot only detect fully methylated target but also targets with just two or three methylation sites. However, it remains a question whether modulating the

positioning and number of methylation sites on the fully complementary auxiliary probe can restrict signal responses to a specific positioning and number of methylation sites on the target. It would be fascinating to develop a sensor that directly responds to the positioning of DNA methylation sites and quantify their distribution.

Apart from analytical development, the biological insights of modulation of hMBD binding to different patterns of DNA methylation are unknown. Only a few publications have studied MBD-binding dynamics to dsDNA at the single-molecule level in vitro[136,138,139,142], not to mention its biological importance in vivo. Next-generation sequencing and methylation microarrays are still the dominant if not the only approaches for studying biological functions of DNA methylation. Our results and conclusions in Chapter 3 were merely the first step in a detailed characterization of the binding kinetics of hMBD to methyl-CpG in different sequence and methylation contexts using single-molecule approaches. We hope to see more follow-up work in this field.

5.3 Two-second DNA biomarker detection through FG-SiMREPS

In Chapter 4, we conceived of a solution to break the limit of detection of SMFKF-based detection by combing total accepted counts of hundreds of FOVs into a single readout. To implement this strategy, two practical concerns need to be addressed: (i) false positives coming from off-target signals should not increase as more and more FOVs are collected; (ii) the total acquisition time should be within 1 h for a good analytical time efficiency. Inspired by Chung *et al.* in 2022[140], we utilized a fluorogenic DNA imager that allowed us to use a much higher imager concentration at the μM level, enabling ultrafast detection of DNA disease biomarkers within seconds.

To demonstrate the idea of FG-SiMREPS (fluorogenic single molecule recognition by equilibrium through Poisson sampling), we rationally designed the sensor constructs for detecting three DNA cancer biomarkers: T790M, L858R and HPV. A programming-assisted design pipeline was established by screening and ranking all possible fluorogenic imager sequences by scores calculated from their predicted target binding affinity and low probability of forming secondary self-structures. Top-ranked imager sequences for T790M, L858R and HPV were able to achieve two-second detections after step-by-step optimization of the capture and imaging conditions without the need for further redesign, demonstrating great robustness of our semi-automatic design pipeline for a mismatch-containing fluorogenic imager at the single-molecule level. Finally, we applied the optimized imaging conditions to a 104-FOV detection of each of the three targets. However, a linear increase in false positives in wildtype and blank control experiments was observed in all cases. Moreover, only the distribution of accepted counts across 104 FOVs for T790M detection exhibited a distinct peak for mutant T790M. Detection of L858R and HPV both showed a "uniform"-like distribution.

The non-Poissonian broad distributions across 104 FOVs in the detection of both L858R and HPV might result from uneven surface capture. In fact, this type of distribution was also accidentally observed in detecting T790M. Due to the small size of the bottom hole of 3D-printed sample wells used for the 104-FOV detection, repeated pipetting might scratch and damage the detection surface, causing significant loss of captured molecules. Another reason was likely to be the significant number of false positives in wildtype and blank control. Populations of mutant-specific signals highly overlapped for the wildtype and blank signals, forcing filtering thresholds to reject many likely true mutant signals in order to maintain a low level of false

positives. Therefore, a significant loss of rejected true signals might also contribute to the "uniform" distribution.

The motivation of using 3D-printed sample wells was to achieve detection of the entire capture surface by using a small bottom hole. However, this practice does not increase sensitivity since there is no enrichment of target molecules. In other words, the surface density of captured molecules remained the same regardless of size of the capture surface. Leakage by epoxy when gluing sample wells on coverslips as well as pipetting scratches also caused uncontrolled disturbance to the surface. Switching back to the cut pipette tips is an easier and better practice. Another problem is the significant number of false positives in the wildtype and blank samples. One solution we had in mind was to add a fluorogenic imager with the same sequence but labeled with a different fluorophore. We expect that colocalization in both emission channels should allow us to remove off-target interactions by tandem kinetic filtering in individual channels, as the combined channels should be strictly confined to mutant-specific signals. More optimization to reach specificity is necessary before developing FG-SiMREPS into a mature analytical tool. Standard curves and quantifications in biological samples are needed for demonstrating its diagnostic and prognostic promises.

