Springer Theses
Recognizing Outstanding Ph.D. Research

Anandhakumar Chandran

Advancing Development of Synthetic Gene Regulators

With the Power of High-Throughput Sequencing in Chemical Biology

Springer

Springer Theses

Recognizing Outstanding Ph.D. Research

Aims and Scope

The series “Springer Theses” brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student’s supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today’s younger generation of scientists.

Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

More information about this series at http://www.springer.com/series/8790

Anandhakumar Chandran

Advancing Development of Synthetic Gene Regulators

With the Power of High-Throughput Sequencing in Chemical Biology

Doctoral Thesis accepted by
Kyoto University, Kyoto, Japan

Springer

Author
Dr. Anandhakumar Chandran
Department of Chemistry,
Graduate School of Science
Kyoto University
Kyoto
Japan

Supervisor
Prof. Hiroshi Sugiyama
Department of Chemistry,
Graduate School of Science
Kyoto University
Kyoto
Japan

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-981-10-6546-0 ISBN 978-981-10-6547-7 (eBook)
https://doi.org/10.1007/978-981-10-6547-7

Library of Congress Control Number: 2017952918

© Springer Nature Singapore Pte Ltd. 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

This book is lovingly and sincerely dedicated to

My brother
Alex V.F. Paul Menon, IAS
Who is my great source of inspiration and keep me motivated every time

and

My love
Dr. Junetha Syed
Who is always supporting, helping and standing by me.

Supervisor's Foreword

We are witnessing a golden era of genetics, which has led to development of target-specific drug discovery. Next-generation sequencing (NGS) plays an important role in this improvement and is used commonly in the field of molecular biology. The divergent application of NGS technologies is at the interface of chemistry and biology, with a special emphasis on small molecule development and screening.

The development of small molecules targeting specific genomic sequence is an attractive goal in personalized medicine. Hairpin *N*-methylpyrrole (P)–*N*-methylimidazole (I) polyamides (PIPs) are a class of small molecule that can bind in the minor groove of DNA.

In this thesis by Dr. Anandhakumar Chandran, the author reviewed multitasking high-throughput sequencing technologies applications for chemical biologists, together with NGS-based methods to identify small molecules genomic effect. The author developed Bind-n-seq, a cost-effective high-throughput sequencing-based method, to identify the binding specificity of PIP conjugates in a randomized DNA library. This aids the specific redesigning of PIP conjugates.

Further, the biological effects of PIPs primarily rely on its genome-wide binding preferences. The author established a new genome-wide assessment method using high-throughput sequencing to map the differential binding sites and relative enriched regions of non-cross-linked SAHA-PIPs throughout the human genome. SAHA-PIPs' binding motifs were identified, and our genome-level mapping of SAHA-PIPs' enriched region provided insightful evidence of the SAHA-PIPs' working mechanism on the silenced gene network.

The author also developed a method using high-throughput sequencing to map the binding sites and relative enriched regions of alkylating PIP throughout the human genome. Our genome-level mapping of alkylating PIP-enriched region and the binding sites on human genome identifies significant genomic targets of breast cancer.

This thesis demonstrates the potential of high-throughput sequencing technologies in small molecule design and valuation. Understanding the binding specificity of DNA binding small molecules will be beneficial for the development of small molecule drugs.

Kyoto, Japan
August 2017

Prof. Hiroshi Sugiyama

Part of this Thesis have been Published in the Following Journal Articles:

1. Chandran A, Syed J, Li Y, Sato S, Bando T, Sugiyama H (2016) "Genome-Wide Assessment of the Binding Effects of Artificial Transcriptional Activators by High-Throughput Sequencing." Chembiochem 17:1905–1910. doi: 10.1002/cbic.201600274.
2. Chandran A, Syed J, Taylor RD, Kashiwazaki G, Sato S, Hashiya K, Bando T, Sugiyama H (2016) "Deciphering the genomic targets of alkylating polyamide conjugates using high-throughput sequencing." Nucleic Acids Res 44:4014–4024. doi: 10.1093/nar/gkw283.
3. Anandhakumar C, Kizaki S, Bando T, Pandian GN, Sugiyama H (2015) "Advancing small-molecule-based chemical biology with next-generation sequencing technologies." ChemBioChem 16:20–38.doi: 10.1002/cbic.20140 2556.
4. Anandhakumar C, Li Y, Kizaki S, Pandian GN, Hashiya K, Bando T, Sugiyama H (2014) "Next-generation sequencing studies guide the design of pyrrole-imidazole polyamides with improved binding specificity by the addition of β-alanine." ChemBioChem 15:2647–2651. doi: 10.1002/cbic.201402497.

Acknowledgements

> *Better than a thousand days of diligent study is one day with a great teacher*
>
> —Japanese proverb.

First and foremost, I would like to express my sincere gratitude to my supervisor **Prof. Hiroshi Sugiyama**, for his motivation, patience and immense knowledge. His guidance, research discussions and suggestions helped me throughout my Ph.D.

Besides my supervisor, I would like to thank Dr. G.N. Pandian & Dr. T. Bando for their valuable discussions, insightful comments and suggestions, it improved my study in a better shape. I also thank Dr. M. Endo & Dr. S. Park for their support.

I sincerely believe that without my present and past fellow laboratory members, the study would not be possible. So I genuinely thank each and every one of my laboratory mates for their support, timely help and discussions. I also thank my friends Dr. Vasanthan Jayakumar, (Keio University) and Dr. Koichiro higasa (Centre for Genomic Medicine, Kyoto University) for their continuous support and encouragement. Now, I take the opportunity to extent my gratitude to Mr. A. Varadhas who recognized my potency and Mr. Arun Prasad IFS, Dr. Rajkumar Rathinavelu, Dr. G. Dhinakar Raj, Dr. S. Uma, Dr. Arun Kumar and Dr. K.G. Tirumurugaan for their continuous moral support.

I would like to thank JEES Mitsubishi Corporation Scholarship for financial support and iCeMS for their iCeMS overseas visit travel grant.

To thank my parents, I bow before them; they were always supporting me and encouraging me with their best wishes.

Finally, I thank all my friends and family members for their love, motivation and support during the study.

I also place on record, my sense of gratefulness to one and all, who directly or indirectly, keep me motivated.

Contents

1 Overview of Next-Generation Sequencing Technologies and Its Application in Chemical Biology 1
1.1 General Introduction: Next Generation Sequencing (NGS) Principles and Platforms 1
1.1.1 Sequencing by Synthesis (SBS) 3
1.1.2 Single-Molecule Sequencing (SMS) 6
1.1.3 Sequencing by Ligation (SBL) 7
1.2 Applications of NGS in Chemical Biology 10
1.2.1 Genome-Wide Localization of Non-B DNA Using Small Molecules 10
1.2.2 Decoding of DNA Base Modification in Single Molecule Level 12
1.2.3 Aptamer Selection Using Massively Parallel Sequencing 18
1.3 Development of *N*-Methylpyrrole (P)—*N*-Methylimidazole (I) Polyamides (PIP) and Its Conjugates 20
1.4 PIP Based Gene Regulation 22
1.5 Utilization of NGS in PIP Based Small Molecule Studies 23
1.5.1 Guiding the Design and Screening of Small Molecule Using NGS 23
1.5.2 Analysis of Gene Expression Induced by PIP Based Small Molecules 26
1.6 Conclusion and Future Prospects 31
References 33

2 Next Generation Sequencing Studies Guide the Design of Pyrrole-Imidazole Polyamides with Improved Binding Specificity by the Addition of β-Alanine 43
2.1 Introduction 43
2.2 Results and Discussion 46

2.2.1 PIP and PIP Conjugates Synthesis and Bind-n-Seq Analysis 46
2.2.2 PIP Redesign, ββ-PIP Synthesis and Bind-n-Seq Analysis 49
2.2.3 Binding Affinity Conformation by SPR. 51
2.2.4 Discussion. 52
2.3 Materials and Methods 55
2.3.1 General Materials and Synthesis 55
2.3.2 Oligomer Sequences Used in Bind-n-Seq Method. 57
2.3.3 Bind-n-Seq 58
2.3.4 Surface Plasmon Resonance (SPR) 59
References. 60

3 Genome-Wide Assessment of the Binding Effects of Artificial Transcriptional Activators by Utilizing the Power of High-Throughput Sequencing 63
3.1 Introduction 63
3.2 Results and Discussion. 65
3.2.1 Revealing the Unique Gene Set Activation Mechanism of PIP Conjugates in the Human Genome. 69
3.2.2 Identification of the Preferential Binding Region of PIP Conjugates in the Tightly Packed Heterochromatin Region 70
3.2.3 Discussion. 71
3.3 Materials and Methods 73
3.3.1 Nucleus Extraction and PIP Conjugate Incubation 73
3.3.2 MNase Digestion 73
3.3.3 Affinity Purification. 74
3.3.4 High-Throughput Sequencing Library Construction and Sequencing. 74
Appendix 76
References. 77

4 Deciphering the Genomic Targets of Alkylating Polyamide Conjugates Using High-Throughput Sequencing. 81
4.1 Introduction 81
4.2 Results and Discussion. 84
4.2.1 Bind-n-Seq with PIP-Indole-*Seco*-CBI Conjugate 1 84
4.2.2 Bind-n-Seq with PIP-Indole-*Seco*-CBI Conjugate 2 86
4.2.3 Bind-n-Seq with PIP-Conjugate 5. 86
4.2.4 Identification of PIP-Indole-*Seco*-CBI Conjugate 2 High Affinity DNA-Alkylating Site in Chromatinized Human Genome 86

4.2.5 Identification of PIP-Indole-*Seco*-CBI Conjugate 2 Representative Enriched Sites in Chromatinized Human Genome 90
4.2.6 Genome-Wide Analysis of PIP-Indole-*Seco*-CBI Conjugate 2 Enriched Sites Distribution 91
4.2.7 Discussion 92
4.3 Materials and Methods 93
4.3.1 Synthesis of Biotin-Conjugated Alkylating Polyamide Conjugates 93
4.3.2 Bind-n-Seq Experiment and High-Throughput Sequencing 96
4.3.3 Affinity Purification-Based High-Throughput Sequencing of Human Genomic Regions Enriched with PIP-Indole-*Seco*-CBI Conjugate 2 97
4.3.4 Validating PIP-Indole-*Seco*-CBI Conjugate 2 Bound and Enriched Region in the Human Genome 99
Appendix 100
References 108

Curriculum Vitae 113

Chapter 1
Overview of Next-Generation Sequencing Technologies and Its Application in Chemical Biology

Abstract Next-generation-sequencing (NGS) technologies enable us to obtain extensive information by deciphering millions of individual DNA sequencing reactions simultaneously. The new DNA-sequencing strategies exceed their precursors in output by many orders of magnitude, resulting in a quantitative increase in valuable sequence information that could be harnessed for qualitative analysis. Sequencing on this scale has facilitated significant advances in diverse disciplines, ranging from the discovery, design, and evaluation of many small molecules and relevant biological mechanisms to maturation of personalized therapies. NGS technologies that have recently become affordable allow us to gain in-depth insight into small-molecule-triggered biological phenomena and empower researchers to develop advanced versions of small molecules. In this chapter we focus on the overlooked implications of NGS technologies in chemical biology, with a special emphasis on small-molecule development and screening.

Keywords Next generation sequencing · Chemical biology · Small molecule · DNA modification · Aptamer

1.1 General Introduction: Next Generation Sequencing (NGS) Principles and Platforms

Strategies to decipher DNA sequences storing huge amounts of genetic instruction create paradigm-shifting opportunities in a wide range of scientific disciplines. The Human Genome Project was completed in 2003 using a first-generation sequencing technique based almost entirely on Sanger's method. In 1977, Sanger et al. [1] described dideoxy nucleotide sequencing of DNA. In the same year, Maxam and Gilbert developed a sequencing technique based on chemical modification of DNA and consequent cleavage [2]. These two methods represent the first generation of sequencing. Sequencing has undergone steady progress from a cottage industry to a large-scale production enterprise that requires a specialized and devoted infrastructure of robotics, a modern chemical approach, bioinformatics, computer

© Springer Nature Singapore Pte Ltd. 2018

A. Chandran, *Advancing Development of Synthetic Gene Regulators*,
Springer Theses, https://doi.org/10.1007/978-981-10-6547-7_1

databases, and instrumentation. The introduction of high-throughput sequencing method in 2007 took DNA sequencing to the next level. It got established upon the notion that millions of autonomous chemical reactions taking place simultaneously, thereby a distinct molecule could be decoded in a quantifiable mode with deep coverage of sequencing reads. This strategy was called deep sequencing, next-generation sequencing (NGS), high-throughput sequencing, or massively parallel sequencing. Shortly afterward, in 2008, NGS was effectively utilized for sequencing the first individual human genome [3]. The database of the Human Genome Project led to a deeper knowledge of several disease processes at the genetic level [4, 5]. During this progress, the cost per reaction of DNA sequencing reduced, mainly because of the efforts to sequence the human genome. Hui [6] has extensively reviewed the evolution and chemistry of various sequencing technologies.

High-throughput sequencing may need only one or two machines runs to complete the experiment; hence NGS technologies are now competitive with the microarray platform for genome analysis. The routine use of microarray-based approaches are limited by the requirement for customized arrays, and these notable technical obstacles led to the transition of core genomic studies to high-throughput sequencing-based platforms [7, 8]. High-throughput sequences are generated from fragmented and adapter-ligated DNA/RNA/amplicon 'libraries' that have never been subjected to conventional vector-based cloning. As such, some of the sequencing bias of cloned DNA sequences that affect genome identification in sequencing projects may be avoided.

Sequencing technologies have a standard workflow regardless of the sequencing platforms, in brief, (1) preparation of sequencing library from the nucleic acid, (2) sequencing and data collection, and (3) data analysis (Fig. 1.1).

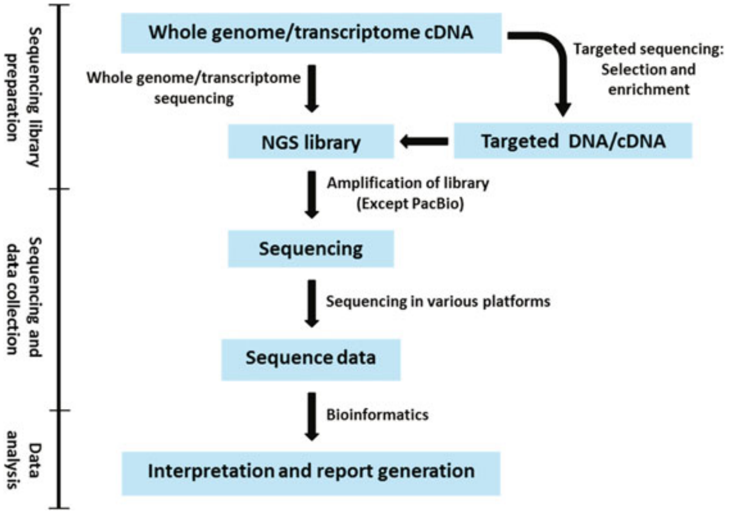

Sequencing library preparation
Sequencing and data collection
Data analysis
Whole genome/transcriptome cDNA
Whole genome/transcriptome sequencing
Targeted sequencing: Selection and enrichment
NGS library
Targeted DNA/cDNA
Amplification of library (Except PacBio)
Sequencing
Sequencing in various platforms
Sequence data
Bioinformatics
Interpretation and report generation

Fig. 1.1 Schematic representation of general working pipeline for next-generation sequencing. This working protocol is typically used in all kind of sequencing approaches

The exclusive reagents used in specific protocols differentiate one technology from another and define the type of data generated from each platform. All these protocols come under three major categories based on the sequencing chemistry (1) sequencing by synthesis, (2) single-molecule sequencing and (3) sequencing by ligation [9].

1.1.1 Sequencing by Synthesis (SBS)

SBS technology is similar to the Sanger sequencing method, it defines the nucleotide arrangement in the template by various signal detection method that is generated during the addition of a new base to the newly synthesized complementary DNA strand by the DNA polymerase. In Sanger sequencing, dideoxynucleotides were used in the chain termination reaction, whereas, in SBS, chemical/fluorescence detection of nucleotide addition is determined by different approaches using altered chemistry. Leading commercial platforms are clustered under the SBS methodology but they differ in sequence read length and template preparation.

1.1.1.1 Illumina

Currently, the Illumina platform is taking over the NGS market. It uses clonal or bridge amplification for template preparation and SBS technologies with cyclic reversible termination (CRT) during sequencing of the DNA template, including nucleotide addition by DNA polymerase, fluorescence detection, and cleavage of the extension termination site (Fig. 1.2a) [10–13]. 3′-Blocking terminators such as 3′-O-azidomethyl-dNTPs (Fig. 1.2b) [14] have been used effectively in CRT. Removal of two chemical bonds from the terminator leads to the detachment of the fluorophore from the nucleotide base to reinstate the 3′-OH group for the next cycle of sequencing. The template-fixing, primer-attached slide is divided into eight channels, meaning that several individual samples can be run simultaneously. Here, substitution of nucleotides is the likely error type.

1.1.1.2 Roche 454

The Roche 454 sequencing platform utilizes a type of SBS method called pyrosequencing to determine the nucleotide incorporation during sequencing [15]. In pyrosequencing (Fig. 1.3), the sequencing adapter-ligated DNA templates adhere to a microbead and are multiplexed by emulsion PCR. Each amplified bead is incubated with DNA polymerase, adenosine 5′-phosphosulfate (ASP), luciferase, ATP sulfurylase, and apyrase in a picotiter plate well. Pyrophosphate is released when the DNA polymerase adds the correct dNTP to the growing strand, which will

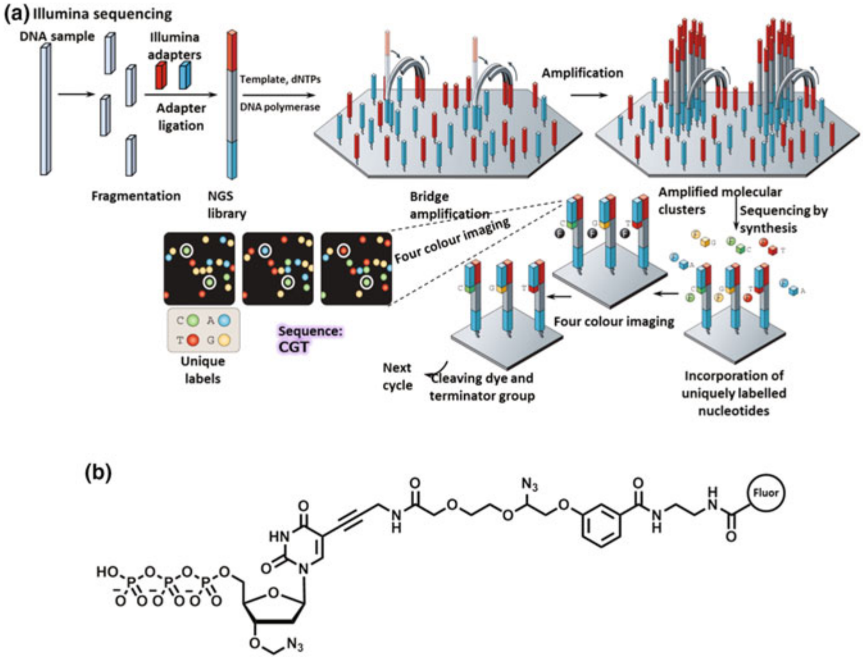


Fig. 1.2 **a** Illumina sequencing: Bridge amplification followed by sequencing by synthesis [10–13] sequencing approaches. **b** Modified nucleotides used in Illumina [10–13]

be converted into ATP by ATP sulfurylase in the presence of ASP. Luciferase reacts with ATP and releases light, which can be measured by an imaging system [16, 17]. Unused dNTPs are washed out with the apyrase. Repetition of the above reaction can define the details of the targeted DNA sequence. As the pyrosequencing does not have any termination moiety chemistry, multiple bases can be incorporated during a single sequencing cycle, which in turn can lead to erroneous homopolymer production. Further developments have been made to improve the sequencing performance and resolve the homopolymer issue [18].

1.1.1.3 Ion Torrent

Ion semiconductor sequencing [19] uses SBS. This technology is both rapid and cost-effective. In contrast to the other platforms, Ion torrent decode the template DNA sequence by detecting the pH changes that occur with the release of hydrogen ion upon the incorporation of nucleotides to the new DNA strand. Template-attached beads are incubated in a micro well with DNA polymerase and a particular type of dNTP. If the incubated dNTP matches the growing template strand, the DNA polymerase adds it. The incorporation of dNTP leads to the discharge of a hydrogen ion that activates an ion-sensitive field-effect transistor (ISFET) ion

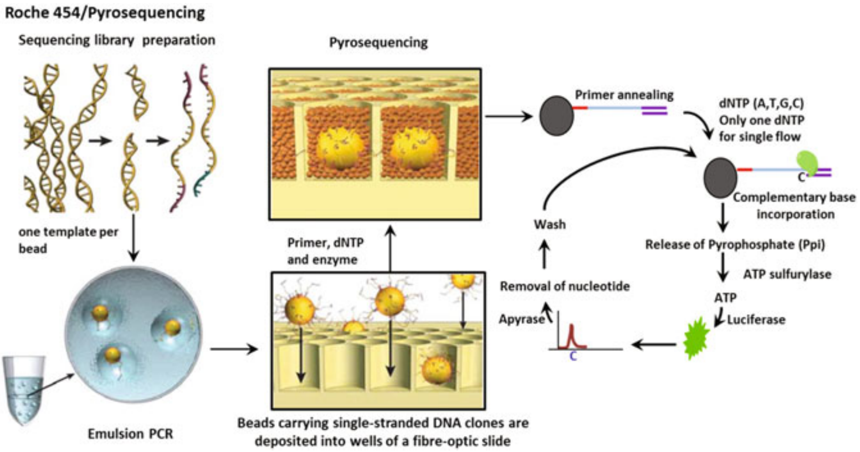

Roche 454/Pyrosequencing
Sequencing library preparation
one template per bead
Emulsion PCR
Pyrosequencing
Primer, dNTP and enzyme
Beads carrying single-stranded DNA clones are deposited into wells of a fibre-optic slide
Primer annealing
dNTP (A,T,G,C) Only one dNTP for single flow
C
Complementary base incorporation
Release of Pyrophosphate (Ppi)
ATP sulfurylase
ATP
Luciferase
Wash
Removal of nucleotide
Apyrase
C

Fig. 1.3 Roche 454 Sequencing: Emulsion PCR amplification followed by pyro sequencing [15]

sensor, which can detect the nucleotide by converting the electric signal into base sequences (Fig. 1.4). The remaining nucleotides are washed out and the next cycle continues. If repeated bases are present in the template, multiple nucleotides will be added in a single cycle; in this case, a stronger electric signal will be detected based

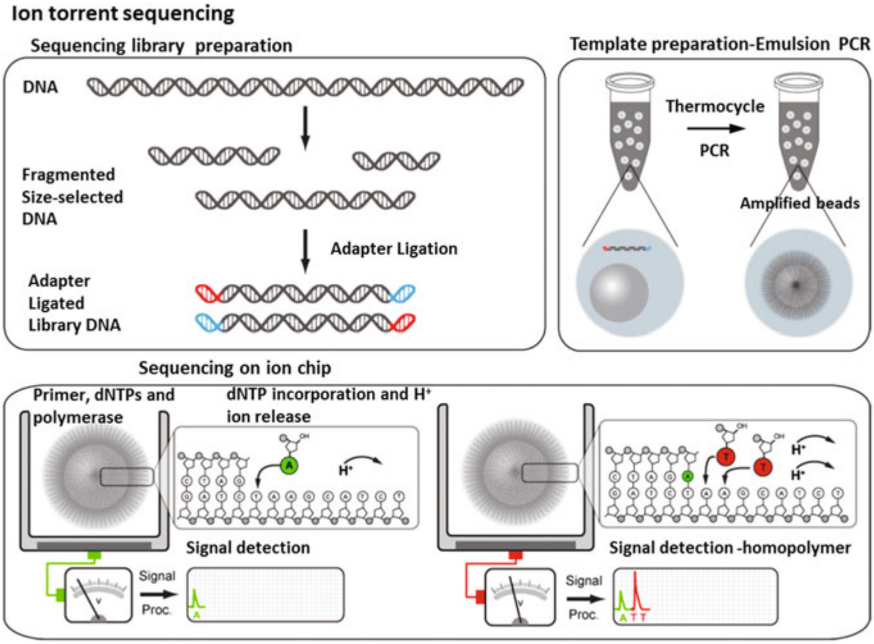

Ion torrent sequencing
Sequencing library preparation
DNA
Fragmented Size-selected DNA
Adapter Ligation
Adapter Ligated Library DNA
Template preparation-Emulsion PCR
Thermocycle
PCR
Amplified beads
Sequencing on ion chip
Primer, dNTPs and polymerase
dNTP incorporation and H⁺ ion release
A
H⁺
Signal detection
Signal
Proc.
A
T
T
H⁺
H⁺
Signal detection -homopolymer
Signal
Proc.
A T T

Fig. 1.4 Ion torrent: Same as B till the emulsion PCR next the enriched libraries sequenced by polymerization [19]

on the hydrogen ion release. Because of this multiple nucleotide addition, homopolymer error can occur, which is one disadvantage with this technology. Ion torrent has two types of sequencing systems: (1) Ion Personal Genome Machine (PGM) for small-scale usage, and (2) Ion Proton, which can generate higher throughput of sequencing data.