5.4 Outlook: in situ methyl-CpG profiling by expansion localization microscopy

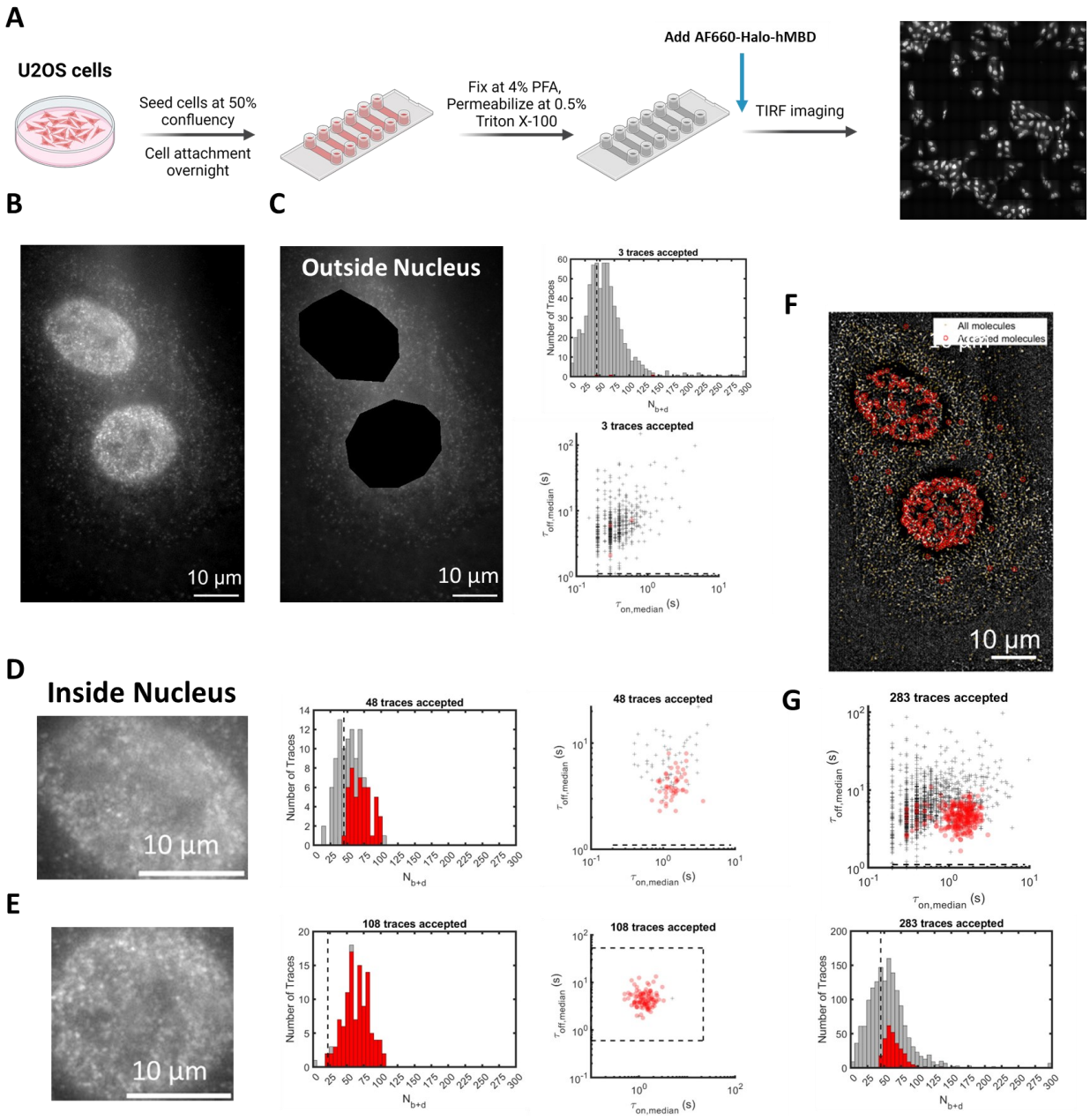

Figure 5.1. In situ methyl-CpG profiling using AF660-Halo-hMBD with fixed U2OS cells. Imaging conditions: 10 nM AF660-Halo-hMBD, 50 mM Tris-HCl pH 8.0, 100 ms exposure, 5 min, 20% 640 nm excitation. Filtering parameters are optimized with signals outside nucleus as negative training dataset and signals inside nucleus as positive training dataset. (A) Pipeline of fixation, permeabilization and TIRF imaging of U2OS cells with AF660-Halo-hMBD. (B) Screenshot of a 5-min movie containing 2 adjacent U2OS cells. (C) After parameter optimizations, distribution of N_{b+d} and dwell times of signals outside nucleus. (D)&(E) After parameter optimizations, distribution of N_{b+d} and dwell times of signals inside nucleus. (F)&(G) After parameter optimizations, distribution of N_{b+d} and dwell times of signals across entire cells. In panel F, red circles are accepted molecules and yellow spots are all identified molecules.

This entire dissertation focuses on the development of analytical tools for the quantitative measurement of DNA disease biomarkers using the principle of SiMREPS. In fact, SiMREPS is essentially a kinetics measurement and can be directly applied to any biological system. As one potential application, **Figure 5.1** pilots the direct application of AF660-Halo-hMBD described in Chapter 3 in situ for quantification of DNA methylation in fixed U2OS cells. By optimizing kinetic filtering criteria for separating signals inside and outside the nucleus, we clearly showed that SiMREPS can distinguish them. However, it was not clear whether these signal differences were due to the affinity of AF660-Halo-hMBD to methyl-CpG or to DNA itself. Nonetheless, these preliminary data supported the notion that SiMREPS observation of binding kinetics in situ is possible.