1.1.2 Single-Molecule Sequencing (SMS)

SMS uses a fluorescence emission detection method for the decoding of DNA. Platforms using this method can generate a measurable signal of fluorescence emission from a single nucleic acid by the addition of a fluorescently labeled nucleotide. Therefore, SMS does not require template amplification and obviates PCR errors. These methods can directly sequence RNA without cDNA amplification [20]. SMS platforms are differentiated based upon the method of immobilization of the template and the other molecules during the sequencing cycle, by the method of detection of emitted light.

1.1.2.1 Helicos

This was the first commercially available SMS system in the NGS market [21]. In this system, the fragmented DNA templates are denatured and attached with 3′-polyadenosine (A) and a terminal fluorescent A. These fragments are either hybridized to surface-attached primers in a flow cell or directly covalently attached to the flow-cell surface and annealed with a universal primer. Then, within the flow cell, fluorescently labeled nucleotides (virtual terminator) are sequentially added (a single dye system). Subsequently, nucleotide incorporation by DNA polymerase gives an image of the details of the template sequence. The cycle is repeated until an appropriate read length is reached; the terminator is removed at the end of every cycle.

1.1.2.2 Pacific Biosciences (PacBio)

Pacific BioSciences released a new single-molecule real-time sequencing (SMRT) technology (Fig. 1.5) [22, 23]. SMRT is based on the observation of DNA polymerization reactions in real time by capturing the light pulses produced during each nucleotide addition event. In this system, DNA polymerase is attached at the bottom of a Zero Mode Waveguide (ZMW) [24] with a single DNA template. By supplying uniquely fluorescently labeled nucleotides (A, T, G, and C) (Fig. 1.5b), the system can image the DNA polymerase-incorporated nucleotide fluorescence. The ZMW ensures the added nucleotides emit the strongest fluorescence. This ZMW-attached DNA polymerase can produce the longest read length in real-time mode [13]. The

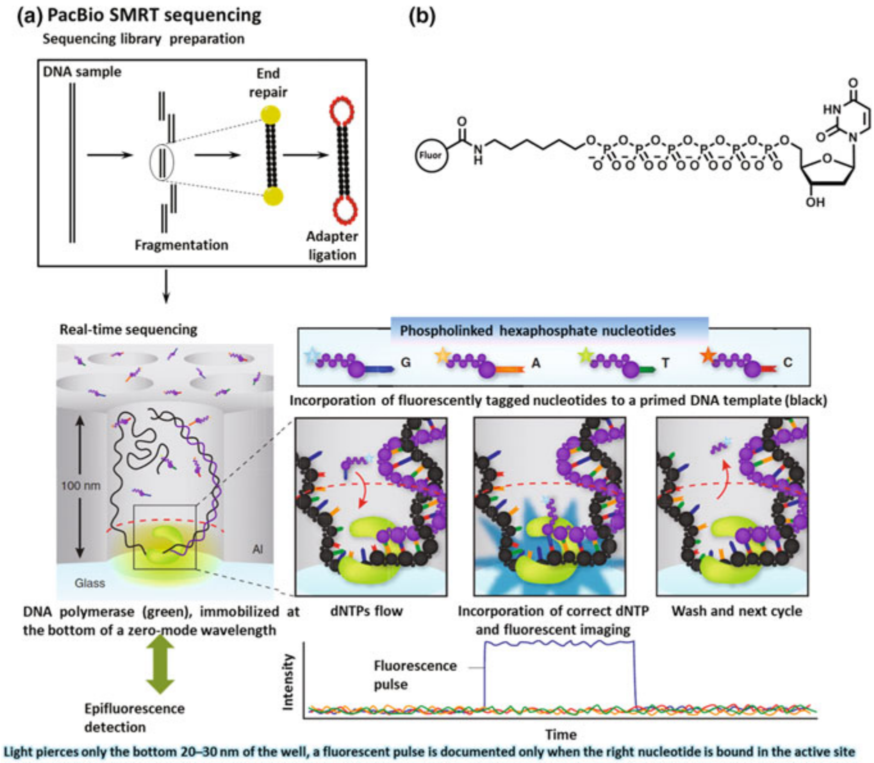

(a) PacBio SMRT sequencing
Sequencing library preparation
DNA sample
End repair
Fragmentation
Adapter ligation
Real-time sequencing
100 nm
Al
Glass
DNA polymerase (green), immobilized at the bottom of a zero-mode wavelength
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
Incorporation of fluorescently tagged nucleotides to a primed DNA template (black)
dNTPs flow
Incorporation of correct dNTP and fluorescent imaging
Wash and next cycle
Intensity
Fluorescence pulse
Time
Light pierces only the bottom 20–30 nm of the well, a fluorescent pulse is documented only when the right nucleotide is bound in the active site
(b)
Fluor
O
N
H
O
P
HN
N
OH

Fig. 1.5 **a** PacBIO: Single-molecule real-time [13, 22, 23] (SMRT) sequencing. **b** Terminally labelled polyphosphate used in PacBio sequencing [22], this type of nucleotides are added specifically to the template DNA, after imaging cleaved efficiently, and stretched as modified or natural nucleotide in next cycles

preparation of a circularized template can enable repeated sequencing of the template and an increased base accuracy.

1.1.3 Sequencing by Ligation (SBL)

Generally, DNA ligase enzyme ligation depicts the linking of two pairs of ends, however, it can also ligate the ends of one strand of the double-stranded DNA (while missing the terminal phosphate essential for ligation or the complementary strand is unbroken). This single strand ligation depends on the DNA ligase sensitivity towards the complementary bases of the two strands if there are mismatches between them the enzyme's ligation efficiency become very low. This mismatch sensitivity of the DNA ligase is being utilized in SBL methodology to define the sequences in the DNA molecule [25]. Various lengths of fluorescent tags labelled oligonucleotide probes were used here. The DNA sequencing library was already ligated with known adapter sequence, this can serve as an anchor sequence, where the primer can be annealed. Addition of DNA ligase to the flow cell can ligate the fluorescence tagged probes to primer with respect to the template sequence. Incorporation of a specific probe to the template can be identified using fluorescence imaging. By repeating this process with various groups of probes, it is possible to interrogate with the template DNA can evaluate the bases in the sequence. Platforms using this technology differ in their read length and the usage of probe.

1.1.3.1 Polonator and SOLiD (Support Oligonucleotide Ligation Detection)

SOLiD system [26] is based on the polonator technology [27]. Polonator is open source and so permits the researchers to advance in highly precise procedures and applications that do not depend on any kit. Both the systems use sequencing by ligation. In SOLiD platform, anchor sequence primed sequencing libraries were amplified on microbeads using emulsion PCR then the amplified beads were attached to a glass slide. After sequencing, primer annealing with the anchor sequence a set of unique fluorophore tagged probes is supplied to the flow cell. The probe contains various possible combinations of complementary bases, the fluorophore probes are partly the degenerated DNA octamers with the first two positions being complementary to the recognition core. DNA ligase can ligate the matching probe to the primer. After fluorescence imaging, 5′ phosphate groups were regenerated by cleaving the phosphorothiolate link with silver ions for the next ligation. A new cycle continues with ligation, detection, and cleavage. After appropriate read length is reached, the first sequencing product is peeled off and the second primer is allowed to anneal at $n - 1$ site to the DNA template. Various types of primer were utilized with the annealing site of n, $n - 1$, $n - 2$, $n - 3$, and $n - 4$. To improve the sequencing, precision DNA template is sequenced twice.

With so many applications and sequencing platforms available on the market, the general issue is how to identify the best available platform for a given chemical biological experiment. A comparison of each of the NGS technologies summarized in Table 1.1 facilitates the recognition by the chemical biologist of the ideal platform for their targeted research.

The major dominant commercial platforms currently on the market are the Illumina Genome Analyzer/HiSeq2500, the Roche 454 Genome Sequencer, the Life Technologies Ion Torrent Personal Genome Machine (PGM)/Ion proton, and the PacBio-SMRT.

Advancement in these new sequencing technologies and its impact on genomics is in turn causing an increase in chemical biological studies. However, significant methodological interpretations need to be explored to harness NGS in a better way in chemical biological studies. Here, we are discussing some of the key methodologies and analytical strategies of NGS applications in chemical biology.

Table 1.1 Comparison of Next Generation Sequencing technologies based on their data production and time duration for a run

Sequencing platform	Developers	Sequencing principle	Maximum data obtained per run (gb)	Maximum read length (bp)	Run time	Key features	Limitations
HiSeq 2500	Illumina	Sequencing by synthesis	120	100	27 h	Larger read number	Higher cost per read
PGM/Ion proton	Ion Torrent	Polymerization	~60	300	5–8 h	Simple detection method	Low reads number per run
Genome Sequencer FLX system	Roche	Pyro sequencing	0.5	400	8 h	Long read length	High cost per Mb
SOLiD	Life Technologies	Sequencing by ligation	50	50	7 days	Base-calling accuracy	High error rates and Low reads number per run
Pacific Biosciences	PacBio *RS/RS II*	Fragment/single molecule real time (SMRT)	13	4200–8500	–	Single molecule detection, long read length	High error rates and Low reads number per run
Polonator G.007	Dover	Ligation	12	26	5 days	Lower instrument cost, open source platform	Shortest NGS read lengths
GridION	Oxford Nanopore	Nanopore sensing	–	–	–	Longer read length, single molecule detection and label free	Not yet available; No data publicly available; 4% error-rates

1.2 Applications of NGS in Chemical Biology

In the interface of chemistry and biology, small molecules have advantages over larger molecules because they can be cell permeable, mostly nontoxic, cost-effective, and more easily synthesized, stored, and optimized. Furthermore, their capabilities to switch 'ON' and 'OFF' the function of specific genes or gene networks are easily alterable and can be precisely tuned. They have the capability of reprogramming somatic cells into pluripotent stem cells [28], but they have limitations, including the requirement for longer times and additional manpower for their selection and validation on a genomic scale.

The field of high-throughput sequencing and application development is a fast-moving area of genomic research. The NGS technologies have extended to an impressive array of applications beyond just genomic sequencing and its large-scale performance in chemical biological research (Fig. 1.6).

1.2.1 *Genome-Wide Localization of Non-B DNA Using Small Molecules*

Under common physiological conditions, the right-handed double-helical B-form of DNA is abundant. However, under specific conditions, DNA can also form a variety of alternate non-B DNA structures such as the four-stranded G-quadruplex,

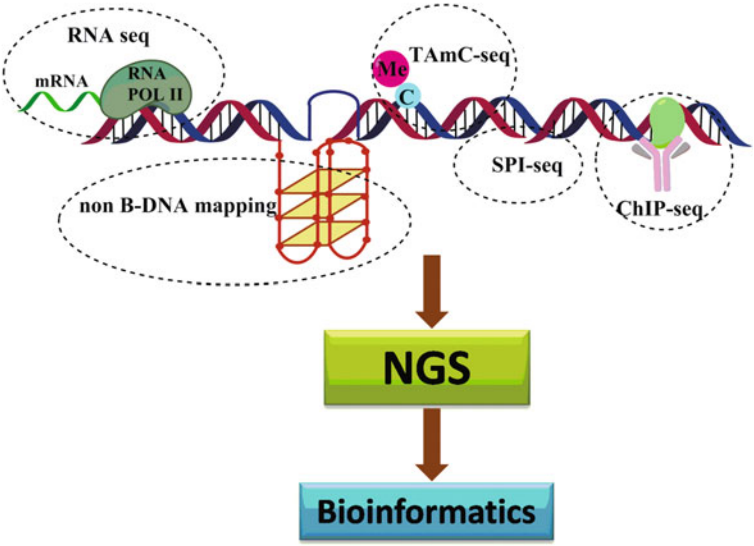
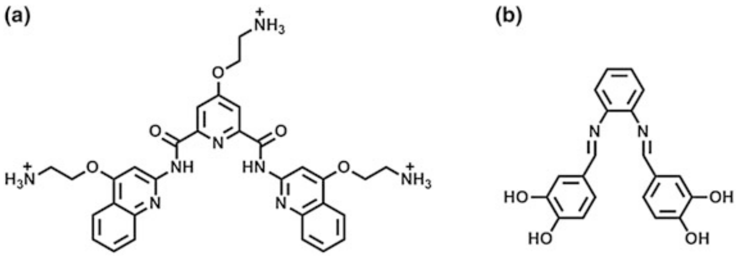


Fig. 1.6 Applications of Next Generation Sequencing in chemical biology

left-handed Z-DNA, cruciform, and others [29, 30]. The G-quadruplex, formed by Hoogsteen hydrogen bonds, is one of the most significant DNA structures, and is always formed in a G-rich region in the presence of some monovalent cations [31–33]. In the mammalian genome, the G-quadruplex is thought to be functionally significant for gene regulation, replication, and genome stability. Small-molecule ligands perturb cellular functions associated with the formation of this structure [34]. G-rich regions are randomly scattered in some sections of the mammalian genome, including telomeric ends and regulatory elements in some promoters, including c-myc and c-kit [35, 36]. G-quadruplex-associated small molecules like pyridostatin have been shown to localize the G-quadruplex in cells [37] and have been used to enrich human telomeric DNA (Scheme 1.1a) [38]. we have reported it in realtime observation [39]. Besides the telomeric G-rich sequence, other regions are also of significance. For example, the G-quadruplex in the promoter region of oncogenes has a close association with gene expression. G-quadruplex formation in these regions may play a vital role in gene regulation, hence this structure is considered an important therapeutic target. Lam et al. [40]. used a G-quadruplex-specific antibody to enrich genomic DNA fragments holding folded G-quadruplex structures and then the deep sequencing of the isolated DNA was performed.

They used a modified single chain hf2 antibody capable of enriching the stable G-quadruplex structures in the genomic DNA. Sequencing spots from independent NGS libraries were aligned to the human genome and peaks were called using the Model-based Analysis of ChIP-Seq (MACS) algorithm [42]. This study gave useful evidence about the presence of an exemplary set of G-quadruplex structures enriched by the hf2 antibody, which were mapped in the genome using deep sequencing. The identification and localization of stable G-quadruplexes in various gene regions of functional importance further strengthens the evidence for a potentially broad role of these structures.

Even though many small molecules are known to stabilize the G-quadruplex, most of them could not target the G-quadruplex in vivo because of noncovalent binding that result in weakened efficacy. Yuan et al. [41] identified the presence of G-quadruplex structures in an oncogenic promoter region using a G-quadruplex DNA cross-linking strategy. In this study, to illustrate the existence of the G-quadruplex in vivo, a new set of Schiff base catechol derivatives (Scheme 1.1b)



Scheme 1.1 Small molecules used to identify non B-DNA: **a** pyridostatin (PDS). **b** Schiff-base catechol derivatives [38, 41]

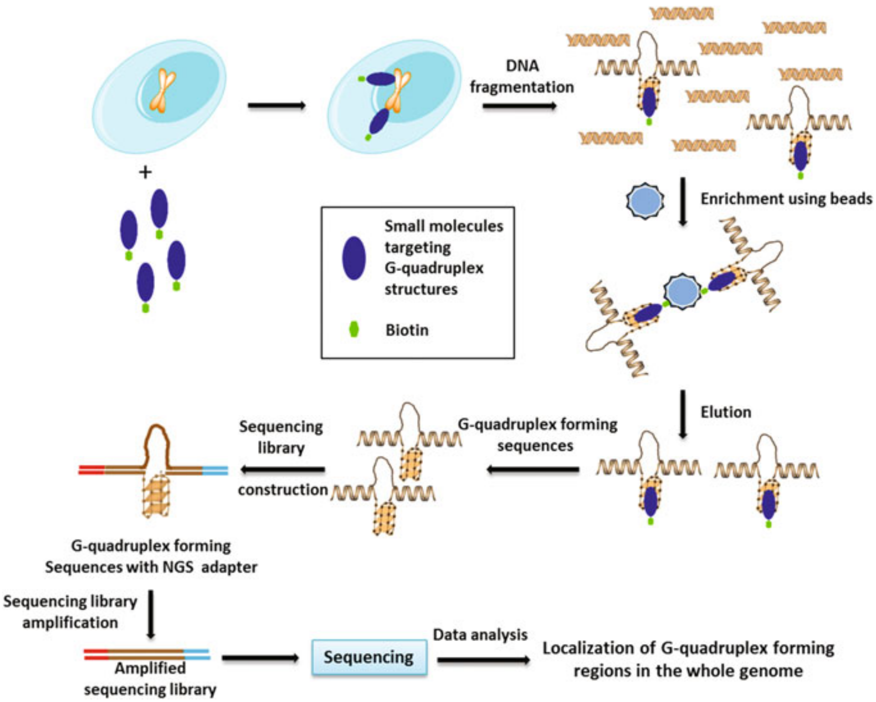

DNA fragmentation
Small molecules targeting G-quadruplex structures
Biotin
Enrichment using beads
Elution
G-quadruplex forming sequences
Sequencing library construction
G-quadruplex forming Sequences with NGS adapter
Sequencing library amplification
Amplified sequencing library
Sequencing
Data analysis
Localization of G-quadruplex forming regions in the whole genome

Fig. 1.7 Experimental procedure for detecting G-quadruplex forming region and sequencing [41]

were used as G-quadruplex cross-linking agents. The group then used a biotin tag for affinity purification of the targeted regions for further discovery [39, 43].

Figure 1.7 illustrates the work flow for extraction and sequencing of G-quadruplex forming regions. To further elucidate the exact sequence of these regions encountered by small molecules and their positions on the chromosome, deep sequencing was used effectively. As a result of high-throughput sequencing, these small molecules were identified as the first example of a G-quadruplex cross-linking agent that can efficiently target G-rich regions in the promoter of oncogenes in vivo. These deep-sequencing-associated methods may prove to be valuable new strategies for the rapid evaluation of the G-quadruplex on a genome-wide scale. They can also be useful in identifying G-quadruplex-mediated transcriptional regulation.

1.2.2 *Decoding of DNA Base Modification in Single Molecule Level*

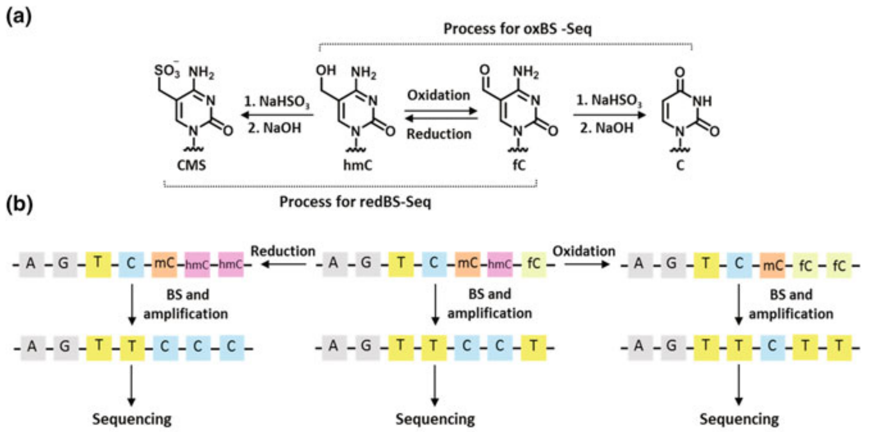
Cancer is known to be a disease process where somatic mutations drive the evolution of more virulent phenotypes. Nonetheless, high-throughput sequencing has unveiled a surprising degree of genetic alteration or base modification [5]. Several studies that compare diseased genomes with healthy ones uncovered tens of thousands of single or dinucleotide differences, epigenetic modifications, and hundreds of genomic rearrangements in the diseased genome. Personalized genome therapy with small-molecular chemistry is a promising future approach to open up drug development for genome-based diseases like cancer. Advancement in NGS technologies has forced small-molecule drug developers grapple with the problem of patient/tumor selection and personalized therapies. Deep sequencing can improve the quality of small-molecule chemical research from guiding the design of small molecules to genome-scale measurement of efficacy.

1.2.2.1 Studies Targets Epigenetically Modified Bases

In mammals, genomic DNA 5-methylcytosine (5mC) plays a vital role in variety of biological process through epigenetic gene regulation. It is an epigenetic modification caused by the action of DNMTs. In the progression of several diseases like cancer, the CpG islands of gene promoters become abnormally hypermethylated, which leads to transcriptional silencing that can be transferred to daughter cells following cell division. In general, hypomethylation occurs earlier in the disease process and is associated with chromosomal instability and loss of imprinting, but hypermethylation related to promoters can silence the gene (oncogene suppressor), so it could be a target for epigenetic therapy with small molecules. Information about the DNA methylation patterns and distribution in the human genome is undoubtedly important for developing small-molecule therapeutics.