Finally, a bold but exciting direction is using AF660-Halo-hMBD to identify the spatial distribution of methyl-CpG clusters in the nucleus and characterize their underlying 3D chromosome structures. Approximately 60 million methyl-CpG base pairs existed in a single human cell and are highly compacted inside just a small nucleus of 5-20 μm diameter. Current super-resolution microscopy can resolve around 10 nm structural features laterally and around 50 nm axially but is still insufficient for resolving single methyl-CpG dinucleotides (assuming a uniformly 3D-distributed methyl-CpG inside nucleus, 2 neighboring methyl-CpG are separated by just 12 nm ~ 50 nm). One solution would be to combine expansion microscopy with super-resolution imaging. AF660-Halo-hMBD might be the perfect imager for expansion microscopy due to its simple imaging condition with low-salt. In fact, I made several attempts of a single-round expansion of U2OS cells in 10 mM Tris-HCl pH 8.0 and achieved around 2-fold expansion. Iterative expansion can also be used to further increase the size of a nucleus, and an expansion factor of >4 would make the nucleus ready for spatial methyl-CpG profiling using

AF660-Halo-hMBD. In the future, I hope to obtain the first full spatial map of methyl-CpG and its underlying high-resolution chromatin structure independently of sequencing-based approaches by using such a direct single-molecule methyl-CpG imager.

Bibliography

1. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum Genet* **122**, 565–581 (2008).

2. Levene, P. A. THE STRUCTURE OF YEAST NUCLEIC ACID: IV. AMMONIA HYDROLYSIS. *Journal of Biological Chemistry* **40**, 415–424 (1919).

3. Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* **6**, 201–209 (1950).

4. Griffith, F. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27**, 113–159 (1928).

5. Avery, O. T., Macleod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med* **79**, 137–158 (1944).

6. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

7. McCarty, M. Discovering genes are made of DNA. *Nature* **421**, 406–406 (2003).

8. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **69**, 89–95 (2001).