Commonly, three established approaches are used to analyze genome-wide DNA methylation patterns in eukaryotic cells [44]. The first strategy involves restriction-enzyme-based approaches, using restriction enzymes that are not able to digest the recognition sequence at the site of DNA methylation, so 5mC can be identified in selected sequences. However, these methods are limited to the specific restriction sites in the genome. In the second strategy, fragmented DNA containing 5mC is captured using an affinity- based capture with 5mC-binding proteins (MBD-Seq) and antibody-based approaches (methylated DNA immunoprecipitation or MeDIP-Seq). Thirdly, in bisulfite sequencing (BS-Seq), denatured DNA is subjected to bisulfite treatment during which the normal cytosine is converted to uracil, but a methylated cytosine remains unchanged, thus permitting base-resolution detection of cytosine methylation. All of these methods have their limitations when utilized on a genomic scale. The major constraint is that these strategies cannot distinguish 5mC from 5-hydroxymethylcytosine (5hmC) [45–48]. 5hmC was discovered in 2009, as another relatively abundant form of cytosine modification [49, 50]. It may be an intermediate in active DNA demethylation, but it can also identify an epigenetic mark [51].

In 2012, a new strategy, “oxidative bisulfite” sequencing (oxBS-Seq), (Fig. 1.8) was developed, producing only Cs at 5mC sites, which in turn allows the clarification of the amount of 5hmC at a particular nucleotide position by comparing these



(c)

Bases	Sequence	BS-Seq	oxBS-Seq	redBS-Seq
A	A	A	A	A
G	G	G	G	G
T	T	T	T	T
C	C	T	T	T
mC	C	C	C	C
hmC	C	C	T	C
fC	C	T	T	C

Fig. 1.8 Oxidative bisulfite sequencing (oxBS-Seq) and reduced bisulfite sequencing (redBS-Seq) for the sequencing of 5hmC and 5fC, respectively. **a** Oxidative and reduced bisulfite reaction scheme. **b** Experimental principle of the BS-Seq, oxBS-Seq, and redBS-Seq. **c** Differential identification of 5mC, 5hmC, and 5fC from C by comparing BS-Seq, oxBS-Seq, and redBS-Seq data [45, 52]

data with BS-Seq data. In this method, dsDNAs containing C, 5mC, or 5hmC are oxidized with KRuO4 and then subjected to BS-Seq. 5hmC in the genomic DNA of mouse embryonic stem (ES) cells was mapped at high resolution using this method; it can also reliably map 5mC. Because of the fundamental mechanism of this method, it can be compatible with any sequencing platform [45].

In addition to oxBS-Seq, a new strategy named TET-assisted bisulfite sequencing (TAB-Seq) was developed, based on the principle that 5hmC can be oxidized to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by TET proteins [53–55]. In this technique, β-glucosyltransferase (β-GT) was used to attach glucose to 5hmC to protect 5hmC from further TET oxidation. After 5hmC protection, all 5mC is converted to 5caC by oxidation with Tet1 protein. Bisulfite treatment of the oxidized DNA then converts all C and 5caC (resulting from 5mC) to uracil or 5caU; however, the original 5hmC bases remain secured as 5gmC. Additional sequencing reveals 5hmC as C, and combined analysis of these data with traditional BS-Seq results delivers an accurate calculation of the modification at each cytosine [56].

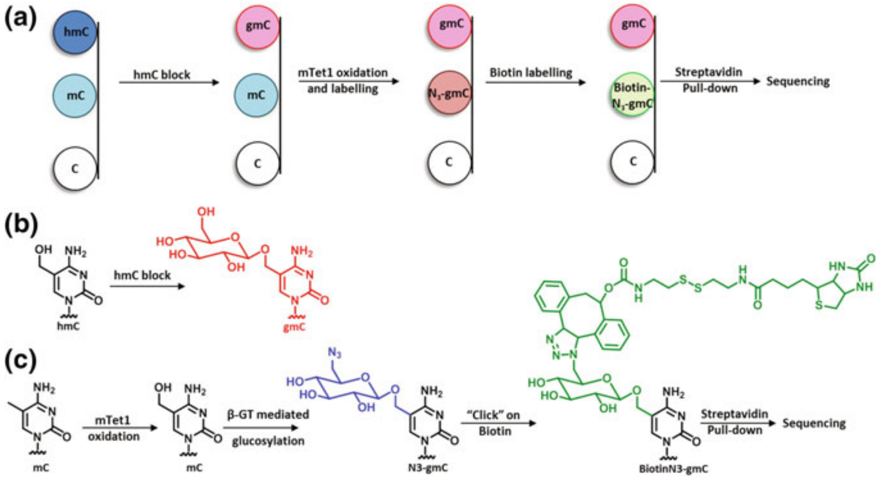


Fig. 1.9 Selective labelling of 5mC in DNA and sequencing using a one-pot mTet1/b-GT protocol. **a** Work flow of TAmC-Seq. **b** The internal 5hmC in gDNA can be secured by glycosylation with regular glucose. **c** The 5mC can be altered to 5hmC by mTet1-catalysed oxidation, and then 6-N3-b-glucosyl-5-hydroxymethyl-cytosine (N3-5gmC) conversion with 6-N3-Glucose (modified glucose moiety) via b-GT-mediated glucosylation, which could be further labelled with biotin moiety using click chemistry subsequent detection, affinity purification and sequencing [57]

Most of the approaches for 5mC analysis still have the restrictions of being density-biased, deficient in robustness and consistency, or incapable of analyzing 5mC specifically. The chemically inert methyl group inhibits direct tagging for subsequent affinity purification and detection. Zhang et al. developed a new approach, TET-assisted 5mC sequencing (TAmC-Seq), in which 5mCs are selectively labeled with an azide functionality that can be further tagged with biotin for affinity purification (Fig. 1.9). In brief, first, 5hmC is protected with glucose, then the mouse Tet1 is allowed to oxidize 5mC to 5hmC. These newly generated 5hmCs are trapped by β-GT-mediated transfer of a modified glucose moiety (6-N3-glucose) to generate 6-N3-b-glucosyl-5-hydroxymethyl-cytosine (N3-5gmC). Using click chemistry, a biotin tag is then inserted via the azide group of N3-5gmC for selective pull-down of the original 5mC and for subsequent sequencing [57]. As a result of these 5hmC sequencing strategies, the important role of the oxidized form of 5mC in epigenetic gene regulation has been recognized, even though it is present in small amount. To detect the oxidation dynamics of 5mC in DNA methylation, it is vital to identify the distribution of 5fC or 5caC at a genome level. C. He et al. developed two significant protocols using NGS for the peculiarity of 5fC in genomic DNA: [58] (1) the 5fC-selective chemical labeling (fC-Seal) method for genome-wide profiling of 5fC, and (2) the 5fC chemically assisted bisulfite sequencing (fCAB-seq) technique for the base-resolution detection of 5fC. Using both of these methods, genome-wide profiling of 5fC identified significant properties of

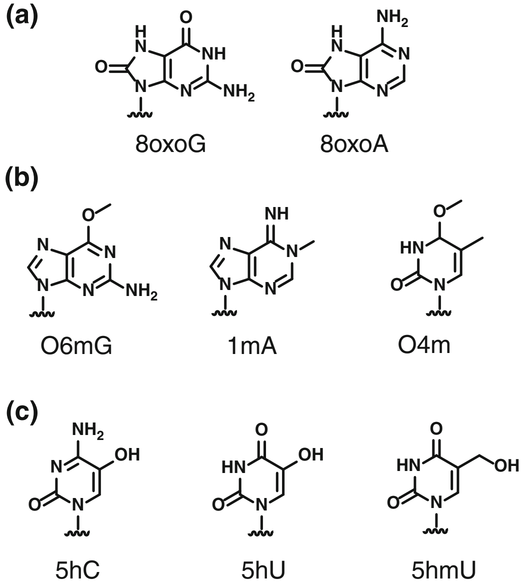
5mC/5hmC oxidation of several gene regulatory elements in the genome. A distinct indication of DNA demethylation is transformation of 5hmC into 5fC [59]. To identify the detail of these modifications, it is necessary to distinguish them more accurately at the level of single-base resolution. After the oxBS-Seq invention, the same group developed a quantitative method called reduced bisulfite sequencing (redBS-Seq) (Fig. 1.8) [52] to detect 5fC in the genome. Here, they used the principle that bisulfite treatment causes deformylation of 5fC, which then deaminates to U to be read as T by sequencing [45] and 5hmC is modified to a cytosine-5-methylsulfonate (CMS) adduct to be detected as C [48].

Using a precise reductant, sodium borohydride, 5fC was reduced to 5hmC in DNA prior to bisulfite treatment, these reduced 5fC bases being read as C (CMS). By comparing redBS-Seq (5fC detected as C) and BS-Seq (5fC detected as T), the group quantitatively detected 5fC at the level of single-base resolution. In mouse ES cells, using a combination of BS-Seq, oxBS-Seq, and redBS-Seq, they made the first quantitative single-base-resolution map of 5mC, 5hmC, and 5fC. These high-throughput sequencing-associated technology developments can offer a robust and reliable tool for the effective enrichment and epigenetic profiling of modified DNA bases.

1.2.2.2 DNA Strand Breakage and Damaged Bases

DNA is under constant stress from both exogenous and normal metabolic factors in the cell. Bases in the DNA can show limited chemical stability and are susceptible to chemical alterations through various types of damage, including alkylation, oxidation, radiation, and hydrolysis [60–62]. DNA damage happens at a frequency of 1000–1,000,000 molecular cuts per cell per day [63]. The majority of DNA damage disturbs the structure of the Watson and Crick double helix; that is, the bases are themselves chemically modified (8-oxoguanine, 8-oxoadenine, 1-methyladenine, 6-O-methylguanine, pyrimidine dimers, 5-hydroxycytosine, 5-hydroxyuracil, 5-hydroxymethyluracil, and thymine glycol). Derivatives resulting from different forms of DNA damage have been associated with progression of diverse but significant biological conditions including cancer, aging, and neurodegenerative diseases. In the eukaryotic genome, some sites are prone to breakage under stress, so the genome faces challenges to DNA stability. Consequently, developing approaches for examining damaged DNA in the framework of sequencing has gained increasing attention. Clark et al. [64] established a method using single-molecule real-time (SMRT) DNA sequencing to directly recognize damaged/modified DNA bases in the DNA template. They investigated alterations in the kinetics of DNA polymerase (stretches of fluorescent signals represent the dynamics of DNA polymerization) during the occurrence of the modified bases (Scheme 1.2).

Commonly used methods to detect DNA damage such as PCR assays, electrochemistry, radioactive labeling, immunochemical methods, mass spectrometry, comet assays, and chromatographic techniques are inadequate for DNA strand-breakage mapping on a genome-wide scale [65] as they cannot identify new hotspots



Scheme 1.2 SMRT sequencing and direct identification of Damaged DNA Bases: Modified bases identified using SMART sequencing based on the Kinetic change of polymerase. **a** Products of oxidative Damage. **b** Products of alkylation DNA damage. **c** Products of ionizing radiation DNA damage [64]

or breakage sites. Currently, various high-throughput sequencing platforms directly deliver a means to measure these multiple processes through massively parallel sequencing of DNA molecules from damaged DNA. Leduc et al. [66] developed an up-front strategy called 'damaged DNA immunoprecipitation' or 'dDIP.' This method combines the immunoprecipitation of biotin-modified nucleotides added by the terminal deoxynucleotidyl transferase (TdT)-mediated dUTP-biotin end-labeling (TUNEL) at sites of DNA damage. Immunoprecipitated DNA from the dDIP can be used in microarray analysis (ChIP-chip) or next-generation sequencing (ChIP-seq). Because of its greater resolution and lower costs, ChIP-seq is replacing ChIP-chip and is evolving as the preferred method to locate DNA-binding proteins. To map genomic hotspots of ssDNA damage, a strategy was developed using ssDNA-binding protein immunoprecipitation followed by sequencing (SPI-Seq) [67]. SPI-Seq was evaluated using Rad52, which is capable of binding to ssDNA formed at DNA lesions. In yeast, Rad52 is important for DNA strand-breakage repair and homologous recombination [68]. Rad52 is recruited to ssDNA exposed by resection during DNA replication. Therefore, mapping Rad52-associated DNA-binding sites is expected to be an alternate method for mapping ssDNA damage in yeast. It can be easily implemented with other proteins such as the DNA repair proteins in human cells (RAD51, RAD52, FANCD2, and BRCA2) and the checkpoint signaling proteins (ATR and ATRIP) that accumulate on ssDNA during DNA damage [69]. DNA damage can degrade important information in the genome. Double-strand breaks (DSB), in which both strands are damaged, are particularly hazardous to the cell. The mechanisms of DSB sensing and repair are well known, although the methods for genome-scale mapping of DSBs in various cells still lack resolution. ChIP-on-chip has been used to map DSBs [70–72]. DSB can be identified indirectly using antibodies to particular DSB-bound proteins. However, this raises a significant source of bias: for example, the commonly used DSB marker, phosphorylated histone variant H2A.X (γH2A.X), can also recognize ssDNA breakage [73–75]. Crosetto et al. [76] developed an experimental and computational methodology to directly map DSBs genome wide, which uses direct in situ DSB labeling, avidin enrichment, and deep sequencing (BLESS).

This labeling avoids the risk of false positives because BLESS cannot tag DSBs that are artificially formed during genomic DNA isolation, and T4 ligase enzyme is used for the ligation, which can ligate dsDNA but not ssDNA breaks. This strategy demonstrated a false positive DSB tagging proportion of less than 1% and is very effective for recognition of DSBs at various genomic sites. DSBs are the foremost drivers of chromosomal translocations. However, analysis of the global effect of genome organization on translocations needs a broad representation of a three-dimensional (3D) genome. Zhang et al. [77] carried out high-throughput genome-wide translocation sequencing using a high-resolution Hi-C spatial organization map generated from the G1-arrested mouse pro-B cell genome and mapped the translocations from target DNA double-strand breaks (DSBs) within it. The power of Hi-C facilitates us to assess the influence on translocation of DSB location inside the 3D genome and to identify the translocations formed with an induced DSB in these cells.

High-throughput sequencing methods open up the landscape (‘breakome’) of DNA breakage throughout the genome and are a simple, rapid, and cost-effective methodology applicable on a genome-wide scale.

1.2.3 Aptamer Selection Using Massively Parallel Sequencing

Aptamers are nucleic-acid-based oligomers that can be chemically synthesized and modified to target molecules with highly selective affinity binding [78–81]. They have been synthesized against a variety of molecular targets including proteins, small molecules, and cell-surface markers [82–87]. Recently, aptamers have found uses in a wide range of applications including diagnostics, molecular imaging, therapeutics, gene delivery, and drug delivery [88–94].

The commonly used systematic evolution of ligands by exponential enrichment (SELEX) method has been used to prepare aptamers targeting proteins and small molecules [78, 95, 96]. It takes months to assess and optimize just a handful of aptamers. NGS can greatly speed up this process. Studies [97, 98] utilizing deep sequencing to select aptamers shortened the time required for initial aptamer selection. Both the studies cited were able to obtain aptamers that bound to the proposed target, but they either were dependent on many rounds of selection or restricted the flexibility available in the sequence space. Hoon et al. [99] developed

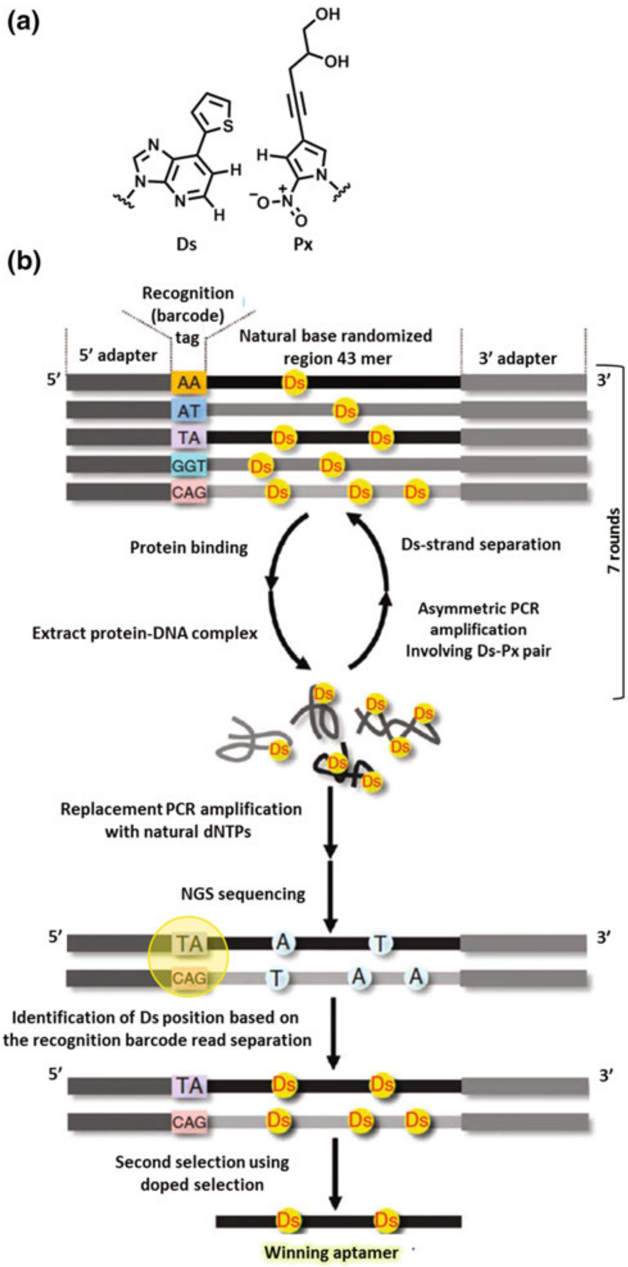

(a)
Ds
Px
(b)
Recognition (barcode) tag
5′ adapter
Natural base randomized region 43 mer
3′ adapter
5′
3′
AA
AT
TA
GGT
CAG
Ds
Protein binding
Ds-strand separation
Extract protein-DNA complex
Asymmetric PCR amplification Involving Ds-Px pair
7 rounds
Replacement PCR amplification with natural dNTPs
NGS sequencing
Identification of Ds position based on the recognition barcode read separation
Second selection using doped selection
Winning aptamer

Fig. 1.10 High-affinity DNA aptamers selection using an additional genetic alphabet. **a** Structure of an unnatural nucleotide with the hydrophobic base 7-(2-thienyl) imidazo [4,5-b] pyridine (Ds) and its base pairing diol-modified 2-nitro-4-propynylpyrrole (Px). **b** Design of oligomer library used in the aptamer selection and sequening scheme of the SELEX procedure: DNA aptamer selection using libraries with five different bases (A, T, G, C and Ds). Recognition tag: barcode sequences used to sequence multiple samples in a single run, Doped selection: aptamer selection as above explained using partially randomized oligo libraries [102]

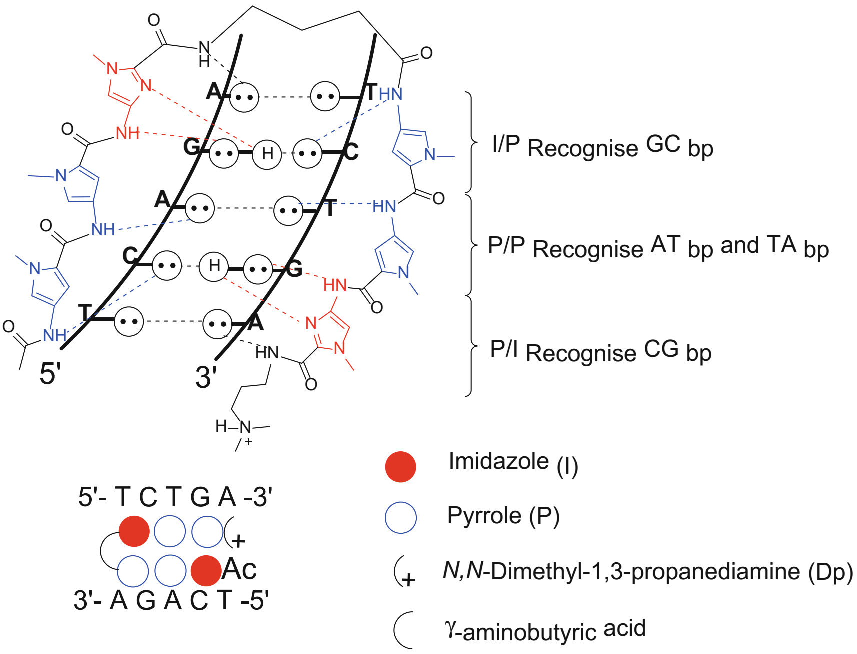
novel aptamers against thrombin using a different method, aptamer selection by K-mer analysis of sequences (ASKAS). This needs only one round of positive selection followed by deep sequencing and data analysis [mainly using cluster-seq (http://code.google.com/p/biopieces)].

In some cases, modified natural nucleotides have been incorporated into aptamers to increase their efficiency [100, 101] In contrast to modified natural nucleotides, there are ‘unnatural’ nucleotides, which pair with each other but not with any of the four (A, T, G, and C) natural nucleotides. These unnatural bases could improve the utility of aptamers by providing added chemical and structural diversity.

Kimoto et al. [102] used an unnatural nucleotide (Fig. 1.10) with a hydrophobic base 7-(2-thienyl) imidazo[4,5-b]pyridine (Ds) to select aptamers for two target proteins, vascular endothelial cell growth factor-165 (VEGF-165) and interferon-γ (IFN-γ). The Ds-base-incorporated aptamers were selected through SELEX followed by high-throughput sequencing to select the optimal aptamers. With the hydrophobic base 7-(2-thienyl) imidazo[4,5-b]pyridine (Ds) to select aptamers for two target proteins, vascular endothelial cell growth factor-165 (VEGF-165) and interferon-γ (IFN-γ). The Ds-base-incorporated aptamers were selected through SELEX followed by high-throughput sequencing to select the optimal aptamers.

1.3 Development of *N*-Methylpyrrole (P)—*N*-Methylimidazole (I) Polyamides (PIP) and Its Conjugates

Small molecules are organic compounds with well-defined chemical structure less than 900 Da in size. Typically, they must exhibit pharmacologically active properties such as easy absorption and metabolism in the human body with little or no toxicity. Macro molecules are large molecules consisting mainly of proteins and carbohydrates of larger molecular weight. Small molecules are relatively stable inside cells compared with large molecules, and they usually do not elicit an immune response [103]. DNA-binding small molecules have gained importance because of their possible application in cancer chemotherapy. Small molecules bind DNA through either intercalation, covalent interactions, or by interacting with the DNA groove [104]. PIPs, derivatives of the naturally occurring anti-cancer agent Distamycin A, are a well-known group of small molecules capable of binding in the DNA minor groove with notable cell permeability and stability [105]. Most of the naturally available chemical products such as actinomycin, echinomycin, daunomycin, and chromomycin pocess complex structure. Dicker son et al. revealed the X-ray crystal structure of 1:1 netropsin binding to DNA, the crescent pyrrole-pyrrole (P-P) bound to consecutive A,T in the minor groove DNA. Further, crystallographic analysis identified that NHs of the carboxamides keen toward the minor groove of the helical structure forming hydrogen bonds with the A-T and



Scheme 1.3 Hairpin PIP recognition of the DNA minor groove

T-A base pairs [106]. The results of selective AT rich sequences recognition by Pyrrole (P) conjugates leads to the synthesis of synthetic analogue of netropsin, imidazole (I) containing 1-methylimidazole-2-carboxamide netropsin [107]. It was revealed Imidazole (I) could selectively bind to G [108]. The antiparallel arrangement of P/I distinguishes CG from GC base pairs [109]. Therefore, an antiparallel arrangement of I/P recognizes a GC base pair but P/P cannot differentiate AT base pair from TA base pair, so it recognizes AT and TA base pairs (Scheme 1.3) [110–112].