9. Ziegler, A., Koch, A., Krockenberger, K. & Großhennig, A. Personalized medicine using DNA biomarkers: a review. *Hum Genet* **131**, 1627–1638 (2012).

10. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68 (2015).

11. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* **48**, D941–D947 (2020).

12. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

13. Eichler, E. E. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N Engl J Med* **381**, 64–74 (2019).

14. Saiki, R. K. *et al.* Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science* **230**, 1350–1354 (1985).

15. Rees, D. C., Williams, T. N. & Gladwin, M. T. Sickle-cell disease. *The Lancet* **376**, 2018–2031 (2010).

16. Gray, I. C. Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics* **9**, 2403–2408 (2000).

17. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet* **32**, 650–654 (2002).

18. Collins, F. S. *et al.* Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**, 1046–1049 (1987).

19. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics* **19**, R131–R136 (2010).

20. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet* **24**, 238–245 (2008).

21. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125–138 (2013).

22. Weigele, P. & Raleigh, E. A. Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. *Chem. Rev.* **116**, 12655–12687 (2016).

23. Sood, A. J., Viner, C. & Hoffman, M. M. DNAmod: the DNA modification database. *Journal of Cheminformatics* **11**, 30 (2019).

24. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23–38 (2013).

25. Kumar, S., Chinnusamy, V. & Mohapatra, T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front. Genet.* **9**, 640 (2018).

26. Johnson, T. B. & Coghill, R. D. RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS [1]. *J. Am. Chem. Soc.* **47**, 2838–2844 (1925).

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

28. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**, 116–117 (1998).

29. Esteller, M. *et al.* Inactivation of the DNA-Repair Gene *MGMT* and the Clinical Response of Gliomas to Alkylating Agents. *N Engl J Med* **343**, 1350–1354 (2000).

30. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* **21**, 1592–1600 (2011).

31. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).

32. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucl Acids Res* **10**, 2709–2721 (1982).

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

34. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **8**, a019505 (2016).

35. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010–1022 (2011).

36. Locke, W. J. *et al.* DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Frontiers in Genetics* **10**, (2019).

37. Mikeska, T. & Craig, J. DNA Methylation Biomarkers: Cancer and Beyond. *Genes* **5**, 821–864 (2014).

38. Papanicolau-Sengos, A. & Aldape, K. DNA Methylation Profiling: An Emerging Paradigm for Cancer Diagnosis. *Annual Review of Pathology: Mechanisms of Disease* **17**, 295–321 (2022).

39. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat Rev Cancer* **11**, 726–734 (2011).

40. Hegi, M. E. *et al. MGMT* Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N Engl J Med* **352**, 997–1003 (2005).

41. Du, Y. & Dong, S. Nucleic Acid Biosensors: Recent Advances and Perspectives. *Anal. Chem.* **89**, 189–215 (2017).

42. Bhattacharjee, R., Moriam, S., Umer, M., Nguyen, N.-T. & Shiddiky, M. J. A. DNA methylation detection: recent developments in bisulfite free electrochemical and optical approaches. *Analyst* **143**, 4802–4818 (2018).

43. Lu, K. *et al.* A Review of Stable Isotope Labeling and Mass Spectrometry Methods to Distinguish Exogenous from Endogenous DNA Adducts and Improve Dose–Response Assessments. *Chem. Res. Toxicol.* **35**, 7–29 (2022).

44. Kumar, R. R., Kumar, A., Chuang, C.-H. & Shaikh, M. O. Recent Advances and Emerging Trends in Cancer Biomarker Detection Technologies. *Ind. Eng. Chem. Res.* **62**, 5691–5713 (2023).

45. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51 Pt 1**, 263–273 (1986).

46. The Nobel Prize in Chemistry 1993. *NobelPrize.org* https://www.nobelprize.org/prizes/chemistry/1993/press-release/.

47. Chen, B. *et al.* Evaluation of Droplet Digital PCR Assay for the Diagnosis of Candidemia in Blood Samples. *Front. Microbiol.* **12**, 700008 (2021).