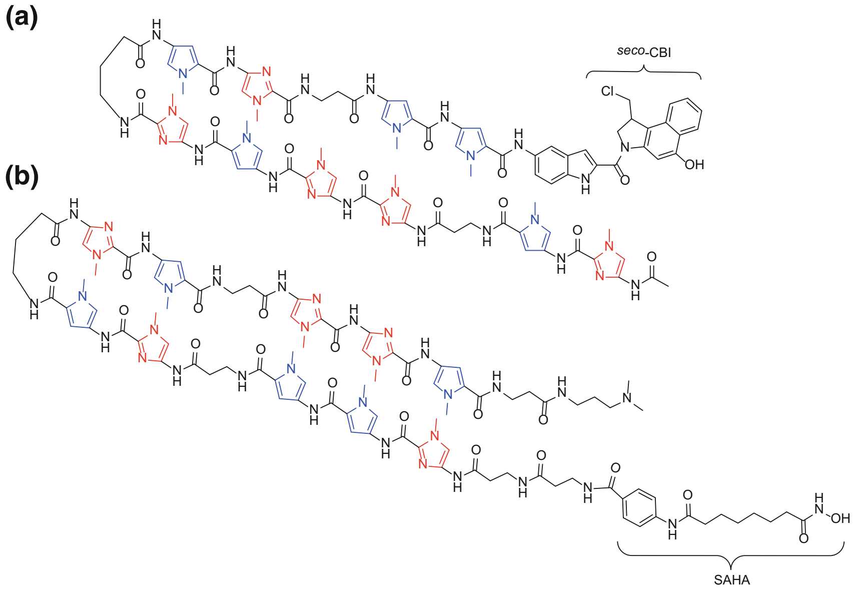
Among the reported PIPs, the most extensively studied type is hairpin PIP, formed by covalently linking two antiparallel PIP ligands, which preferably bind in a forward amino-carboxyl (N-C) orientation with regard to the 5′-3′ direction of the double strand DNA [113]. This hairpin PIP binding to DNA is as good as the natural transcription factor proteins binding [101]. Later development in PIPs designing showed incorporation of flexible aliphatic β-alanine (β) in the place of pyrrole (P) can provide efficient binding, with P/β, β/P or a β/β pairings can recognize W (W = A/T) and β/I recognizing a CG base pairs [114–116]. The hairpin PIPs having efficient cell membrane penetration [117], and it can be directly conjugate to DNA alkylating agents for specific gene silencing [118], histone deacetylase (HDAC) inhibitors such as SAHA [119], and transcription activating domains [120] for controlled gene regulation.

1.4 PIP Based Gene Regulation

PIPs are widely used in gene silencing and gene activation. There are two type of PIP design used to inhibit gene expression. (1) Hairpin PIPs can induce gene silencing by designing PIPs binding to the transcription factor binding site at the promoter region or directly to the promoter region of target genes. In prostate cancer, PIP was used to inhibit the expression of androgen receptor (AR) regulating genes in mouse xenografts. This PIP was designed for androgen response element (ARE) to target number of AR-regulating genes [121]. We have recently reported the inhibition of human ectopic viral integration site 1 (EVI1), an oncogenic transcription factor; over expression observed in metastatic breast cancer cells. Our design of PIP to the transcription factors binding site caused the inhibition of REL/ELK1 transcription factor binding in EVI1 promoter, resulting in the repression of EVI1 gene and its downstream gene targets [122]. (2) Usually some genes are regulated by multiple transcription factors (TF), these TFs can also control multiple genes. In such a case, designing a PIP to unique sequences in the coding region of a gene can selectively silence a single gene. Though PIPs bound in the coding region, it can be easily removed by RNA polymerase during transcription. To overcome this complication, an alkylating moiety such as *seco*-CBI (1,2,9,9a-tetrahydrocyclopropa[1,2-*c*]benz[1,2-*e*]indol-4-one) can be conjugated to the PIPs, which form a covalent can adduct with a purine (mostly adenine) at the N3 position. These alkylating PIPs can selectively inhibit gene expression [123, 124], and target cancer-associated mutation sequences [125]. A PIP-indole-*seco*-CBI conjugate KR12 (Scheme 1.4a), with unique sequence recognition was tested for effective antitumor activity by selective silencing of codon 12 mutated KRAS gene, KR12 alkylation at specific adenine causes DNA strand cleavage and growth inhibition in human colon cancer cells with G12D or G12V mutations, it leads to the senescence and apoptosis [126].

Gene activation can be achieved by two important types of PIP conjugates (1) Attaching transcription activation domain to the PIP by a linker domain. Activation domains are proteins that can recruits transcription complex on the promoter region and induce corresponding gene expression. Mapp and co-workers developed a synthetic transcriptional activator by conjugating hairpin PIP with alpha helical (AH) peptide activation domain through 36-atom chain linker [114]. Later studies with this conjugate showed an improved transcriptional activation by reducing the linker atom length from 36 to 8. Further improvement was achieved by replacing 20-mer AH peptide with 16-mer peptide derived from viral activator VP16 [128]. Uesugi et al., developed a synthetic transcription factor with two functional domains: DNA binding hairpin PIP and non-peptidic wrenchnolol that can bind to the subunit of human mediator complex Sur-2. The developed PIP-wrenchnolol conjugate showed target gene activation [129].

(2) Attaching PIP with chromatin modifier which could activate epigenetically silenced genes. In this regards, our group reported a novel class of synthetic transcriptional activator with PIP and histone deacetylase (HDAC) inhibitor SAHA



Scheme 1.4 Chemical structure of synthetic gene regulators **a** PIP-indole-*seco*-CBI conjugate KR12 [126]. **b** SAHA-PIP designed to target p16 promoter region [127]

(suberoyl anilide hydroxamic acid). The first SAHA-PIP conjugate was designed to target the promoter region of the p16 tumor suppressor gene (Scheme 1.4b). The designed SAHA-PIP conjugate selectively induced Histone acetylation (H3K9) around p16 promoter and showed morphological changes in HeLa cells [127]. Later we reported SAHA-PIP K which could trigger *PIWI* pathway genes (specific to germ cell activation) in fibroblast cells [130]. Recently we developed a library containing 32 SAHA-PIPs, global gene expression screening of this library in human dermal fibroblast cells showed that each SAHA-PIP conjugates can activate unique set of genes (Fig. 1.11) [119]. Follow-up studies on this library revealed that SAHA-PIP I can activate pluripotency inducing genes [131] and SAHA-PIP X can trigger retinal tissue related genes in human dermal fibroblasts [132].

1.5 Utilization of NGS in PIP Based Small Molecule Studies

1.5.1 Guiding the Design and Screening of Small Molecule Using NGS

PIPs are modifiable synthetic oligomers that can bind to the DNA minor groove based on the DNA recognition rules [133]. This group of small molecules as shown

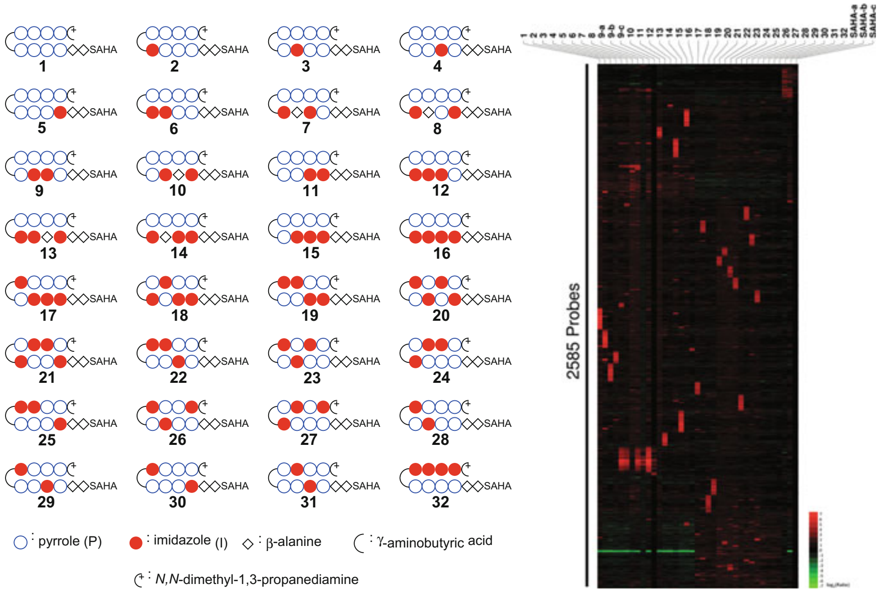


Fig. 1.11 Chemical structure representation of SAHA-PIP library and its corresponding global gene expression hierarchical clustering analysis proposes that each SAHA-PIP activate a unique cluster of genes in human fibroblast cells. For SAHA-PIP 9 (I) and SAHA, biological triplicates data is shown [119]

earlier, it can act as an efficient synthetic gene regulator in either activation [134] or inhibition [135] of a gene/gene network. Defining the binding site of DNA-binding small molecules on a whole-genome-sequence scale could be useful in accomplishing the challenging task of targeting particular regions of dsDNA. In this regard, the biological applications of PIPs could be enhanced with the knowledge of their sequence specificity in a large sequence framework. Meier et al. [136] studied the sequence selectivity and canonical pairing rules of PIP DNA binding in a broad sequence context, using affinity purification coupled with massively parallel sequencing (Fig. 1.12a, b). The study followed the methodologies of Bind-n-seq [137] a high-throughput method for analyzing protein–DNA interactions in vitro. The major steps in this process commence with the synthesis of biotinylated PIP and 21mer randomized oligonucleotide with sequence-specific adapter sequences on both sides. Each PIP-biotin conjugate is allowed to equilibrate with the 21mer randomized region and the bound and unbound sequences are separated via affinity purification. Next, PIP-enriched sequences are subjected to high-throughput sequencing, and finally motifs among the sequences are identified with the motif-finding program DREME [138]. The technique permits fast, quantitative identification of the PIP-binding sites and their direction. This method correlates well with restriction endonuclease protection, selection, and amplification (REPSA) [139] and microarray-based binding site identification [140–142]. This unbiased

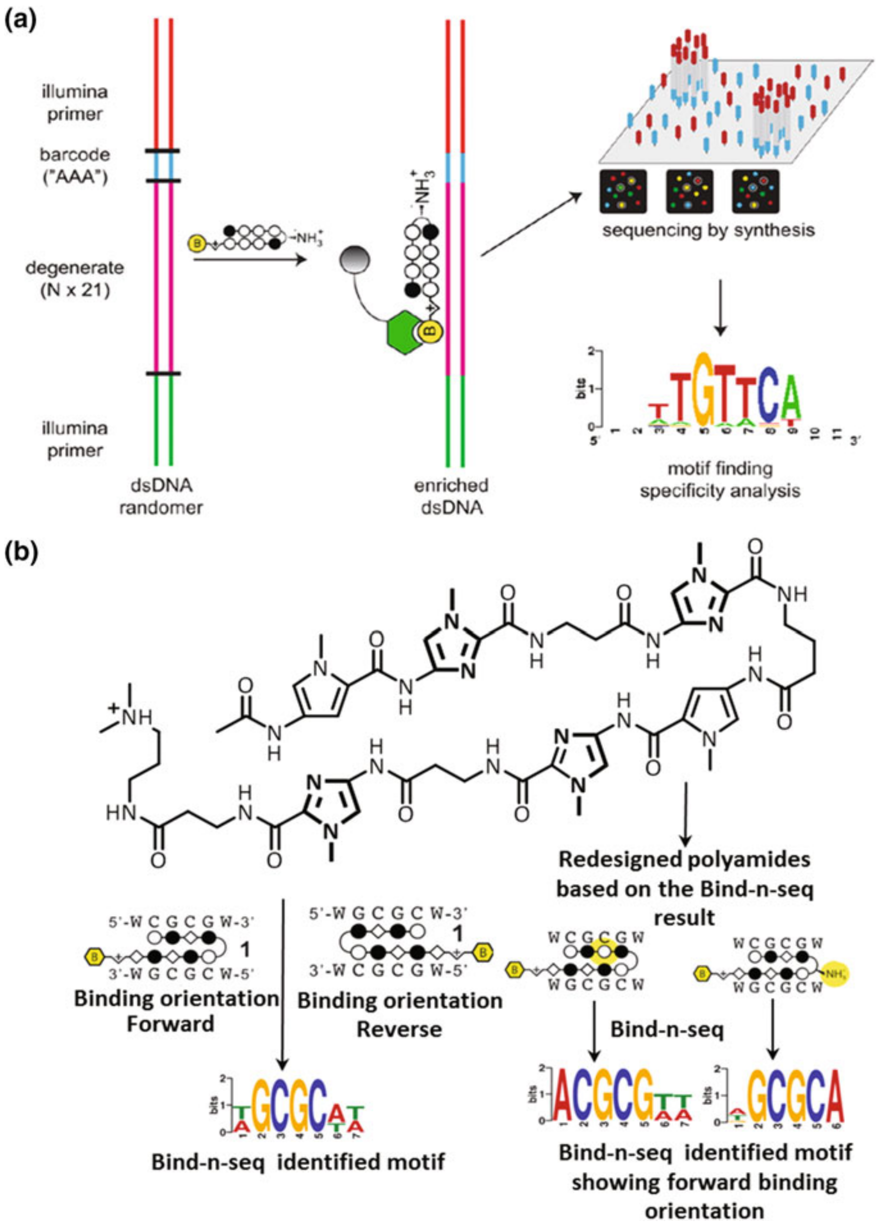

(a)
illumina primer
barcode ("AAA")
degenerate (N x 21)
illumina primer
dsDNA randomer
enriched dsDNA
sequencing by synthesis
motif finding specificity analysis
bits
5'
3'
(b)
Redesigned polyamides based on the Bind-n-seq result
5'-W C G C G W-3'
3'-W G C G C W-5'
1
Binding orientation Forward
5'-W G C G C W-3'
3'-W C G C G W-5'
Binding orientation Reverse
W C G C G W
W G C G C W
Bind-n-seq
Bind-n-seq identified motif
Bind-n-seq identified motif showing forward binding orientation

Fig. 1.12 High-throughput sequencing guided designing and validation of DNA binding small molecule. **a** Workflow for Bind-n-Seq analysis of PIPs recognition motif identification. **b** Methylated 5-CGCG-3 targeted PIP structure (PIP could potentially bind in the forward orientation or the reverse orientation), Bind-n-seq analysis of PIP showed the reverse orientation binding 5′-GCGC-3 and possible modification in the PIP (Positions are highlighted in yellow) and Bind-n-seq analysis of redesigned polyamide showing forward orientation of binding 5′-CGCG-3′ [136, 143]

technique revealed unanticipated binding/orientation sites of the PIPs, which was useful in improving the study of sequence-selective inhibition of CpG methylation [143]. In this study, PIPs targeting the sequence 5′-CGCG-3′ [144] were used. Bind-n-seq analysis could discern structure–activity relationships and could guide the authors in designing improved CpG methylation antagonists (Fig. 1.12b). Recent studies substantiate the potential of PIP conjugates in gene regulation [119, 145]. The biological activity of these gene-regulating PIP conjugates could be enhanced with improved sequence recognition. Such complex feats could be achieved with the development of strategies like Bind-n-seq. Thus, NGS offers a greater scope in designing next-generation DNA-binding small molecules. To understood about the binding principle of PIPs in actual chromatin packed DNA, "Crosslinking of small molecules for isolation of chromatin" (COSMIC) was a study to determine the PIP high-affinity binding in nucleus by conjugating photo-crosslinker with PIP. COSMIC-qPCR was further extended for the binding conformation, but no NGS was carried out [146]. Thus, the ability to genome-wide mapping of the PIP-DNA-binding sites throughout the chromatin packed genome utilizing NGS could provide a more detailed understanding of the mechanism of gene regulation by PIP conjugates.

1.5.2 Analysis of Gene Expression Induced by PIP Based Small Molecules

1.5.2.1 Transcriptome Studies

Transcriptome analysis gives an account of the complete spectrum of mRNAs in a cell and their magnitude of expression for a specific physiological condition or type of cell [39]. RNA high-throughput sequencing (RNA-Seq) is a modern approach to transcriptome profiling that uses high-throughput sequencing technologies [147–149]. It can analyze the expressed sequences in a spatiotemporal manner and is rapidly replacing other methods of profiling gene expression such as microarrays. Microarray expression studies have been effective in interpreting the expression of mRNAs within cells and tissues; however, there are a number of limitations to this technology, including low sensitivity and specificity. More importantly, microarray constrains the expression-profiling data to specific annotations and content. Gene expression studies using RNA-Seq offer the possibility of reducing and/or in some cases eliminating these drawbacks. Once a transcriptome has been sequenced, we can use the data to evaluate gene regulation. This type of study mainly focuses on: (1) listing all the transcripts with respect to cell type for the species, including all varieties of RNAs (mRNAs, noncoding RNAs, and small RNAs); (2) resolution of the transcriptional organization of genes, based on their transcription start sites, pattern of splicing, and other posttranscriptional modifications; and (3) enumeration of the differential expression levels of each transcript

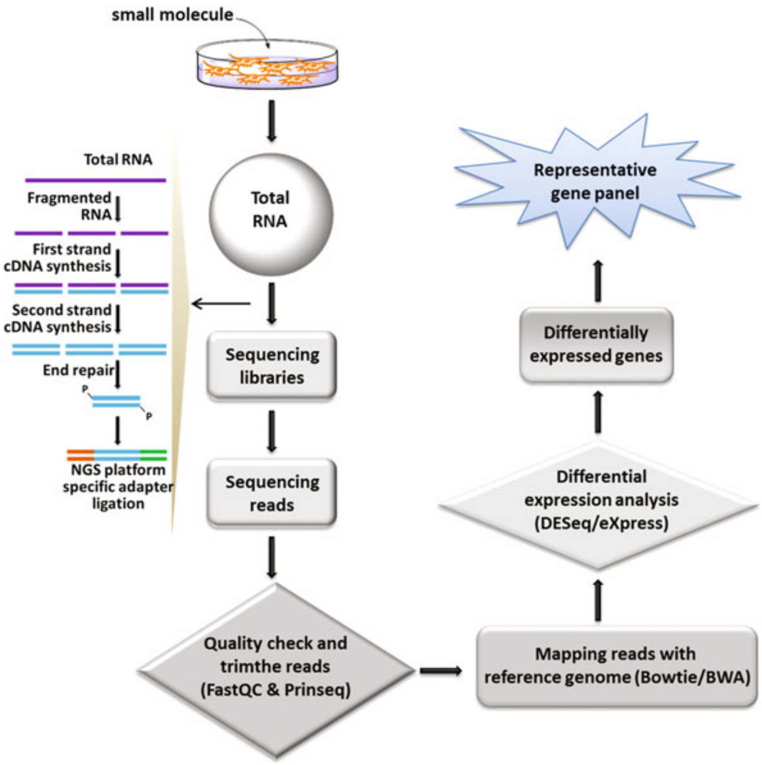

small molecule
Total RNA
Total RNA
Fragmented RNA
First strand cDNA synthesis
Second strand cDNA synthesis
End repair
P
P
NGS platform specific adapter ligation
Sequencing libraries
Sequencing reads
Quality check and trimthe reads (FastQC & Prinseq)
Mapping reads with reference genome (Bowtie/BWA)
Differential expression analysis (DESeq/eXpress)
Differentially expressed genes
Representative gene panel

Fig. 1.13 Schematic representation of the pipeline for RNA-Seq analysis of small molecule-regulated RNA

(Fig. 1.13). With the availability of faster and more cost-effective NGS platforms, ample transcriptome analyses can be performed to check the effect of gene regulation by DNA-binding small molecules.

Transcriptional regulation by DNA-binding small molecules could have important therapeutic uses. Successive studies have shown that pyrrole–imidazole polyamides (PIPs) can repress DNA binding by transcription factors such as the androgen receptor (AR) [121], hypoxia inducible factor 1 alpha (HIF-1α) [150], the glucocorticoid receptor (GR) [151], and nuclear factor kappa B (NF-κB) [152] in live cells. RNA-Seq of small-molecule-targeted cells, tissues, or animal models has allowed the identification of further alterations in gene expression. Raskatov et al. [153] investigated the effect of a PIP synthesized to bind with the DNA sequence 5′-WGGWWW-3′ (W = A or T) in a xenograft tumor model. The study primarily focused on the evaluation of the effect of PIPs on gene expression in vivo. To quantify the global effect on gene expression of PIP in a xenograft environment, RNA from PIP-treated and untreated mice was measured using RNA-Seq. A panel of representative genes was selected from the list of differentially expressed genes and confirmed using reverse transcription–quantitative polymerase chain reaction (RT-qPCR). Differentially expressed genes including *CCL2*, *NPTX1*, *SERPINE1*, and *MMP28* were identified. A similar PIP with a different recognition site (5′-WGGWCW-3′) was used to study the global transcriptome expression changes in breast cancer cells using RNA-Seq [154]. These transcriptome studies demonstrate the crucial importance of deep-sequencing strategies for the rapid validation of small-molecule potency.

1.5.2.2 Protein-DNA Interaction Studies

Recent progress in high-throughput sequencing technology has facilitated the identification of DNA binding protein's target sites in genome scale. A combined chromatin immune precipitation and high-throughput sequencing (ChIP-Seq) method has been used extensively to determine the DNA-binding patterns of DNA-binding proteins and the epigenetic modification marks on chromatin [155–158]. Theoretically, this technology can distinctively recognize in an unbiased manner various sections of DNA in the genome that are physically associated with a specific DNA-binding protein. This permits clear mapping of the interactions between particular proteins and their transcriptional targets to suggest interconnections of gene regulatory networks. Furey [159] reviews current studies using the transcription-factor-binding ChIP together with high-throughput sequencing and its full downstream analysis pipeline. Figure 1.14 represents the complete workflow for ChIP-Seq. The studies on DNA-interacting proteins most frequently target transcription factors (e.g., p53 or NFκB), chromatin-modifying enzymes (e.g., DNA methyltransferases (DNMTs), histone deacetylases), modified histones interacting with genomic DNA (e.g., histone 3 trimethylated on lysine 4), and the basal transcriptional machinery apparatus (Example RNA polymerase II). They can govern when genes are switched on or off/transcribed. Some of the DNA-interacting proteins can act as repressors and some as activators. Furthermore, a single protein sometimes directly controls multiple downstream genes, resulting in the highly diverse gene regulatory networks that control numerous biological processes. Pyridostatin is a highly selective G-quadruplex-associated small molecule [38, 160] known to inhibit the growth of human cancer cells by inducing replication/transcription dependent on DNA damage.

Rodriguez et al. [161] used ChIP-Seq methods to analyze genome-wide pyridostatin-induced DNA damage with the DNA damage-marker protein γH2AX. By comparing ChIP-Seq data from pyridostatin-treated and untreated control cells, it was identified that in the human genome, each distinct chromosome includes ~60 γH2AX domains that are induced by pyridostatin (example enrichment were shown in Fig. 1.15a. This indicates the impact of pyridostatin on gene expression including that of the *SRC* gene (proto-oncogene) (Fig. 1.15b). The study showed that the small molecule reduced the level of SRC protein and its dependent cellular activity in human breast cancer cells. As a result, a previously unknown

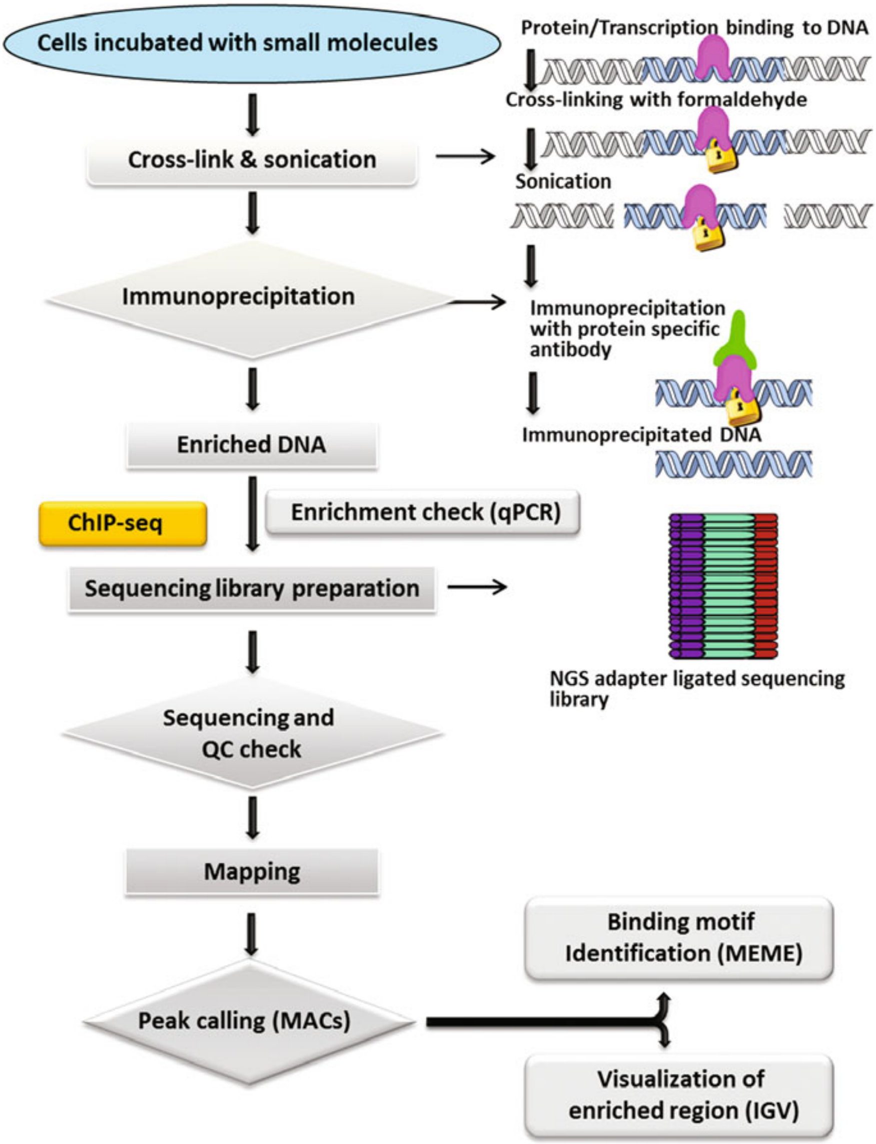

Cells incubated with small molecules
Cross-link & sonication
Immunoprecipitation
Enriched DNA
ChIP-seq
Enrichment check (qPCR)
Sequencing library preparation
Sequencing and QC check
Mapping
Peak calling (MACs)
Binding motif Identification (MEME)
Visualization of enriched region (IGV)
Protein/Transcription binding to DNA
Cross-linking with formaldehyde
Sonication
Immunoprecipitation with protein specific antibody
Immunoprecipitated DNA
NGS adapter ligated sequencing library

Fig. 1.14 Complete work flow of ChIP-Seq and its downstream analysis pipeline

small-molecule pyridostatin-binding region in the genome was identified that may lead to drug discovery for identifiable genomic targets.

PIP–DNA binding causes allosteric changes in the DNA helix that can interfere with protein–DNA interactions [163, 164] Yang et al. [162] investigated the effect of PIPs targeted to the RNAP2 transcription machinery. ChIP-Seq was used to map

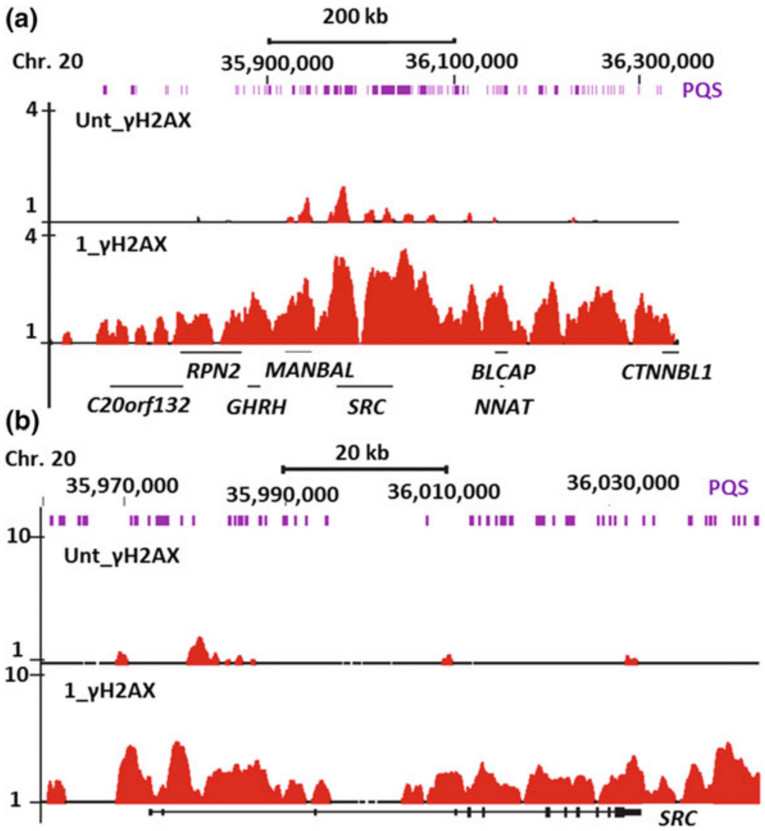


Fig. 1.15 Genome wide mapping of small molecule induced gene. **a** ChIP-Seq recognized γH2AX regions containing putative G-quadruplex–forming sequence (PQS) clusters in oncogenes and tumor suppressor genes. MRC MRC-5-SV40 cells treated with 2 µM pyridostatin (1_ γH2AX) compared to untreated MRC MRC-5-SV40 cells (Unt_ γH2AX). **b** A zoomed window size showing the enriched sequencing read distribution in the chromosomal region containing *SRC* gene [161]. **c** Structure of polyamides (PIP) and d) ChIP-Seq result of genome level RNAP2 occupancy from control(NT), DHT-treated (DHT), and DHT + PIP-treated (DHT + 1) samples over (A) an ARdriven gene, *KLK3* [162]. Axis details: x-axis represent the chromosome positions and transcripts are displayed, y-axis corresponds to sequencing read depth which represent the number of times an antibody bound to that specific sequence during the experiment; purple bars represent mapped PQS [161, 162]

the global occupancy of RNAP2 in LNCaP cells under dihydrotestosterone (DHT) induction. The results indicate that androgen receptor (AR)-driven genes such as KLK3 show increased RNAP2 binding to their DNA, but this was decreased in the presence of PIP (Fig. 1.15c, d). Although RNAP2 binding across constitutively expressed genes such as *GAPDH* did not change with DHT treatment, there was a reduction in binding after PIP treatment. This reduction in RNAP2 occupancy induced by PIP was in line with a global reduction of RNAP2

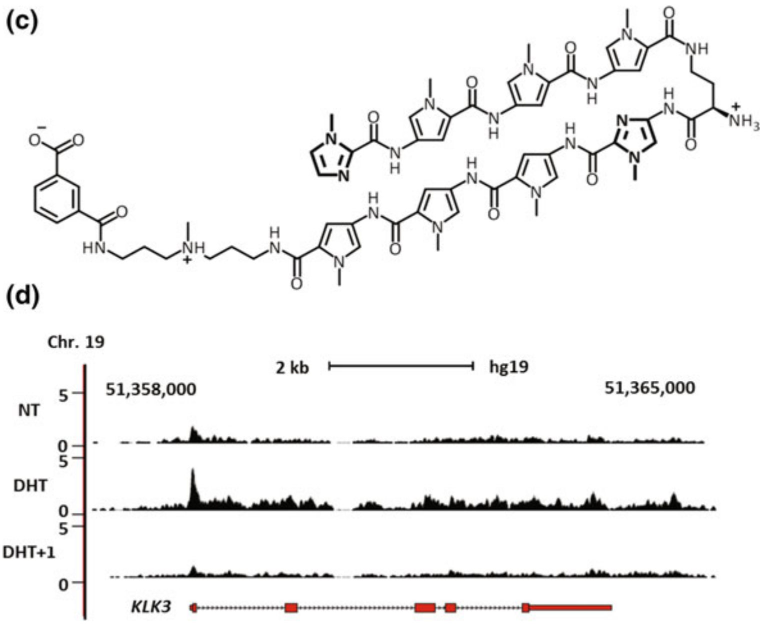


Fig. 1.15 (continued)

occupancy across genic regions. By using this high-throughput ChIP-Seq, Yang et al. derived a conclusive genome-wide mapping of RNAP2 binding showing reduced affinity to DNA preferentially at transcription start sites, while the occupancy at enhancer sites was unchanged. Treatment with PIP caused a time- and dose-dependent weakening of the binding of RNAP2 large subunit RPB1 that is avoidable with proteasome inhibition. Similarly, transcriptional activator PIPs [SAHA-PIP = HDAC inhibitor suberoylanilide hydroxamic acid (SAHA) + hairpin PIP] increased the level of PIWIL1 (associated with germ-cell development) in the H3Ac-occupied regions in a genome-wide PIP-induced epigenetic study [130] Overall, high-throughput sequencing delivers ideal tools to unravel many interactions that make up these gene regulatory networks. ChIP-Seq could be a promising strategy for drug discovery to identify the specific role of chemical compounds.

1.6 Conclusion and Future Prospects

As deep sequencing opens up a new era in small-molecule development, it can allow for rapid validation of small molecules with minimum hands-on work. As we have described, new innovations in sequencing applications and data processing are being developed regularly (Table 1.2) [165].

Table 1.2 List of studies used next-generation sequencing technologies beyond just genomic sequencing and its large scale operations in chemical biological research

S. N	Target study	Deep sequencing Methods	Small molecule/antibody used in the study	Platform used	Analysis pipeline	Data produced
1	Protein-DNA interaction	ChIP-Seq	Py-Im polyamide	Illumina GAIIx sequencer	MACS	25–30 million post filtered reads per library
2	hmC seq	1. oxBS-Seq 2. TAB-Seq	1. Potassium perruthenate 2. β-glucosyl transferase/tet1	1. Illumina GAIIx 2. Ion PGM/Illumina HiSeq2000	1. Bismark v0.6.4/comparison of oxBS-seq with BS-seq 2. MACS/comparison of TAB-seq with BS-seq	1. Single-end read with 40 bp sequecing 2. Paired end sequencing, total 509.8 M reads
3	mC seq	TAmC-Seq	N3-5gmC labelling with biotin via click chemistry	Illumina	Bowtie v0.17.2/comparison of TAmC-Seq data with MeDIP-Seq data and BS-seq	Single-end read with 51 bp sequencing
4	Transcriptional regulation by small molecules	RNA-Seq	Py-Im polyamide	Illumina GAIIx sequencer	Bowtie/eXpress 1.0.0 (bio.math.berkeley.edu/A/index.html)	25–30 million post filtered reads per library
5	Non-B-DNA detection	ChIP-Seq with G-quadruplex specific antibody/G-quadruplex DNA cross-linking strategy	G-quadruplex specific hf2 antibody	Illumina MiSeq	MACS/MEME	8 million reads were used
6	Aptamer selection	ASKAS	Thrombin-coated magnetic beads	Illumina GAIIx	K-mer analysis (Tallymer software)/cluster-seq (http://code.google.com/p/biopieces)	Single-end read with 36 cycle sequencing
7	DNA strand breakage	SPI-Seq	ssDNA binding Rad52	Illumina GA-II	SOAP2/kernel density estimation/IGV	1–1.5 M reads per library

Innovative sequencing technologies, such as SMS and nanostructure-based sequencing, hold great promise to achieve ever faster, cheaper, more accurate, and more reliable ways to design and produce advanced or next-generation small molecules for genome-level applications. This prospective application of high-throughput sequencing in chemical biology will be a great milestone in the field of PIP based synthetic gene regulators development. Generally DNA-binding small molecules design has been guided by advanced analytical screening methods. However, applications of PIPs and its various types of conjugates in biological aspects such as personalized medicine will need the understanding of sequence-specificity in actual genomic space. It is also important to develop large scale screening techniques that can investigate sequence-selectivity of PIP conjugates in a higher throughput with less time and source in an intensive manner than gel-based assays.

Scope of the research

As compiled in the general introduction, multitasking NGS technologies can accurately measure genomic data in different physiological conditions, and this accuracy could lead to value-added research by chemical biologists in innovative ways. All chemical biological research depending on DNA/RNA sequence data has been profoundly improved, driven by the powerful NGS tools. Various studies in the interface of chemistry and biology (small-molecule screening, artificial transcription controller development, and epigenetic modifier development) need genome-wide quantitative analysis, and the advancements in NGS could simplify this laborious process. It is also important to design PIPs in an established sequence-specific manner to precisely control gene regulation. To develop binding specificity established synthetic gene regulators, we have precisely studied various types of PIPs and its conjugates such as SAHA-PIPs and alkylating PIPs binding mechanism using our newly developed high-throughput sequencing based methods. We also studied the poorly understood phenomenon of PIP conjugate-DNA binding in the native chromatinized DNA containing nucleus from live cells. It revealed the actual binding mechanism of PIP conjugates and unknown high-affinity target sites in the actual chromatin context. These key findings and the newly developed high-throughput sequencing based methods will significantly contribute to advancement in development of synthetic gene regulators, focusing on the personalized therapies.

References

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467
2. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564. doi:10.1073/pnas.74.2.560
3. Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876. doi:10.1038/nature06884

4. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. doi:10.1038/nature03001
5. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696. doi:10.1038/nrg2841
6. Hui P (2012) Next generation sequencing: chemistry, technology and applications. In: TripleC. pp 1–18
7. Hurd PJ, Nelson CJ (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Briefings Funct Genomics Proteomics 8:174–183. doi:10.1093/bfgp/elp013
8. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680. doi:10.1038/nrg2641
9. Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99:175–185. doi:10.3732/ajb.1200020
10. Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. doi:10.1038/nature07517
11. Barnes C, Balasubramanian S, Liu X, Swerdlow H, Milton J (2006) Labelled nucleotides. US Patent 7,057,026, 6 June 2006
12. Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15:1767–1776. doi:10.1101/gr.3770505
13. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11 (1):31–46. doi:10.1038/nrg2626
14. Guo J, Xu N, Li Z et al (2008) Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A 105:9145–9150. doi:10.1073/pnas.0804023105
15. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. doi:10.1038/nature03959
16. Ronaghi M, Uhlén M, Nyrén P (1998) DNA SEQUENCING: a sequencing method based on real-time pyrophosphate. Science 281(5375):363–365. doi:10.1126/science.281.5375.363
17. Ahmadian A, Ehn M, Hober S (2006) Pyrosequencing: history, biochemistry and future. Clin Chim Acta 363:83–94
18. Smith AM, Heisler LE, St. Onge RP et al (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. doi:10.1093/nar/gkq368
19. Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352. doi:10.1038/nature10242
20. Ozsolak F, Platt AR, Jones DR et al (2009) Direct RNA sequencing. Nature 461:814–818. doi:10.1038/nature08390
21. Harris TD, Buzby PR, Babcock H et al (2008) Single-molecule DNA sequencing of a viral genome. SOM. Science 320:106–109. doi:10.1126/science.1150427
22. Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. doi:10.1126/science.1162986
23. Kielian M (2011) News & views research. Nature 1:8–9. doi:10.1038/487043a
24. Levene MJ, Korlach J, Turner SW et al (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686. doi:10.1126/science.1079700
25. Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. Science 241:1077–1080. doi:10.1126/science.3413476
26. Valouev A, Ichikawa J, Tonthat T et al (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18:1051–1063. doi:10.1101/gr.076463.108
27. Shendure J, Porreca GJ, Reppas NB et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732. doi:10.1126/science.1117389
28. Hou P, Li Y, Zhang X et al (2013) Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science 341:651–654. doi:10.1126/science.1239278

29. Mirkin SM (2008) Discovery of alternative DNA structures: a heroic decade (1979–1989). Front Biosci 13:1064–1071. doi:10.2741/2744
30. Ross PD, Shrake A (1988) Decrease in stability of human albumin with increase in protein concentration. J Biol Chem 263:11196–11202
31. Zahler AM, Williamson JR, Cech TR, Prescott DM (1991) Inhibition of telomerase by G-quartet DNA structures. Nature 350:718–720
32. Rodriguez R, Pantoş GD, Gonçalves DPN et al (2007) Ligand-driven G-quadruplex conformational switching by using an unusual mode of interaction. Angew Chem Int Ed 46:5405–5407. doi:10.1002/anie.200605075
33. Wang X, Huang J, Zhou Y et al (2010) Conformational switching of G-quadruplex DNA by photoregulation. Angew Chem Int Ed 49:5305–5309. doi:10.1002/anie.201002290
34. Balasubramanian S, Neidle S (2009) G-quadruplex nucleic acids as therapeutic targets. Curr Opin Chem Biol 13:345–353
35. Bugaut A, Jantos K, Wietor J-L et al (2008) Exploring the differential recognition of DNA G-quadruplex targets by small molecules using dynamic combinatorial chemistry. Angew Chem Int Ed 47:2677–2680
36. Balasubramanian S, Hurley LH, Neidle S (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? Nat Rev Drug Discov 10:261–275. doi:10.1038/nrd3428
37. Granotier C, Pennarun G, Riou L et al (2005) Preferential binding of a G-quadruplex ligand to human chromosome ends. Nucleic Acids Res 33:4182–4190. doi:10.1093/nar/gki722
38. Müller S, Kumari S, Rodriguez R, Balasubramanian S (2010) Small-molecule-mediated G-quadruplex isolation from human cells. Nat Chem 2:1095–1098. doi:10.1038/nchem.842
39. Rajendran A, Endo M, Hidaka K et al (2014) G-quadruplex-binding ligand-induced DNA synapsis inside a DNA origami frame. RSC Adv 4:6346. doi:10.1039/c3ra45676e
40. Lam EYN, Beraldi D, Tannahill D, Balasubramanian S (2013) G-quadruplex structures are stable and detectable in human genomic DNA. Nat Commun 4:1796. doi:10.1038/ncomms2792
41. Yuan L, Tian T, Chen Y et al (2013) Existence of G-quadruplex structures in promoter region of oncogenes confirmed by G-quadruplex DNA cross-linking strategy. Sci Rep 3:1811
42. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137. doi:10.1186/gb-2008-9-9-r137
43. Song C-X, Szulwach KE, Fu Y et al (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol 29:68–72. doi:10.1038/nbt.1732
44. Beck S, Rakyan VK (2008) The methylome: approaches for global DNA methylation profiling. Trends Genet 24:231–237. doi:10.1016/j.tig.2008.01.006
45. Booth MJ, Branco MR, Ficz G et al (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science 336:934–937. doi:10.1126/science.1220671
46. Jin SG, Kadam S, Pfeifer GP (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res. doi:10.1093/nar/gkq223
47. Lister R, O'Malley RC, Tonti-Filippini J et al (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536. doi:10.1016/j.cell.2008.03.029
48. Huang Y, Pastor WA, Shen Y et al (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS ONE. doi:10.1371/journal.pone.0008888
49. Tahiliani M, Koh KP, Shen Y et al (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science (80-) 324:930–935. doi:10.1126/science.1170116
50. Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science (80-) 324:929–930. doi:10.1126/science.1169786

51. Branco MR, Ficz G, Reik W (2012) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet 13:7–13. doi:10.1038/nrg3080
52. Booth MJ, Marsico G, Bachman M et al (2014) Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat Chem 6:435–440. doi:10.1038/nchem.1893
53. Ito S, Shen L, Dai Q et al (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science (80-) 333:1300–1303. doi:10.1126/science.1210597
54. He Y-F, Li B-Z, Li Z et al (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science (80-) 333:1303–1307. doi:10.1126/science.1210944
55. Pfaffeneder T, Hackner B, Truss M et al (2011) The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew Chem Int Ed 50:7008–7012
56. Yu M, Hon GC, Szulwach KE et al (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149:1368–1380. doi:10.1016/j.cell.2012.04.027
57. Zhang L, Szulwach KE, Hon GC et al (2013) Tet-mediated covalent labelling of 5-methylcytosine for its genome-wide detection and sequencing. Nat Commun 4:1517. doi:10.1038/ncomms2527
58. Song CX, Szulwach KE, Dai Q et al (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell 153:678–691. doi:10.1016/j.cell.2013.04.001
59. Spruijt CG, Gnerlich F, Smits AH et al (2013) Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives. Cell 152:1146–1159. doi:10.1016/j.cell.2013.02.004
60. Geacintov NE, Broyde S (2010) Chemical biology of DNA damage. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 3–126
61. Preston BD, Albertson TM, Herr AJ (2010) DNA replication fidelity and cancer. Semin Cancer Biol 20:281–293
62. Kelley MR (2012) DNA repair in cancer therapy. Academic Press, Elsevier Science, pp 1–16
63. Lodish H, Berk A, Matsudaira P et al (2004) Molecular biology of the cell, 5th ed. WH Freeman, New York, p 963
64. Clark TA, Spittle KE, Turner SW, Korlach J (2011) Direct detection and sequencing of damaged DNA bases. Genome Integr 2:10. doi:10.1186/2041-9414-2-10
65. Kumari S, Rastogi RP, Singh KL et al (2008) DNA damage detection strategies. EXCLI J, 44–62
66. Leduc F, Faucher D, Bikond Nkoma G et al (2011) Genome-wide mapping of DNA strand breaks. PLoS ONE 6:e17353. doi:10.1371/journal.pone.0017353
67. Zhou ZX, Zhang MJ, Peng X et al (2013) Mapping genomic hotspots of DNA damage by a single-strand-DNA-compatible and strand-specific ChIP-seq method. Genome Res 23:705–715. doi:10.1101/gr.146357.112
68. Mortensen UH, Lisby M, Rothstein R (2009) Rad52. Curr Biol 19:R676–R677. doi:10.1016/j.cub.2009.06.001
69. Bekker-Jensen S, Lukas C, Kitagawa R et al (2006) Spatial organization of the mammalian genome surveillance machinery in response to DNA strand breaks. J Cell Biol 173:195–206. doi:10.1083/jcb.200510130
70. Szilard RK, Jacques P-E, Laramée L et al (2010) Systematic identification of fragile sites via genome-wide location analysis of gamma-H2AX. Nat Struct Mol Biol 17:299–305. doi:10.1038/nsmb.1754
71. Seo J, Kim K, Chang DY et al (2014) Genome-wide reorganization of histone H2AX toward particular fragile sites on cell activation. Nucleic Acids Res 42:1016–1025. doi:10.1093/nar/gkt951
72. Harrigan JA, Belotserkovskaya R, Coates J et al (2011) Replication stress induces 53BP1-containing OPT domains in G1 cells. J Cell Biol 193:97–108. doi:10.1083/jcb.201011083