48. Berden, P. *et al.* Amplification Efficiency and Template Accessibility as Distinct Causes of Rain in Digital PCR: Monte Carlo Modeling and Experimental Validation. *Anal. Chem.* **94**, 15781–15789 (2022).

49. Fakruddin, M. *et al.* Nucleic acid amplification: Alternative methods of polymerase chain reaction. *J Pharm Bioall Sci* **5**, 245 (2013).

50. Sang, P. *et al.* Nucleic Acid Amplification Techniques in Immunoassay: An Integrated Approach with Hybrid Performance. *J. Agric. Food Chem.* **69**, 5783–5797 (2021).

51. Ravan, H., Kashanian, S., Sanadgol, N., Badoei-Dalfard, A. & Karami, Z. Strategies for optimizing DNA hybridization on surfaces. *Analytical Biochemistry* **444**, 41–46 (2014).

52. Chatterjee, T. *et al.* Direct kinetic fingerprinting and digital counting of single protein molecules. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22815–22822 (2020).

53. Meddeb, R. *et al.* Quantifying circulating cell-free DNA in humans. *Sci Rep* **9**, 5220 (2019).

54. Warnecke, P. M. *et al.* Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Research* **25**, 4422–4426 (1997).

55. Taryma-Lesniak, O., Kjeldsen, T. E., Hansen, L. L. & Wojdacz, T. K. Influence of Unequal Amplification of Methylated and Non-Methylated Template on Performance of Pyrosequencing. *Genes* **13**, 1418 (2022).

56. Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications. *Nat Microbiol* **2**, 1–9 (2017).

57. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* **18**, 67–83 (2020).

58. Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nat Biomed Eng* **5**, 643–656 (2021).

59. Ghouneimy, A., Mahas, A., Marsic, T., Aman, R. & Mahfouz, M. CRISPR-Based Diagnostics: Challenges and Potential Solutions toward Point-of-Care Applications. *ACS Synth. Biol.* **12**, 1–16 (2023).

60. East-Seletsky, A. *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270–273 (2016).

61. Hajian, R. *et al.* Detection of unamplified target genes via CRISPR–Cas9 immobilized on a graphene field-effect transistor. *Nat Biomed Eng* **3**, 427–437 (2019).

62. Dai, Y. *et al.* Exploring the Trans-Cleavage Activity of CRISPR-Cas12a (cpf1) for the Development of a Universal Electrochemical Biosensor. *Angew Chem Int Ed* **58**, 17399–17405 (2019).

63. Bruch, R. *et al.* CRISPR/Cas13a-Powered Electrochemical Microfluidic Biosensor for Nucleic Acid Amplification-Free miRNA Diagnostics. *Advanced Materials* **31**, 1905311 (2019).

64. Fozouni, P. *et al.* Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **184**, 323-333.e9 (2021).

65. Ying, Y.-L. *et al.* Nanopore-based technologies beyond DNA sequencing. *Nat. Nanotechnol.* **17**, 1136–1146 (2022).

66. Șoldănescu, I., Lobiuc, A., Covașă, M. & Dimian, M. Detection of Biological Molecules Using Nanopore Sensing Techniques. *Biomedicines* **11**, 1625 (2023).

67. Chen, X. *et al.* Nanopore single-molecule analysis of biomarkers: Providing possible clues to disease diagnosis. *TrAC Trends in Analytical Chemistry* **162**, 117060 (2023).

68. Burck, N. *et al.* Nanopore Identification of Single Nucleotide Mutations in Circulating Tumor DNA by Multiplexed Ligation. *Clinical Chemistry* **67**, 753–762 (2021).

69. Bhatti, H., Lu, Z. & Liu, Q. Nanopore Detection of Cancer Biomarkers: A Challenge to Science. *Technol Cancer Res Treat* **21**, 153303382210766 (2022).