73. Marti TM, Hefner E, Feeney L et al (2006) H2AX phosphorylation within the G1 phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. Proc Natl Acad Sci U S A 103:9891–9896. doi:10.1073/pnas.0603779103
74. Chadwick BP, Lane TF (2005) BRCA1 associates with the inactive X chromosome in late S-phase, coupled with transient H2AX phosphorylation. Chromosoma 114:432–439. doi:10.1007/s00412-005-0029-1
75. Tuduri S, Crabbé L, Conti C et al (2009) Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat Cell Biol 11:1315–1324. doi:10.1038/ncb1984
76. Crosetto N, Mitra A, Silva MJ et al (2013) Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods 10:361–365. doi:10.1038/nmeth.2408
77. Zhang Y, McCord RP, Ho YJ et al (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148:908–921. doi:10.1016/j.cell.2012.02.002
78. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510. doi:10.1126/science.2200121
79. Mairal T, Cengiz Özalp V, Lozano Sánchez P et al (2008) Aptamers: molecular tools for analytical applications. Anal Bioanal Chem 390:989–1007. doi:10.1007/s00216-007-1346-4
80. Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346:818–822. doi:10.1038/346818a0
81. Jayasena SD (1999) Aptamers: an emerging class of molecules that rival antibodies in diagnostics. Clin Chem 45:1628–1650. doi:10.1038/mtna.2014.74
82. Lupold SE, Hicke BJ, Lin Y, Coffey DS (2002) Identification and characterization of nuclease-stabilized RNA molecules that bind human prostate cancer cells via the prostate-specific membrane antigen. Cancer Res 62:4029–4033. doi:10.1126/science.2200121
83. Huizenga DE, Szostak JW (1995) A DNA aptamer that binds adenosine and ATP. Biochemistry 34:656–665. doi:10.1021/bi00002a033
84. Bock LC, Griffin LC, Latham JA et al (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin. Nature 355:564–566. doi:10.1038/355564a0
85. Mallikaratchy P, Tang Z, Kwame S et al (2007) Aptamer directly evolved from live cells recognizes membrane bound immunoglobin heavy mu chain in Burkitt's lymphoma cells. Mol Cell Proteomics 6:2230–2238. doi:10.1074/mcp.M700026-MCP200
86. Green LS, Jellinek D, Jenison R et al (1996) Inhibitory DNA ligands to platelet-derived growth factor B-chain. Biochemistry 35:14413–14424. doi:10.1021/bi961544
87. Mann D, Reinemann C, Stoltenburg R, Strehlitz B (2005) In vitro selection of DNA aptamers binding ethanolamine. Biochem Biophys Res Commun 338:1928–1934
88. Zhou J, Li H, Li S et al (2008) Novel dual inhibitory function aptamer-siRNA delivery system for HIV-1 therapy. Mol Ther 16:1481–1489. doi:10.1038/mt.2008.92
89. Swensen JS, Xiao Y, Ferguson BS et al (2009) Continuous, real-time monitoring of cocaine in undiluted blood serum via a microfluidic, electrochemical aptamer-based sensor. J Am Chem Soc 131:4262–4266. doi:10.1021/ja806531z
90. Tong GJ, Hsiao SC, Carrico ZM, Francis MB (2009) Viral capsid DNA aptamer conjugates as multivalent cell-targeting vehicles. J Am Chem Soc 131:11174–11178. doi:10.1021/ja903857f
91. Li W, Yang X, Wang K et al (2008) Real-time imaging of protein internalization using aptamer conjugates. Anal Chem 80:5002–5008. doi:10.1021/ac800930q
92. Chen HW, Medley CD, Sefah K et al (2008) Molecular recognition of small-cell lung cancer cells using aptamers. ChemMedChem 3:991–1001. doi:10.1002/cmdc.200800030
93. Xiao Y, Lubin AA, Heeger AJ, Plaxco KW (2005) Label-free electronic detection of thrombin in blood serum by using an aptamer-based sensor. Angew Chem Int Ed 44:5456–5459. doi:10.1002/anie.200500989

94. McNamara JO, Kolonias D, Pastor F et al (2008) Multivalent 4-1BB binding aptamers costimulate CD8+ T cells and inhibit tumor growth in mice. J Clin Invest 118:376–386. doi:10.1172/JCI33365
95. Nimjee SM, Rusconi CP, Sullenger B (2005) Aptamers: an emerging class of therapeutics. Annu Rev Med 56:555–583. doi:10.1146/annurev.med.56.062904.144915
96. Famulok M (1999) Oligonucleotide aptamers that recognize small molecules. Curr Opin Struct Biol 9:324–329
97. Cho M, Xiaoa Y, Niec J et al (2011) Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. Proc Natl Acad Sci U S A 108:4105–4110. doi:10.1073/pnas.1015181108
98. Kupakuwana GV, Crill JE, McPike MP, Borer PN (2011) Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. PLoS ONE. doi:10.1371/journal.pone.0019395
99. Hoon S, Zhou B, Janda KD et al (2011) Aptamer selection by high-throughput sequencing and informatic analysis. Biotechniques 51:413–416. doi:10.2144/000113786
100. Gold L, Ayers D, Bertino J et al (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS ONE. doi:10.1371/journal.pone.0015004
101. Latham JA, Johnson R, Toole JJ (1994) The application of a modified nucleotide in aptamer selection: novel thrombin aptamers containing-(1-pentynyl)-2′-deoxyuridine. Nucleic Acids Res 22:2817–2822. doi:10.1093/nar/22.14.2817
102. Kimoto M, Yamashige R, Matsunaga K et al (2013) Generation of high-affinity DNA aptamers using an expanded genetic alphabet. Nat Biotechnol 31:453–457. doi:10.1038/nbt.2556
103. Declerck PJ (2012) Biologicals and biosimilars: a review of the science and its implications. Generics Biosimilars Initiat J 1:13–16. doi:10.5639/gabij.2012.0101.005
104. Pindur U, Jansen M, Lemster T (2005) Advances in DNA-ligands with groove binding, intercalating and/or alkylating activity: chemistry, DNA-binding and biology. Curr Med Chem 12:2805–2847. doi:10.2174/092986705774454698
105. Gottesfeld JM, Neely L, Trauger JW et al (1997) Regulation of gene expression by small molecules. Nature 387:202–205
106. Kopka ML, Yoon C, Goodsell D et al (1985) The molecular origin of DNA-drug specificity in netropsin and distamycin. Proc Natl Acad Sci U S A 82:1376–1380. doi:10.1073/pnas.82.5.1376
107. Wade WS, Mrksich M, Dervan PB (1992) Design of peptides that bind in the minor groove of DNA at 5′-(A, T)G(A, T)C(A, T)-3′ sequences by a dimeric side-by-side motif. J Am Chem Soc 114:8783–8794. doi:10.1021/ja00049a006
108. Mrksich M, Wade WS, Dwyer TJ et al (1992) Antiparallel side-by-side dimeric motif for sequence-specific recognition in the minor groove of DNA by the designed peptide 1-methylimidazole-2-carboxamide netropsin. Proc Natl Acad Sci 89:7586–7590. doi:10.1073/pnas.89.16.7586
109. Kielkopf CL, Baird EE, Dervan PB, Rees DC (1998) Structural basis for G.C recognition in the DNA minor groove. Nat Struct Biol 5:104–109
110. Dervan PB (2001) Molecular recognition of DNA by small molecules. Bioorg Med Chem 9:2215–2235
111. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. Curr Opin Struct Biol 13:284–299
112. Dervan PB, Doss RM, Marques MA (2005) Programmable DNA binding oligomers for control of transcription. Curr Med Chem Anticancer Agents 5:373–387. doi:10.2174/1568011054222346
113. White S, Baird EE, Dervan PB (1997) Orientation preferences of pyrrole-imidazole polyamides in the minor groove of DNA. J Am Chem Soc 119:8756–8765. doi:10.1021/ja971569b

114. Turner JM, Swalley SE, Baird EE, Dervan PB (1998) Aliphatic/aromatic amino acid pairings for polyamide recognition in the minor groove of DNA. J Am Chem Soc 120:6219–6226. doi:10.1021/ja980147e
115. Wang CC, Ellervik U, Dervan PB (2001) Expanding the recognition of the minor groove of DNA by incorporation of beta-alanine in hairpin polyamides. Bioorg Med Chem 9:653–657
116. Bando T, Minoshima M, Kashiwazaki G et al (2008) Requirement of β-alanine components in sequence-specific DNA alkylation by pyrrole–imidazole conjugates with seven-base pair recognition. Bioorg Med Chem 16:2286–2291
117. Belitsky JM, Leslie SJ, Arora PS et al (2002) Cellular uptake of N-methylpyrrole/N-methylimidazole polyamide-dye conjugates. Bioorg Med Chem 10:3313–3318. doi:10.1016/S0968-0896(02)00204-3
118. Shinohara KI, Narita A, Oyoshi T et al (2004) Sequence-specific gene silencing in mammalian cells by alkylating pyrrole-imidazole polyamides. J Am Chem Soc 126:5113–5118. doi:10.1021/ja031673v
119. Pandian GN, Taniguchi J, Junetha S et al (2014) Distinct DNA-based epigenetic switches trigger transcriptional activation of silent genes in human dermal fibroblasts. Sci Rep 4:3843. doi:10.1038/srep03843
120. Mapp AK, Ansari AZ, Ptashne M, Dervan PB (2000) Activation of gene expression by small molecule transcription factors. Proc Natl Acad Sci U S A 97:3930–3935. doi:10.1073/pnas.97.8.3930
121. Nickols NG, Dervan PB (2007) Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. Proc Natl Acad Sci U S A 104:10418–10423. doi:10.1073/pnas.0704217104
122. Syed J, Pandian GN, Sato S et al (2014) Targeted suppression of EVI1 oncogene expression by sequence-specific pyrrole-imidazole polyamide. Chem Biol 21:1370–1380. doi:10.1016/j.chembiol.2014.07.019
123. Oyoshi T, Kawakami W, Narita A et al (2003) Inhibition of transcription at a coding sequence by alkylating polyamide. J Am Chem Soc 125:4752–4754
124. Shinohara KI, Sasaki S, Minoshima M et al (2006) Alkylation of template strand of coding region causes effective gene silencing. Nucleic Acids Res 34:1189–1195. doi:10.1093/nar/gkl005
125. Taylor RD, Asamitsu S, Takenaka T et al (2014) Sequence-specific DNA alkylation targeting for kras codon 13 mutation by pyrrole-imidazole polyamide seco-cbi conjugates. Chem Eur J 20:1310–1317. doi:10.1002/chem.201303295
126. Hiraoka K, Inoue T, Taylor RD et al (2015) Inhibition of KRAS codon 12 mutants using a novel DNA-alkylating pyrrole–imidazole polyamide conjugate. Nat Commun 6:6706. doi:10.1038/ncomms7706
127. Ohtsuki A, Kimura MT, Minoshima M et al (2009) Synthesis and properties of PI polyamide-SAHA conjugate. Tetrahedron Lett 50:7288–7292. doi:10.1016/j.tetlet.2009.10.034
128. Ansari AZ, Mapp AK, Nguyen DH et al (2001) Towards a minimal motif for artificial transcriptional activators. Chem Biol 8:583–592. doi:10.1016/S1074-5521(01)00037-0
129. Kwon Y, Arndt HD, Mao Q et al (2004) Small molecule transcription factor mimic. J Am Chem Soc 126:15940–15941. doi:10.1021/ja0445140
130. Han L, Pandian GN, Junetha S et al (2013) A synthetic small molecule for targeted transcriptional activation of germ cell genes in a human somatic cell. Angew Chem Int Ed 52:13410–13413
131. Pandian GN, Sato S, Anandhakumar C et al (2014) Identification of a small molecule that turn ‘ON’ the pluripotency gene circuitry in human fibroblasts. ACS Chem Biol 141024155048006. doi:10.1021/cb500724t
132. Syed J, Chandran A, Pandian GN et al (2015) A synthetic transcriptional activator of genes associated with the retina in human dermal fibroblasts. ChemBioChem 16:1497–1501. doi:10.1002/cbic.201500140

133. Dervan PB, Poulin-Kerstien AT, Fechter EJ, Edelson BS (2005) Regulation of gene expression by synthetic DNA-binding ligands. Top Curr Chem 253:1–31. doi:10.1007/b100440
134. Groth A, Rocha W, Verreault A, Almouzni G (2007) Chromatin challenges during DNA replication and repair. Cell 128:721–733
135. Gottesfeld JM, Neely L, Trauger JW et al (1997) Regulation of gene expression by small molecules. Nature 387:202–205
136. Meier JL, Yu A, Korf I et al (2012) Guiding the design of synthetic DNA-binding molecules with massively parallel sequencing. J Am Chem Soc 134:17814–17822. doi:10.1021/ja308888c
137. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res 37:e151. doi:10.1093/nar/gkp802
138. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659. doi:10.1093/bioinformatics/btr261
139. Gopal YNV, Van Dyke MW (2003) Combinatorial determination of sequence specificity for nanomolar DNA-binding hairpin polyamides. Biochemistry 42:6891–6903. doi:10.1021/bi027373s
140. Warren CL, Kratochvil NCS, Hauschild KE et al (2006) Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A 103:867–872. doi:10.1073/pnas.0509843102
141. Puckett JW, Muzikar KA, Tietjen J et al (2007) Quantitative microarray profiling of DNA-binding molecules. J Am Chem Soc 129:12310–12319. doi:10.1021/ja0744899
142. Carlson CD, Warren CL, Hauschild KE et al (2010) Specificity landscapes of DNA binding molecules elucidate biological function. Proc Natl Acad Sci U S A 107:4544–4549. doi:10.1073/pnas.0914023107
143. Kang JS, Meier JL, Dervan PB (2014) Design of sequence-specific DNA binding molecules for DNA methyltransferase inhibition. J Am Chem Soc 136:3687–3694. doi:10.1021/ja500211z
144. Minoshima M, Bando T, Sasaki S et al (2008) Pyrrole-imidazole hairpin polyamides with high affinity at 5′-CGCG-3′ DNA sequence; influence of cytosine methylation on binding. Nucleic Acids Res 36:2889–2894. doi:10.1093/nar/gkn116
145. Pandian GN, Nakano Y, Sato S et al (2012) A synthetic small molecule for rapid induction of multiple pluripotency genes in mouse embryonic fibroblasts. Sci Rep 2:1–8. doi:10.1038/srep00544
146. Erwin GS, Bhimsaria D, Eguchi A, Ansari AZ (2014) Mapping polyamide-DNA interactions in human cells reveals a new design strategy for effective targeting of genomic sites. Angew Chem Int Ed 53:10124–10128. doi:10.1002/anie.201405497
147. Wilhelm BT, Marguerat S, Watt S et al (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239–1243. doi:10.1038/nature07002
148. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. doi:10.1038/nmeth.1226
149. Morin RD, Bainbridge M, Fejes A et al (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45:81–94. doi:10.2144/000112900
150. Olenyuk BZ, Zhang GJ, Klco JM et al (2004) Inhibition of vascular endothelial growth factor with a sequence-specific hypoxia response element antagonist. Proc Natl Acad Sci U S A 101:16768–16773. doi:10.1073/pnas.0407617101
151. Muzikar KA, Nickols NG, Dervan PB (2009) Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression. Proc Natl Acad Sci U S A 106:16598–16603. doi:10.1073/pnas.0909192106

152. Raskatov JA, Meier JL, Puckett JW et al (2012) Modulation of NF-κB-dependent gene transcription using programmable DNA minor groove binders. Proc Natl Acad Sci U S A 109:1023–1028. doi:10.1073/pnas.1118506109
153. Raskatov JA, Nickols NG, Hargrove AE et al (2012) Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide. Proc Natl Acad Sci U S A 109:16041–16045. doi:10.1073/pnas.1214267109
154. Nickols NG, Szablowski JO, Hargrove AE et al (2013) Activity of a py-im polyamide targeted to the estrogen response element. Mol Cancer Ther 12:675–684. doi:10.1158/1535-7163.MCT-12-1040
155. Han J, Yuan P, Yang H et al (2010) Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature 463:1096–1100. doi:10.1038/nature08735
156. Chen X, Xu H, Yuan P et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133:1106–1117. doi:10.1016/j.cell.2008.04.043
157. Heng JCD, Feng B, Han J et al (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. Cell Stem Cell 6:167–174. doi:10.1016/j.stem.2009.12.009
158. Yuan P, Han J, Guo J et al (2009) Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev 23:2507–2520. doi:10.1101/gad.1831909
159. Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13:840–852. doi:10.1038/nrg3306
160. Rodriguez R, Müller S, Yeoman JA et al (2008) A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. J Am Chem Soc 130:15758–15759
161. Rodriguez R, Miller KM, Forment JV et al (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. Nat Chem Biol 8:301–310. doi:10.1038/nchembio.780
162. Yang F, Nickols NG, Li BC et al (2013) Antitumor activity of a pyrrole-imidazole polyamide. Proc Natl Acad Sci U S A 110:1863–1868. doi:10.1073/pnas.1222035110
163. Chenoweth DM, Dervan PB (2010) Structural basis for cyclic Py-Im polyamide allosteric inhibition of nuclear receptor binding. J Am Chem Soc 132:14521–14529. doi:10.1021/ja105068b
164. Chenoweth DM, Dervan PB (2009) Allosteric modulation of DNA by small molecules. Proc Natl Acad Sci U S A 106:13175–13179. doi:10.1073/pnas.0906532106
165. Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30:1084–1094. doi:10.1038/nbt.2421

Chapter 2
Next Generation Sequencing Studies Guide the Design of Pyrrole-Imidazole Polyamides with Improved Binding Specificity by the Addition of β-Alanine

Abstract The identification of binding sites for small molecules in the genome space is important for various applications. Previously, we demonstrated rapid transcriptional activation by our small molecule SAHA-PIPs. However, it was not clear whether the strong biological effects exerted by SAHA-PIP were due to its binding specificity. Here, we used high-throughput sequencing (Bind-n-seq) to identify the binding specificity of SAHA-PIPs. Firstly sequence specificity bias was determined with SAHA-PIPs (**3** and **4**), which showed enhanced 6-bp sequence-specific binding compared with hairpin PIPs (**1** and **2**). This finding allowed us to investigate the role of β-alanine that links SAHA with PIP, which led to the design of ββ-PIPs (**5** and **6**). ββ-PIPs showed enhanced binding specificity. Overall, we demonstrated the importance of β-moieties for the binding specificity of PIPs, and the utilization of cost-effective high-throughput screening of these small molecules to the minor groove.

Keywords DNA recognition · High throughput screening · Bind-n-seq · Next generation sequencing · Synthetic biology · HDAC inhibitor

2.1 Introduction

Dysregulation of gene expression is a major cause of many diseases. Global gene expression is governed by transcriptional regulation, and genetic and epigenetic machinery plays major roles in this. Alteration of the epigenetic mechanism can provide targeted gene regulation and cellular reprogramming, and could enable the development of artificial genetic switches, governed by DNA-binding small molecules, which could bring back otherwise irredeemable dysfunctions connected with defective transcriptional/epigenetic machinery [1]. Hairpin *N*-methylpyrrole (P)–*N*-methylimidazole (I) polyamides (PIPs) are a class of small molecule that can bind in the minor groove of DNA [2] and can recognize each of the four Watson–Crick base pairs [3, 4]. When arranged side-by-side, I and P discriminate C·G from G·C base pairing [2], whereas the P and P pair can recognize both A·T and T·A [5].

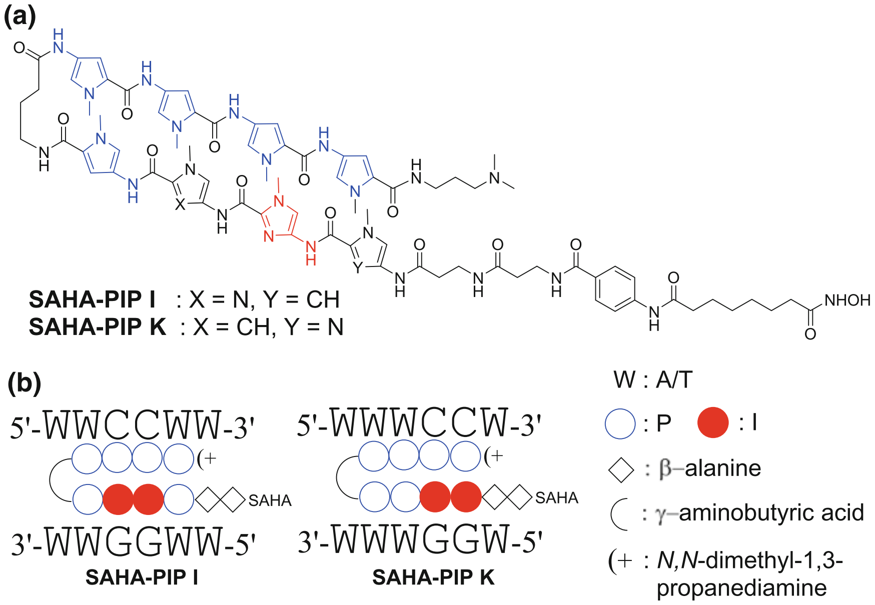
© Springer Nature Singapore Pte Ltd. 2018

A. Chandran, *Advancing Development of Synthetic Gene Regulators*,
Springer Theses, https://doi.org/10.1007/978-981-10-6547-7_2

Chromatin-modifying enzymes can alter transcriptional regulation randomly by epigenome modification [6]. Previously, we linked PIP to an epigenetically active histone deacetylase (HDAC) inhibitor (SAHA) to generate SAHA-PIP [7–10] and demonstrated the effect of SAHA-PIPs in mouse fibroblasts to induce differential activation of pluripotent stem cell-associated genes [8–10]. This study demonstrated the efficiency of SAHA-PIP in intervening in transcriptional activation and showed that the compound may activate the silent gene system in somatic cells. Our recent report on the microarray analysis of a SAHA-PIP library revealed that such compounds can trigger a unique set of gene clusters in human dermal fibroblast (HDF) cells [11]. However, whether the strong biological effects exhibited by SAHA-PIP were due to its binding specificity in a broad context was unclear.

Recently, the binding sites of hairpin PIPs in DNA at the genome level were reported [9]. In this approach, termed Bind-n-seq, PIPs were tagged with biotin and the bound DNA was analyzed by high-throughput sequencing methods [12]. Although the PIPs bound to the DNA with canonical pairing rules as predicted, some PIPs were identified with the reverse binding orientation [12, 13].

In this study, we identify the high-affinity binding sites of the previously reported germ cell gene activating SAHA-PIP K [14] and its structurally similiar counterpart SAHA-PIP I [11] that activate an entirely different set of developmental genes (Scheme 2.1). By using the Bind-n-seq approach, we could identify the binding specificity of SAHA-PIP I (**3**) and SAHA-PIP K (**4**) in a broad context of



Scheme 2.1 **a** Chemical structures, and **b** DNA sequence-specific binding models of SAHA-PIP I and K

Table 2.1 Relative enrichment of 6-base pair “k”mers enriched by **3** in two separate binding and enrichment reactions (only top 30 sequences were shown)

Rank	Sequences	Enrichment ratio (BNS-1)	Enrichment ratio (BNS-1)
1	TACCAA	3.83	4.047
2	ATTACC	3.63	3.89
3	GGTACC	3.231	3.393
4	CTACCA	3.155	3.2
5	GTACCA	3.109	3.166
6	TGGTAA	3.006	3.085
7	AACCAA	2.959	3.169
8	ATGGTA	2.945	2.812
9	AATACC	2.853	3.083
10	ATACCA	2.846	2.918
11	GGTAGA	2.831	2.863
12	TTCCAA	2.795	2.931
13	CTTCCA	2.721	2.827
14	ACCAAT	2.651	2.779
15	GTTCCA	2.634	2.616
16	GGTACA	2.617	2.574
17	AGGTAG	2.559	2.675
18	ACCATC	2.55	2.331
19	CCAACC	2.439	2.503
20	GGAAGA	2.416	2.561
21	ATAACC	2.383	2.59
22	GGAACC	2.359	2.468
23	CCATCC	2.352	2.024
24	AACCAT	2.316	2.369
25	ACCAAG	2.313	2.46
26	GAACCA	2.308	2.38
27	ACTTCC	2.303	2.456
28	ACTACC	2.285	2.442
29	ACCAAC	2.279	2.547
30	AGGTAC	2.276	2.426

Its corresponding graph was shown in Fig. 2.3

oligo libraries. Both compounds showed high-affinity binding compared with PIP I (**1**) and PIP K (**2**) (PI polyamide without SAHA). This prompted us to redesign SAHA-PIP and replace the SAHA moiety with β-alanine (the β-moiety) in the non-core binding region of the 10-ring hairpin polyamide. The results showed that the β-moiety enforces binding of PIP in the closed form (Table 2.1).

Application of the epigenetically active and DNA-binding SAHA-PIPs I and K relies largely on high-affinity sites. SAHA-PIPs are expected to bind in the minor groove of DNA according to well-defined pairing rules (N-terminal to

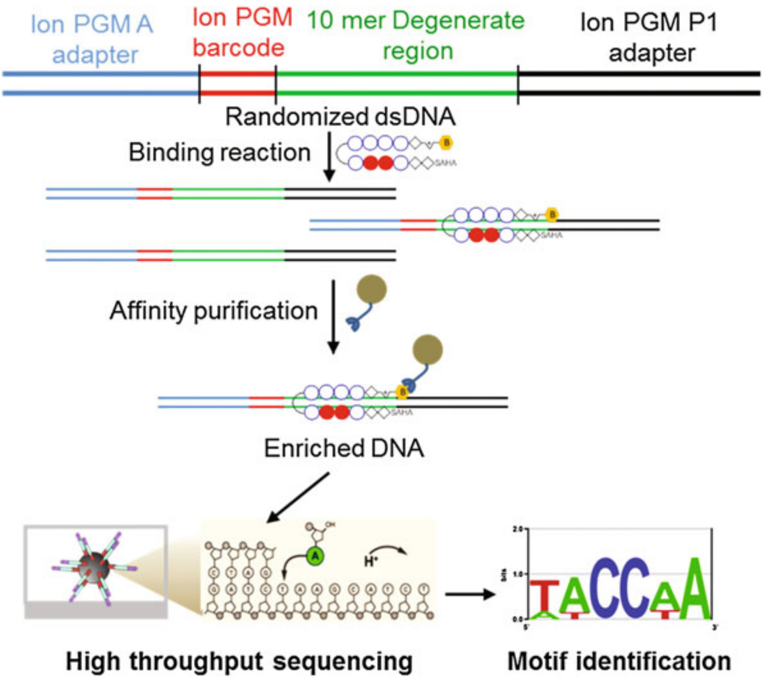


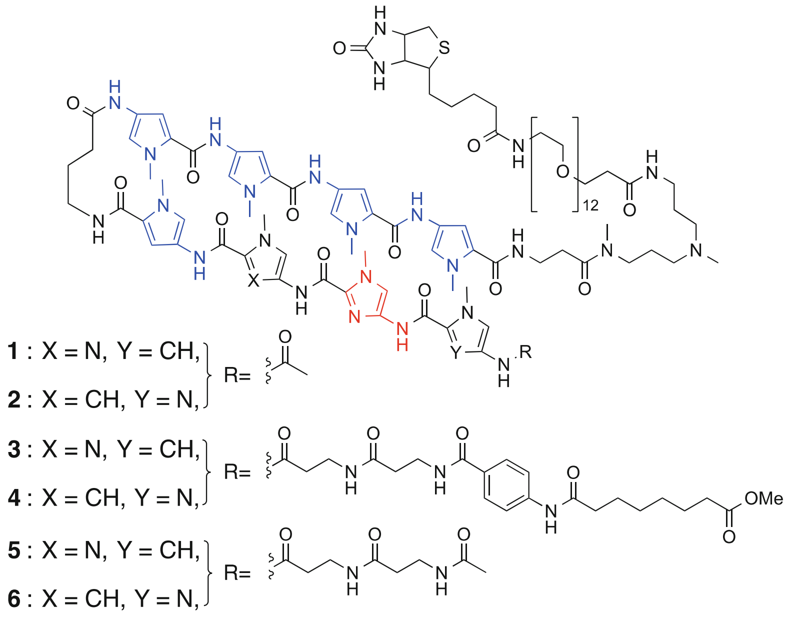
Fig. 2.1 High-throughput sequencing based Bind-n-seq experimental design and binding motif discovery

C-terminal:5′-3′) [7]. Meier et al. studied the sequence selectivity and canonical pairing rules of PIP-DNA binding in a broad sequence context using affinity purification coupled with massively parallel sequencing in an Illumina sequencer [12]. To characterize the binding specificity of the PIP derivatives, we carried out high-throughput sequencing (Bind-n-seq) studies in a randomized DNA sequence space to define the recognition motifs of PIPs and SAHA-PIPs (Fig. 2.1) .

2.2 Results and Discussion

2.2.1 PIP and PIP Conjugates Synthesis and Bind-n-Seq Analysis

A small library of biotin-conjugated PIPs and SAHA-PIPs (**1**–**4**) (Scheme 2.2, **1**–**4**) was first synthesized (in the SAHA moiety of biotin-conjugated SAHA-PIPs (**3** and **4**), we used the methyl ester instead of hydroxamic acid for synthetic convenience. This change was not expected to affect the DNA-binding affinity of the SAHA-PIPs because SAHA does not bind DNA. Synthesis details are provided in the materials



Scheme 2.2 Chemical structures 1–6

and methods) and allowed to bind with 10-bp randomized DNA fragments equipped with an Ion PGM-compatible adapter sequence (for details of the design of the oligos and primer used to make duplex DNA, see the materials and methods). By using affinity purification, the enriched DNA sequences bound by the PIP derivatives were extracted and sequenced with an Ion PGM sequencer. After a quality check and adapter trimming of the sequenced reads, Bind-n-Seq analysis [15] revealed enrichment of "*k*-mer" ($k = 6$), and confirmed that the binding sites of **3** and **4** matched the canonical binding rule. Interestingly, **4** was found to bind in the forward orientation (N-terminal to C-terminal:5′-3′, Fig. 2.2d), but it was difficult to determine the binding orientation of **3** (Fig. 2.2c) because of its symmetrical binding site (5′-WWCCWW-3′/5′-WWGGWW-3′). The results obtained with **4** are significant because the PIPs have the ability to bind in both forward and reverse orientation and hence have a relatively large number of duplex DNA sites as potential targets. When the binding of **4** is principally in the forward orientation, then the number of potential target sites is reduced, which in turn suggests superior specificity over **2**. Motifs were identified by using the DREME [16] and enoLOGOS [17] algorithms, which are motif-binding programs developed to identify the short response elements typically bound by eukaryotic transcription factors. From the analysis described above, a graphical depiction of the strongest motif was generated for the highly enriched binding site of **1**–**4** (Fig. 2.2a); a table of enriched sequences is given in Tables 2.1 and 2.2. Replicate quantification validated the high-affinity binding; for example, the enrichment of DNA sequences

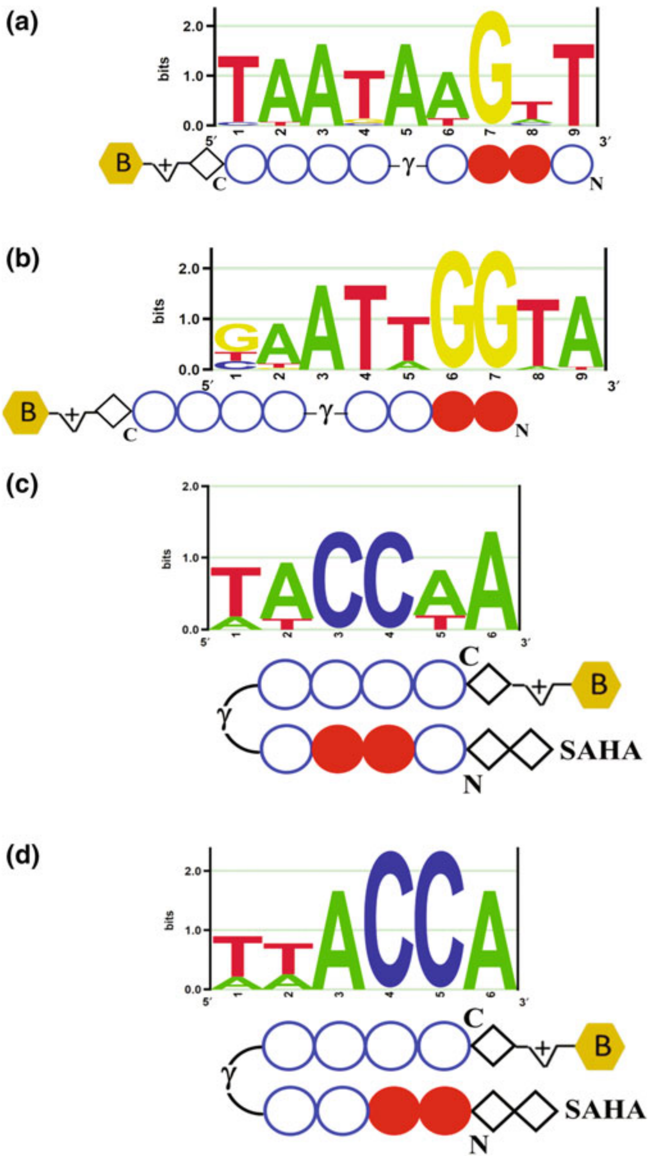


Fig. 2.2 Structure of Polyamides binding region and its primary binding motifs identified using Bind-n-seq through next generation sequencing: **a** 1 **b** 2 **c** 3 **d** 4

bound by **3** was highly reproducible ($R^2 = 0.92$ for two separate binding and enrichment experiments; Fig. 2.3). The enriched sequence data and the similarity of enrichment between two different experiments are shown in the Table 2.1.

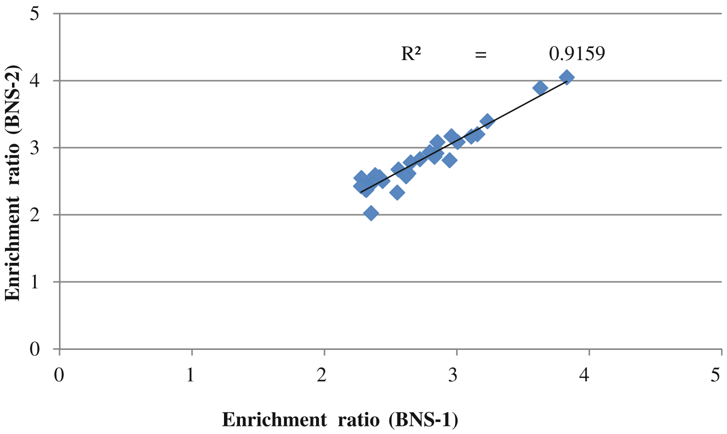


Fig. 2.3 Graphical representation of relative enrichment of two separate binding and enrichment reactions of 3 (sequence details are provided in Table 2.2)

2.2.2 PIP Redesign, ββ-PIP Synthesis and Bind-n-Seq Analysis

The results of the massively parallel sequencing analysis Bind-n-seq showed that **3** and **4** (Fig. 2.2c, d) had high-affinity binding toward the canonical binding sites. However, the results also showed that **1** and **2** bound with randomized DNA in an open conformation (Fig. 2.2a, b). In addition to the SAHA moiety, **3** and **4** contain two β-alanine moieties (used to conjugate the SAHA with PIP) in the non-core binding machinery; compounds **1** and **2** lacks these two β-moieties. The high-affinity binding of **3** and **4** prompted us to investigate the role of the β-alanine moiety in the construction of hairpin PIP. To this end, we designed PIP derivatives **5** and **6** (corresponding to SAHA-PIPs without SAHA) (Scheme 2.2, **5**–**6**), having two β-alanine moieties in the non-core binding domain. Thus, ββ-PIP-biotin conjugates **5** and **6** were synthesized and subjected to binding reaction with randomized DNA with affinity purification followed by massively parallel sequencing.

Bind-n-seq data analyses were performed for sequence reads enriched by **5** and **6**. Assessment of the identified motifs (Fig. 2.4a, b, corresponding sequence details in Tables 2.3 and 2.4) revealed that the β-alanine moiety enforces high-affinity binding of the PIPs in the closed form. Compounds **5** and **6** bind with the WWCCWA and TWACCA/AWTGGT sites, respectively. Since we could not identify the binding orientation of **5**, we considered only the binding orientation of **6**. Interestingly, DNA motifs enriched after binding **6** revealed that this compound bound to the DNA with equal affinity of forward and reverse orientations. This investigation thus indicates that the β-alanine moiety induces binding of hairpin PIP in the closed form with DNA and does not affect the direction of binding.

Table 2.2 Relative enrichment of 6-base pair “k”mers enriched by **4** binding and enrichment reactions

Rank	Sequence	Enrichment
1	TTACCA	4.26
2	ATTACC	3.822
3	AATGGT	3.454
4	ATAACC	3.37
5	ATGGTA	3.205
6	ATTGGT	2.916
7	TAACCA	2.783
8	TTGGTA	2.631
9	TATGGT	2.568
10	CTTACC	2.352
11	ATACCA	2.286
12	ATATGG	2.267
13	GTTACC	2.246
14	TATTGG	2.231
15	TAATGG	2.185
16	AATTGG	2.105
17	TATACC	2.086

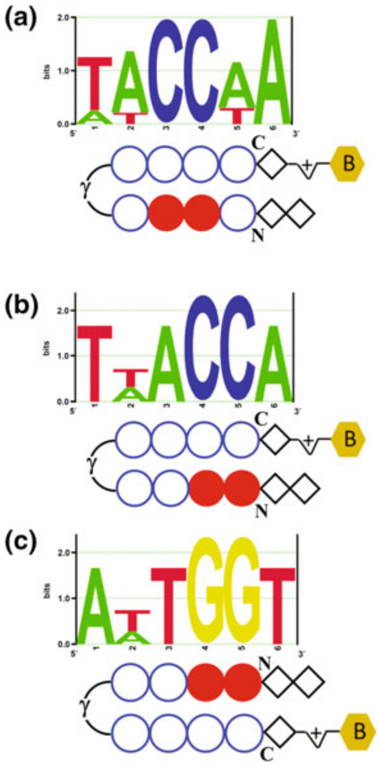


Fig. 2.4 Structural representation of redesigned polyamides binding region and its primary binding motifs identified using Bind-n-seq through next generation sequencing: **a** 5 **b** 6

2.2.3 Binding Affinity Conformation by SPR

To confirm the findings of Bind-n-seq analyses, we performed SPR analysis because previous reports have shown that the rates of association (k_a) and dissociation (k_d), and the dissociation constant (K_D) correlate well with the binding affinity [13]. For the SPR analysis, we synthesized PIPs **7–10** without biotin conjugates (Scheme 2.3). SPR analyses were performed with a Biacore X instrument with the match and mismatch DNA oligonucleotides from the human transcription factor sequences to identify whether PIP binding can selectively trigger human transcription factors (See materials and methods) [18]. SAHA-PIP K and I were shown to distinctively regulate the expression of silent developmental genes in human fibroblasts. Based on our previous reports, for **7** and **9** (similar binding sequence as SAHA-PIP I), we selected the match sequences in the promoter region of the gene encoding the developmental associated POU homeodomain [11] and for **8** and **10** (similar binding sequence as SAHA-PIP K), we selected the match (**8** and **10**) sequences in the promoter region of the gene encoding the germ cell associated *PIWIL 2* [14]. For checking the relative binding affinity, the match sequences of SAHA-PIP I (**7** and **9**) and K (**8** and **10**) were interchanged and used as the mismatch sequences. We compared the binding ability and efficiency of compounds **9** and **10** with those of **7** and **8** in the corresponding match and mismatch binding of DNA. The PIP derivatives were passed through a 5′-biotinylated hairpin DNA immobilized sensor chip by using a biotin–avidin system. The binding affinities of the PIP derivatives were measured by analyzing the data produced from the SPR method. Detailed assessment of the SPR dissociation constant (K_D) values showed that **9** and **10** had higher affinity toward the matching polyamide–DNA binding sites than **7** and **8**, with a K_D range of approximately 5–25-fold. These data correlate best with the binding sites identified by Bind-n-seq. The rates of association (k_a) and dissociation (k_d) and the dissociation constants (K_D) are shown in Table 2.5, and the sensorgrams are shown in the Supplementary Information (Figs. 2.5 and 2.6). Although many methods are under development, the targeting of dsDNA in a genomic sequence perspective by using small synthetic molecules remains difficult. The structural composition of polyamides and DNA sequence-dependent structural variations may decrease the binding specificity of PIP derivatives to the desired specific DNA-binding sites. This study delivers a low-cost sequencing method for designing and screening of sequence-specific DNA-binding molecules with the standardized experimental conditions. We have shown that the addition of a SAHA moiety to the N-terminal of PIPs does not affect the canonical binding rule of polyamides in a wide sequence context. Thus, the HDAC inhibitor SAHA can be used in a sequence-specific manner with the assistance of polyamides as we reported earlier [14].

Table 2.3 Relative enrichment of 6-base pair "k"mers enriched by **5** binding and enrichment reactions. (only top 30 sequences were shown)

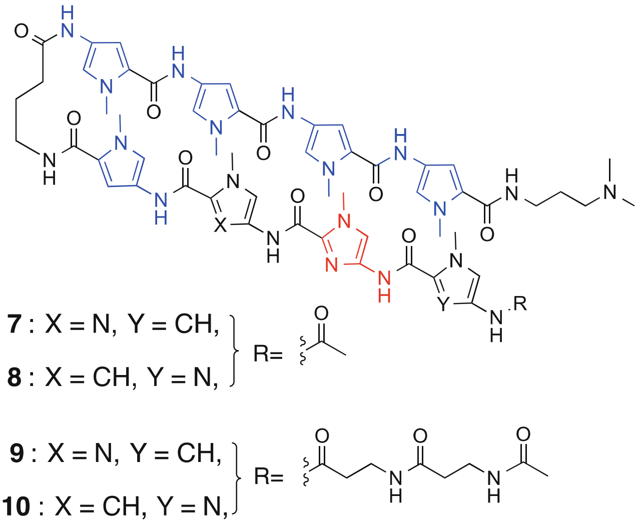
Rank	Sequence	Enrichment
1	TACCAA	3.635
2	ATTACC	3.585
3	CTACCA	2.993
4	TGGTAA	2.785
5	GTACCA	2.982
6	AACCAA	2.745
7	ATGGTA	2.754
8	TACCTA	2.325
9	ACTACC	2.541
10	GGTAGA	2.641
11	ATACCA	2.693
12	ACCAAT	2.539
13	CTTCCA	2.531
14	ACCAAC	2.197
15	GGTACA	2.393
16	TTCCAA	2.513
17	AGGTAG	2.568
18	GTTCCA	2.363
19	AATACC	2.722
20	CAACCA	2.085
21	GGAAGA	2.335
22	ACCAAG	2.215
23	AAGGTA	2.208
24	AGTACC	2.256
25	ATAACC	2.224
26	GAACCA	2.089
27	ACTTCC	2.232
28	ACCTAC	2.034
29	AGGTAC	2.234
30	ACCATC	2.445

2.2.4 Discussion

Despite the finding that polyamides bind according to the canonical binding rule, we have identified that **1** and **2** (without SAHA) can also preferentially bind to DNA in an open conformation instead of in the hairpin form. This phenomenon was not observed for **3** and **4** (PIPs with SAHA), which prompted us to look in more detail at the features leading to polyamide binding in the closed form. Our investigation focused on the two N-terminal β-alanine moieties in the non-core binding machinery. Several biologically active polyamides with β-residues in the core binding sites have been reported, and the importance of β-residue placement in the core binding region of polyamides has been examined [19]. However, none of the

Table 2.4 Relative enrichment of 6-base pair “k”mers enriched by **6** binding and enrichment reactions

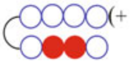
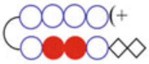
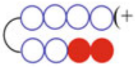
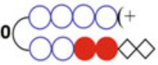
Rank	Sequence	Enrichment
1	AATGGT	3.147
2	TTACCA	3.142
3	ATGGTA	2.851
4	ATTACC	2.835
5	ATAACC	2.476
6	ATTGGT	2.395
7	TTGGTA	2.338
8	TAATGG	2.148
9	TATGGT	2.126
10	TAACCA	2.101



Scheme 2.3 Chemical structures 7–10

studies investigated the effect of β-moieties in the N-terminal non-core binding machinery. We studied the Bind-n-seq offered high-affinity binding of polyamides with ββ-PIP derivatives **5** and **6** and established which sequences are enriched by ββ-PIPs with a high-throughput sequencing method. The identified motifs clearly indicate that the N-terminal β-residue can restore the forward polyamide-binding specificity. This conclusion was further supported by the results of SPR analysis. Although the ββ-PIP derivatives can restore the forward binding affinity, the motifs identified by **6** show both forward and reverse binding orientations, in which the core binding N-terminal of the polyamide is *N*-methylimidazole (I). This result suggests that the ββ-residues direct the polyamide to bind in the hairpin form. Previous studies have shown [12] that the replacement of a β-amino–GABA linker with an α-amino–GABA linker can restore the forward binding of polyamides. Our

Table 2.5 Binding affinities of polyamide 7–10

polyamide		5'-Biotin labeled- GCGCCTTCCTTCCCCT T 3'- CGCGGAAGGAAGGGGT T		5'-Biotin labeled- CGTCCTTTCCAGCAGT T 3'-GCAGGAAAGGTCGTC T T		specificity
7		K_D(M) match **1.2 x 10^{-7}**	*Ka*($M^{-1}s^{-1}$) 1.4 x 10^5 *Kd*(s^{-1}) 1.6 x 10^{-2}	K_D(M) 2-bp mismatch **1.8 x 10^{-7}**	*Ka*($M^{-1}s^{-1}$) 1.0 x 10^5 *Kd*(s^{-1}) 1.8 x 10^{-2}	**1.5**
9		**2.8 x 10^{-8}**	*Ka*($M^{-1}s^{-1}$) 1.2 x 10^6 *Kd*(s^{-1}) 3.3 x 10^{-2}	**1.4 x 10^{-7}**	*Ka*($M^{-1}s^{-1}$) 1.6 x 10^5 *Kd*(s^{-1}) 2.1 x 10^{-2}	**5.0**
8		K_D(M) 2-bp mismatch **1.1 x 10^{-7}**	*Ka*($M^{-1}s^{-1}$) 1.4 x 10^5 *Kd*(s^{-1}) 1.6 x 10^{-2}	K_D(M) match **1.5 x 10^{-8}**	*Ka*($M^{-1}s^{-1}$) 2.5 x 10^5 *Kd*(s^{-1}) 3.7 x 10^{-3}	**7.7**
10		**6.5 x 10^{-8}**	*Ka*($M^{-1}s^{-1}$) 2.6 x 10^5 *Kd*(s^{-1}) 1.7 x 10^{-2}	**2.5 x 10^{-9}**	*Ka*($M^{-1}s^{-1}$) 1.3 x 10^6 *Kd*(s^{-1}) 3.1 x 10^{-3}	**26**

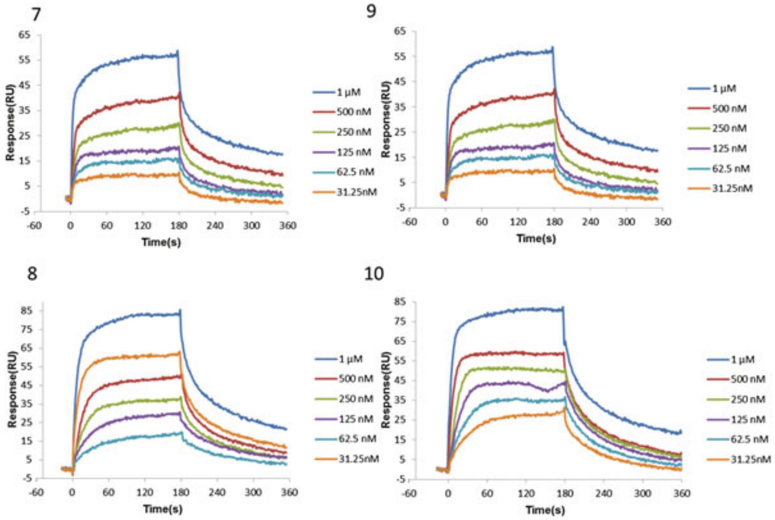


Fig. 2.5 SPR sensorgrams for the interactions of **7** and **9** with its match and **8** and **10** with its mismatch binding sequences in the promoter region of the gene encoding POU homeodomain are shown (Binding specificities were provided in Table 2.1)

results with **5** show that the N-terminal non-core β-residues can also restore the high forward binding affinity of polyamides, in which the core binding N-terminal of the polyamide is *N*-methylpyrrole (P). Further studies need to be carried out to confirm the exact number of β-residues required to restore the forward orientation of polyamide binding.