70. Johnson-Buck, A. *et al.* Kinetic fingerprinting to identify and count single nucleic acids. *Nat Biotechnol* **33**, 730–732 (2015).

71. Su, X. *et al.* Single-Molecule Counting of Point Mutations by Transient DNA Binding. *Sci Rep* **7**, 43824 (2017).

72. Hayward, S. L. *et al.* Ultraspecific and Amplification-Free Quantification of Mutant DNA by Single-Molecule Kinetic Fingerprinting. *J. Am. Chem. Soc.* **140**, 11755–11762 (2018).

73. Khanna, K. *et al.* Rapid kinetic fingerprinting of single nucleic acid molecules by a FRET-based dynamic nanosensor. *Biosensors and Bioelectronics* **190**, 113433 (2021).

74. Li, Z. *et al.* Attomolar Sensitivity in Single Biomarker Counting upon Aqueous Two-Phase Surface Enrichment. *ACS Sens.* **7**, 1419–1430 (2022).

75. Chatterjee, T., Johnson-Buck, A. & Walter, N. G. Highly sensitive protein detection by aptamer-based single-molecule kinetic fingerprinting. *Biosensors and Bioelectronics* **216**, 114639 (2022).

76. Li, Z. *et al.* Probing DNA Hybridization Equilibrium by Cationic Conjugated Polymer for Highly Selective Detection and Imaging of Single-Nucleotide Mutation. *Anal. Chem.* **90**, 6804–6810 (2018).

77. Yu, Y. *et al.* Digestion of Dynamic Substrate by Exonuclease Reveals High Single-Mismatch Selectivity. *Anal. Chem.* **90**, 13655–13662 (2018).

78. Li, L., Yu, Y., Wang, C., Han, Q. & Su, X. Transient Hybridization Directed Nanoflare for Single-Molecule miRNA Imaging. *Anal. Chem.* **91**, 11122–11128 (2019).

79. Mandal, S. *et al.* Direct Kinetic Fingerprinting for High-Accuracy Single-Molecule Counting of Diverse Disease Biomarkers. *Acc. Chem. Res.* **54**, 388–402 (2021).

80. Chatterjee, T. *et al.* Ultraspecific analyte detection by direct kinetic fingerprinting of single molecules. *TrAC Trends in Analytical Chemistry* **123**, 115764 (2020).

81. Johnson-Buck, A., Li, J., Tewari, M. & Walter, N. G. A guide to nucleic acid detection by single-molecule kinetic fingerprinting. *Methods* **153**, 3–12 (2019).

82. Montoya, K. Direct Identification and Counting of MicroRNAs in Single Cells by Transient Binding and Kinetic Fingerprinting. (the University of Michigan, 2023).

83. Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L. & Wiggins, C. H. Learning Rates and States from Biophysical Time Series: A Bayesian Approach to Model Selection and Single-Molecule FRET Data. *Biophysical Journal* **97**, 3196–3205 (2009).

84. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590–607 (2019).

85. Jedi, M., Young, G. P., Pedersen, S. K. & Symonds, E. L. Methylation and Gene Expression of BCAT1 and IKZF1 in Colorectal Cancer Tissues. *Clin Med Insights Oncol* **12**, 1179554918775064 (2018).

86. Luo, H., Wei, W., Ye, Z., Zheng, J. & Xu, R. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends in Molecular Medicine* **27**, 482–500 (2021).

87. Karpf, A. R. & Jones, D. A. Reactivating the expression of methylation silenced genes in human cancer. *Oncogene* **21**, 5496–5503 (2002).

88. Laird, P. W. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **3**, 253–266 (2003).

89. Bruschi, M. The Epigenetic Progenitor Origin of Cancer Reassessed: DNA Methylation Brings Balance to the Stem Force. *Epigenomes* **4**, 8 (2020).

90. Ymd, L., Dsc, H., P, J. & Rwk, C. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science (New York, N.Y.)* **372**, (2021).

91. Pedersen, S. K. *et al.* A Two-Gene Blood Test for Methylated DNA Sensitive for Colorectal Cancer. *PLOS ONE* **10**, e0125041 (2015).

92. Premarket Approval (PMA). https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma.cfm?id=P130001.

93. Taieb, J. *et al.* Analysis of circulating tumour DNA (ctDNA) from patients enrolled in the IDEA-FRANCE phase III trial: Prognostic and predictive value for adjuvant treatment duration. *Annals of Oncology* **30**, v867 (2019).

94. Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175**, 315–332 (1948).

95. Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends Genet* **38**, 676–707 (2022).

96. Herman, J. G., Graff, J. R., Myöhänen, S., Nelkin, B. D. & Baylin, S. B. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A* **93**, 9821–9826 (1996).

97. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827–1831 (1992).

98. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126**, 1189–1201 (2006).

99. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).

100. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).

101. Jacobs, B. K., Goetghebeur, E. & Clement, L. Impact of variance components on reliability of absolute quantification using digital PCR. *BMC Bioinformatics* **15**, 283 (2014).

102. Maden, S. K., Thompson, R. F., Hansen, K. D. & Nellore, A. Human methylome variation across Infinium 450K data on the Gene Expression Omnibus. *NAR Genom Bioinform* **3**, lqab025 (2021).

103. Maden, S. K. *et al.* recountmethylation enables flexible analysis of public blood DNA methylation array data. *Bioinformatics Advances* **3**, vbad020 (2023).

104. Mandal, S., Khanna, K., Johnson-Buck, A. & Walter, N. G. A guide to accelerated direct digital counting of single nucleic acid molecules by FRET-based intramolecular kinetic fingerprinting. *Methods* **197**, 63–73 (2022).

105. Stejskal, P. *et al.* Circulating tumor nucleic acids: biology, release mechanisms, and clinical relevance. *Molecular Cancer* **22**, 15 (2023).

106. Neumann, M. H. D., Bender, S., Krahn, T. & Schlange, T. ctDNA and CTCs in Liquid Biopsy – Current Status and Where We Need to Progress. *Computational and Structural Biotechnology Journal* **16**, 190–195 (2018).

107. Wojdacz, T. K., Borgbo, T. & Hansen, L. L. Primer design versus PCR bias in methylation independent PCR amplifications. *Epigenetics* **4**, 231–234 (2009).

108. Moskalev, E. A. *et al.* Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Research* **39**, e77–e77 (2011).

109. Kaur, D. *et al.* Comprehensive evaluation of the Infinium human MethylationEPIC v2 BeadChip. *Epigenetics Communications* **3**, 6 (2023).

110. Khodadadi, E. *et al.* Current Advances in DNA Methylation Analysis Methods. *Biomed Res Int* **2021**, 8827516 (2021).

111. Clouaire, T. & Stancheva, I. Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? *Cell. Mol. Life Sci.* **65**, 1509–1522 (2008).

112. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics* **7**, 1051–1073 (2015).

113. Furuta, Y. & Kobayashi, I. Restriction-Modification Systems as Mobile Epigenetic Elements. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).

114. Jørgensen, H. F. & Bird, A. MeCP2 and other methyl-cpg binding proteins. *Ment. Retard. Dev. Disabil. Res. Rev.* **8**, 87–93 (2002).

115. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews* **77**, 53–72 (2013).

116. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* **20**, 157–172 (2019).

117. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLOS Genetics* **12**, e1005854 (2016).

118. Casadesús, J. & Sánchez-Romero, M. A. DNA Methylation in Prokaryotes. in *DNA Methyltransferases - Role and Function* (eds. Jeltsch, A. & Jurkowska, R. Z.) 21–43 (Springer International Publishing, Cham, 2022). doi:10.1007/978-3-031-11454-0_2.

119. Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* **18**, 491–498 (2021).

120. Li, L., Chen, B.-F. & Chan, W.-Y. An Epigenetic Regulator: Methyl-CpG-Binding Domain Protein 1 (MBD1). *IJMS* **16**, 5125–5140 (2015).

121. Wood, K. H. & Zhou, Z. Emerging Molecular and Biological Functions of MBD2, a Reader of DNA Methylation. *Front. Genet.* **7**, (2016).

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

123. Inomata, K. *et al.* Kinetic and Thermodynamic Evidence for Flipping of a Methyl-CpG Binding Domain on Methylated DNA. *Biochemistry* **47**, 3266–3271 (2008).

124. Yu, Y. *et al.* Direct DNA Methylation Profiling Using Methyl Binding Domain Proteins. *Anal. Chem.* **82**, 5012–5019 (2010).

125. Heimer, B. W., Tam, B. E. & Sikes, H. D. Characterization and directed evolution of a methyl-binding domain protein for high-sensitivity DNA methylation analysis. *Protein Engineering, Design and Selection* **28**, 543–551 (2015).

126. Buchmuller, B. C., Kosel, B. & Summerer, D. Complete Profiling of Methyl-CpG-Binding Domains for Combinations of Cytosine Modifications at CpG Dinucleotides Reveals Differential Read-out in Normal and Rett-Associated States. *Sci Rep* **10**, 4053 (2020).

127. Serre, D., Lee, B. H. & Ting, A. H. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Research* **38**, 391–399 (2010).

128. Bock, C. *et al.* Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat Biotechnol* **34**, 726–737 (2016).

129. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28**, 1097–1105 (2010).

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

131. Liu, K. *et al.* Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *Journal of Biological Chemistry* **293**, 7344–7354 (2018).

132. Boyd, M. E., Heimer, B. W. & Sikes, H. D. Functional heterologous expression and purification of a mammalian methyl-CpG binding domain in suitable yield for DNA methylation profiling assays. *Protein Expression and Purification* **82**, 332–338 (2012).

133. Lotze, J., Reinhardt, U., Seitz, O. & Beck-Sickinger, A. G. Peptide-tags for site-specific protein labelling in vitro and in vivo. *Mol. BioSyst.* **12**, 1731–1745 (2016).

134. Los, G. V. *et al.* HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis. *ACS Chem. Biol.* **3**, 373–382 (2008).

135. England, C. G., Luo, H. & Cai, W. HaloTag Technology: A Versatile Platform for Biomedical Applications. *Bioconjugate Chem.* **26**, 975–986 (2015).

136. Pan, H. *et al.* CpG and methylation-dependent DNA binding and dynamics of the methylcytosine binding domain 2 protein at the single-molecule level. *Nucleic Acids Research* **45**, 9164–9177 (2017).

137. Qin, J. *et al.* Investigation of the interaction between MeCP2 methyl-CpG binding domain and methylated DNA by single molecule force spectroscopy. *Analytica Chimica Acta* **1124**, 52–59 (2020).

138. Leighton, G. O. *et al.* Densely methylated DNA traps Methyl-CpG–binding domain protein 2 but permits free diffusion by Methyl-CpG–binding domain protein 3. *J Biol Chem* **298**, 102428 (2022).

139. Strauskulage, L. Building a novel single-molecule system to study readers of DNA methylation with high resolution reveals binding preferences of MBD proteins. (the UNIVERSITY OF CALIFORNIA, SAN FRANCISCO, 2023).

140. Chung, K. K. H. *et al.* Fluorogenic DNA-PAINT for faster, low-background super-resolution imaging. *Nat Methods* **19**, 554–559 (2022).

141. Khanna, K. *et al.* Rapid kinetic fingerprinting of single nucleic acid molecules by a FRET-based dynamic nanosensor. *Biosens Bioelectron* **190**, 113433 (2021).

142. Qin, J. *et al.* Investigation of the interaction between MeCP2 methyl-CpG binding domain and methylated DNA by single molecule force spectroscopy. *Analytica Chimica Acta* **1124**, 52–59 (2020).