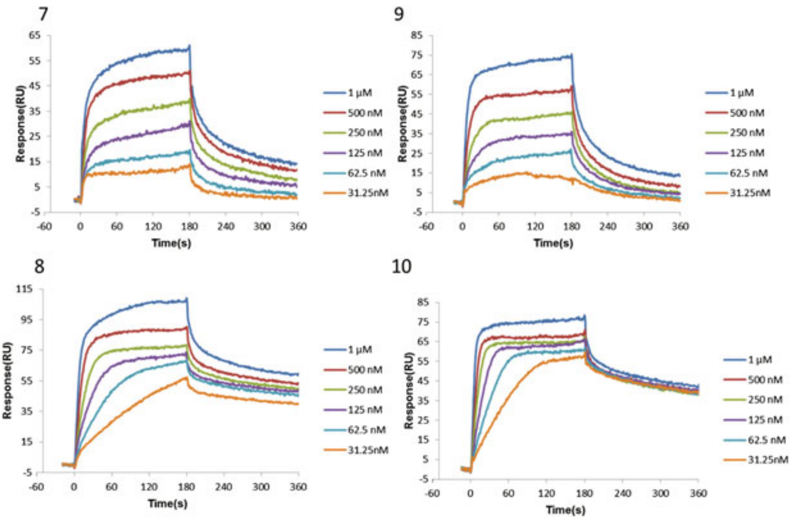


Fig. 2.6 SPR sensorgrams for the interactions of **7** and **9** with its mismatch and **8** and **10** with its match binding sequences in the promoter region of the gene encoding *PIWIL2* are shown (Binding specificities were provided in Table 2.1)

2.3 Materials and Methods

2.3.1 General Materials and Synthesis

Reagents and solvents were purchased from standard suppliers and used without further purification. The EZ-Link NHS-PEG$_{12}$-Biotin was purchased from Thermo Scientific (number 21312). The analytical HPLC was performed with a COSMOSIL 5C18-MS-II reversed phase column (4.6 × 150 mm, Nacalai) in 0.1% TFA in water with CH_3CN as eluent at a flow rate of 1.0 mL/min, and a linear gradient elution of 0–100% CH_3CN over 20 or 40 min with detection at 254 nm. The HPLC purification was performed with a COSMOSIL 5C18-MS-II reversed phase column (10 × 150 mm, Nacalai) in 0.1% TFA in water with CH_3CN as eluent. The final products were analyzed by ESI-TOF-MS (Bruker).

2.3.1.1 General Scheme for Synthesis of Biotin-Conjugated Polyamides

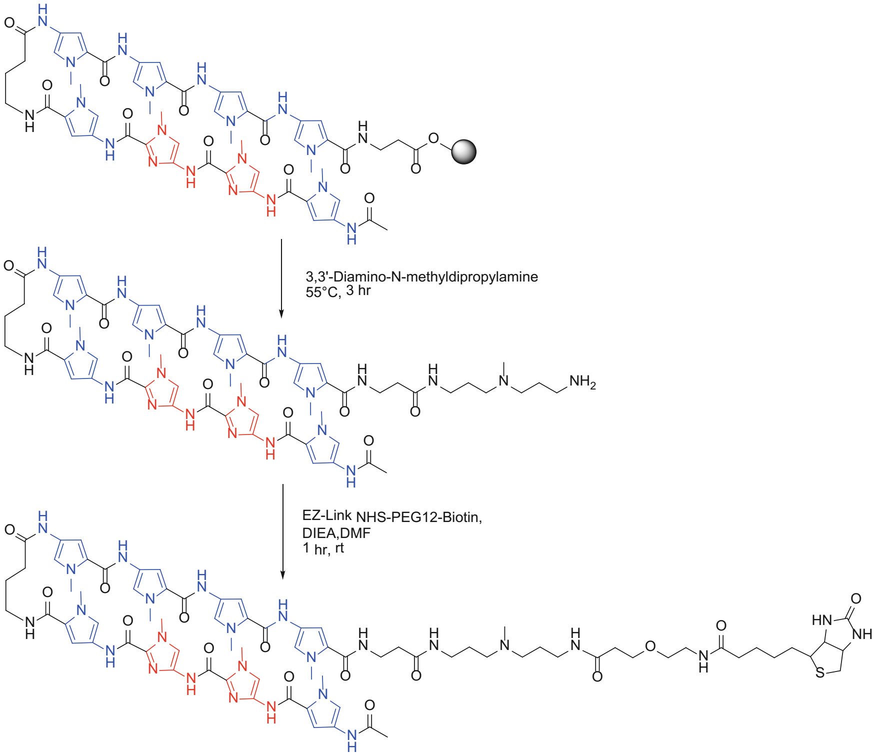
Ac-Py-Im-Im-Py-γ-Py-Py-Py-Py-β-(+)-PEG$_{12}$-Biotin (1)

Py-Im polyamide supported by oxime resin was prepared in a stepwise reaction by a reported Fmoc solid-phase procedure. The product with oxime resin was cleaved with 3,3′-Diamino-*N*-methyldipropylamine (500 μL) at 55 °C for 3 h. After filtration and evaporation, the resulted oil was quenched by Et_2O. The obtained precipitation was washed with Et_2O for three times and dried in vacuum. The product was used for the next coupling steps without further purification.

A solution of polyamide (1.32 mg, 1 μmol), EZ-Link NHS-PEG_{12}-Biotin (1 mg, 1.2 μmol) and DIEA (2 μL, 10 μmol) in DMF (100 μL) was stirred at room temperature for 1 h. After consumption of starting material was confirmed by HPLC, Et_2O was added to the mixture and the resultant was collected by centrifugation, and washed by Et_2O and CH_2Cl_2. The crude product was purified by reverse-phase HPLC. After lyophilization **1** was obtained (0.9 mg, yield 42%). Analytical HPLC: $t_R = 15.5$ min (0.1% TFA-CH_3CN, 0–100%, 40 min). ESI-TOF-MS m/z: calcd for $C_{99}H_{146}N_{26}O_{26}S$ $[M + 2H]^{2+}$ 2149.0622, found 2149.0460.

Ac-Im-Im-Py-Py-γ-Py-Py-Py-Py-β-(+)-PEG_{12}-Biotin (2)

A similar synthetic procedure of **1** was used for the preparation of **2**. Analytical HPLC: $t_R = 15.7$ min (0.1% TFA-CH_3CN, 0–100%, 40 min). ESI-TOF-MS m/z: calcd for $C_{99}H_{146}N_{26}O_{26}S$ $[M + 2H]^{2+}$ 2149.0622, found 2149.1082.



SAHA-β-β-Py-Im-Im-Py-γ-Py-Py-Py-Py-β-(+)-PEG$_{12}$-Biotin (3)

A similar synthetic procedure of **1** was used for the preparation of **3**. Analytical HPLC: t_R = 17.2 min (0.1% TFA-CH_3CN, 0–100%, 40 min). ESI-TOF-MS m/z: calcd for $C_{119}H_{173}N_{29}O_{31}S$ $[M + 2H]^{2+}$ 2538.2573, found 2538.0851.

SAHA-β-β-Im-Im-Py-Py-γ-Py-Py-Py-Py-β-(+)-PEG$_{12}$-Biotin (4)

A similar synthetic procedure of **1** was used for the preparation of **4**. Analytical HPLC: t_R = 17.3 min (0.1% TFA-CH_3CN, 0–100%, 40 min). ESI-TOF-MS m/z: calcd for $C_{119}H_{173}N_{29}O_{31}S$ $[M + 2H]^{2+}$ 2538.2573, found 2538.2536.

Ac-β-β-Py-Im-Im-Py-γ-Py-Py-Py-Py-β-(+)-PEG$_{12}$-Biotin (5)

A similar synthetic procedure of **1** was used for the preparation of **5**. Analytical HPLC: t_R = 9.1 min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{105}H_{156}N_{28}O_{28}S$ $[M + 2H]^{2+}$ 2291.1365, found 2290.8420.

Ac-β-β-Im-Im-Py-Py-γ-Py-Py-Py-Py-β-(+)-PEG$_{12}$-Biotin (6)

A similar synthetic procedure of **1** was used for the preparation of **6**. Analytical HPLC: t_R = 9.1 min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{105}H_{156}N_{28}O_{28}S$ $[M + 2H]^{2+}$ 2291.1365, found 2290.8356.

2.3.2 *Oligomer Sequences Used in Bind-n-Seq Method*

Bind-n-Seq 92mer:

5'CCATCTCATCCCTGCGTGTCTCCGACTCAGXXXXXXXXXXNNNNNNNNNNAAT
CACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG 3`

- All barcoded Bind-n-Seq 92 mers were synthesized by Sigma Aldrich machine mixing, standard desalting purification.
- Barcode in yellow shadow (XXXXXXXXXX), 10-letter barcodes used as per Ion torrent sequencing technologies.

Primer 1:

5′-CCA CTA CGC CTC CGC TTT CCT CTC TA-3′

- Used in initial primer extension reaction.
- Synthesized by Sigma Aldrich, purification by standard desalting.

2.3.3 Bind-n-Seq

Previously reported Bind-n-Seq experiments [12, 13] and subsequent analysis to evaluate small molecule binding affinity towards specific DNA sequences in a broad context sequence pool were customized with the modifications suitable for Ion torrent PGM sequencer. The scheme involves the following three major steps,

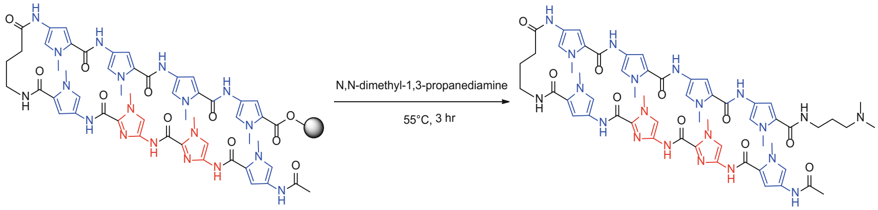
(1) Synthesis of biotinylated PIPs, and randomized oligonucleotides with high-throughput sequencing platform specific adapters (Ion torrent PGM). Adapter ligated oligonucleotides (3 μM) were duplexed by primer extension with adapter specific primer 1 (9 μM) in 25 μL reaction contain 2× goTag PCR master mix with 2 mM Mg^{2+}. Reactions were performed at 95 °C (3 min), 60 °C (2 min), 70 °C (5 min) and then 4 °C using Bio-Rad thermocycler. Biotin conjucated PIPs (100 nM) were allowed to equilibrate with duplex random oligonucleotides for 14 h followed by Streptavidin M-280 Dynabeads (Beads were prepared based on the previous report [12, 13]) separation of the bound and unbound sequences using affinity purification.

(2) Polyamides-enrichment recovered DNA was diluted 1:10 and amplified with sequencing library adapter specific primer for 15 cycles in order to obtain enough sequencing template. After purification enriched libraries were subjected to quality and quantity check with Agilent DNA High sensitivity BioAnalyzer kit (Agilent Technologies, USA). The qualified libraries were used for template preparation using Ion PGM^{TM} template OT2 200 kit in Ion one touch2 system. The templates were then enriched using Ion one touch ES. The enriched libraries were sequenced as a single read sequencing with Ion PGM sequencer using Ion PGMTM sequencing 200 kit v2 and 318 chips (Life Technologies, USA) by following the manufacturer's instructions.

(3) The sequenced reads (composed of A, C, T, or G was then processed to obtain a valid constant region and unique random region) were retained and split into separate files through unique 10-nt ion Xpress-barcode using the Ion torrent suit 3.4.2 as mentioned before. To count the amount of PIP enriched unique DNA sequences, a sliding window of length k (=6) in MERMADE, a new pipeline for Bind-n-Seq analysis (http://korflab.ucdavis.edu/Datasets/BindNSeq.27) were used. Highly enriched motifs were confirmed with DREME primary motif analysis.

2.3.4 Surface Plasmon Resonance (SPR)

2.3.4.1 General Scheme for Synthesis of Polyamides Used for SPR Analysis



Ac-Py-Im-Im-Py-γ-Py-Py-Py-Py-Dp (7)

Py-Im polyamide supported by oxime resin was prepared in a stepwise reaction by a reported Fmoc solid-phase procedure. The product with oxime resin was cleaved with *N*,*N*-dimethyl-1,3-propanediamine (500 μL) at 55 °C for 3 h. After filtration and evaporation, the resulted oil was quenched by Et_2O. The obtained precipitation was washed with Et_2O for three times and dried in vacuo. The crude product was purified by reverse-phase HPLC. Analytical HPLC: $t_R = 9.0$ min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{57}H_{69}N_{21}O_{10}$ $[M + 2H]^{2+}$ 1209.5536, found 1209.3718.

Ac-Im-Im-Py-Py-γ-Py-Py-Py-Py-Dp (8)

A similar synthetic procedure of **7** was used for the preparation of **8**. Analytical HPLC: $t_R = 9.4$ min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{57}H_{69}N_{21}O_{10}$ $[M + 2H]^{2+}$ 1209.5536, found 1209.4150.

Ac-β-β-Py-Im-Im-Py-γ-Py-Py-Py-Py-Dp (9)

A similar synthetic procedure of **7** was used for the preparation of **9**. Analytical HPLC: $t_R = 9.2$ min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{63}H_{79}N_{23}O_{12}$ $[M + 2H]^{2+}$ 1351.6279, found 1351.5146.

Ac-β-β-Im-Im-Py-Py-γ-Py-Py-Py-Py-Dp (10)

A similar synthetic procedure of **7** was used for the preparation of **10**. Analytical HPLC: $t_R = 9.2$ min (0.1% TFA-CH_3CN, 0–100%, 20 min). ESI-TOF-MS m/z: calcd for $C_{63}H_{79}N_{23}O_{12}$ $[M + 2H]^{2+}$ 1351.6279, found 1351.5112.

2.3.4.2 SPR Analysis

SPR analyses were performed as described in previous studies [18] using BIACORE X instrument. In brief, we purchased biotinylated hairpin DNAs with match and mismatch PIP binding sites from JBioS (Tokyo, Japan) (**7** and **9** match/**8** and **10** mismatch: Sequence in the promoter region of the gene encoding POU homeodomain [8]-5′-Biotin- GCG CCT TCC TTC CCC TTT TGG GGA AGG AAG GCG C-3′ and **7** and **9** mismatch/**8** and **10** match: Sequence in the promoter region of the gene encoding PIWIL2 [10]-5′-Biotin- CGT CCT TTC CAG CAG TTTTCT GCT GGA AAG GAC G-3′). The purchased DNAs were diluted to 100 nM. Then immobilization of hairpin biotinylated DNAs was performed on a streptavidin-coated sensor chip SA to obtain the desired immobilization level (approximately 1100–1200 RU rise). HBS-EP buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 3 mM EDTA, and 0.005% surfactant P20) with 0.1% DMSO at room temperature was used to carry out SPR assays. We prepared an order of **7–10** (without biotin) solutions with different concentrations in HBS-EP buffer with 0.1% DMSO and at the flow rate of 20 mL/min, the PIPs were passed on the immobilized oligonucleotide chips. The data processing was performed using BIAevaluation 4.1 program to measure the rates of association (k_a) and dissociation (k_d) and dissociation constant (K_D), with an appropriate fitting model.

References

1. Berdasco M, Esteller M (2010) Aberrant epigenetic landscape in cancer: how cellular identity goes awry. Dev Cell 19:698–711
2. Wade WS, Mrksich M, Dervan PB (1992) Design of peptides that bind in the minor groove of DNA at 5′-(A, T)G(A, T)C(A, T)-3′ sequences by a dimeric side-by-side motif. J Am Chem Soc 114:8783–8794. doi:10.1021/ja00049a006
3. Wemmer DE, Dervan PB (1997) Targeting the minor groove of DNA. Curr Opin Struct Biol 7:355–361. doi:10.1016/S0959-440X(97)80051-6
4. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. Curr Opin Struct Biol 13:284–299
5. Pelton JG, Wemmer DE (1989) Structural characterization of a 2:1 distamycin A.d (CGCAAATTGGC) complex by two-dimensional NMR. Proc Natl Acad Sci U S A 86:5723–5727. doi:10.1073/pnas.86.15.5723
6. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705
7. Ohtsuki A, Kimura MT, Minoshima M et al (2009) Synthesis and properties of PI polyamide-SAHA conjugate. Tetrahedron Lett 50:7288–7292. doi:10.1016/j.tetlet.2009.10.034
8. Pandian GN, Shinohara KI, Ohtsuki A et al (2011) Synthetic small molecules for epigenetic activation of pluripotency genes in mouse embryonic fibroblasts. ChemBioChem 12:2822–2828. doi:10.1002/cbic.201100597
9. Pandian GN, Ohtsuki A, Bando T et al (2012) Development of programmable small DNA-binding molecules with epigenetic activity for induction of core pluripotency genes. Bioorg Med Chem 20:2656–2660. doi:10.1016/j.bmc.2012.02.032

10. Pandian GN, Nakano Y, Sato S et al (2012) A synthetic small molecule for rapid induction of multiple pluripotency genes in mouse embryonic fibroblasts. Sci Rep 2:1–8. doi:10.1038/srep00544
11. Pandian GN, Taniguchi J, Junetha S et al (2014) Distinct DNA-based epigenetic switches trigger transcriptional activation of silent genes in human dermal fibroblasts. Sci Rep 4:3843. doi:10.1038/srep03843
12. Meier JL, Yu A, Korf I et al (2012) Guiding the design of synthetic DNA-binding molecules with massively parallel sequencing. J Am Chem Soc 134:17814–17822. doi:10.1021/ja308888c
13. Kang JS, Meier JL, Dervan PB (2014) Design of sequence-specific DNA binding molecules for DNA methyltransferase inhibition. J Am Chem Soc 136:3687–3694. doi:10.1021/ja500211z
14. Han L, Pandian GN, Junetha S et al (2013) A synthetic small molecule for targeted transcriptional activation of germ cell genes in a human somatic cell. Angew Chem Int Ed 52:13410–13413
15. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res 37:e151. doi:10.1093/nar/gkp802
16. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659. doi:10.1093/bioinformatics/btr261
17. Workman CT, Yin Y, Corcoran DL et al (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. doi:10.1093/nar/gki439
18. Morinaga H, Bando T, Takagaki T et al (2011) Cysteine cyclic pyrrole-imidazole polyamide for sequence-specific recognition in the DNA minor groove. J Am Chem Soc 133:18924–18930. doi:10.1021/ja207440p
19. Minoshima M, Bando T, Sasaki S et al (2008) Pyrrole-imidazole hairpin polyamides with high affinity at 5′-CGCG-3′ DNA sequence; influence of cytosine methylation on binding. Nucleic Acids Res 36:2889–2894. doi:10.1093/nar/gkn116

Chapter 3
Genome-Wide Assessment of the Binding Effects of Artificial Transcriptional Activators by Utilizing the Power of High-Throughput Sequencing

Abstract One of the major goals in DNA-based personalized medicine is the development of sequence-specific small molecules to target the genome by means of synthetic biology; SAHA-PIPs belong to such class of small molecules. In a complex eukaryotic genome, the differential biological effects of SAHA-PIPs remain unclear. These questions can be addressed by directly identifying the binding regions of small molecules across the genome; however, it is a challenge to enrich specifically the small-molecule-bound DNA without chemical cross-linking. Here, we developed a method using high-throughput sequencing to map the binding area of non-cross-linked small molecules throughout the chromatinized human genome. Analysis of the sequenced data confirmed the presence of specific binding sites for SAHA-PIPs among the enriched sequence reads. Mapping the binding sites and enriched regions on the human genome clarifies the origin of the distinctive biological effects of SAHA-PIP. This approach will be useful for identifying the functionality of other small molecules on a large scale.

Keywords SAHA-PIP · Pull-down · Epigenetics · Small molecule · High-throughput sequencing

3.1 Introduction

The competency of each cell type to maintain its precise biological characteristics depends on the inherited differences in chromatin packaging named as nucleosomes. These preset arrangements in the genome enable the epigenome to control the fate of the cell by regulating a specific set of genes [1, 2]. Notably, changes in the epigenetic machinery can lead to cell plasticity, which triggers cellular reprogramming [3]. Although various histone marks are considered to be the gears of the epigenome, methylation and acetylation are the extensively studied histone modifications in governing chromatin dynamics [4]. Epigenetic alterations by these chromatin-modifying enzymes result in the alteration of gene expression. Given that most epigenetic marks are dynamic, some of these post translational histone

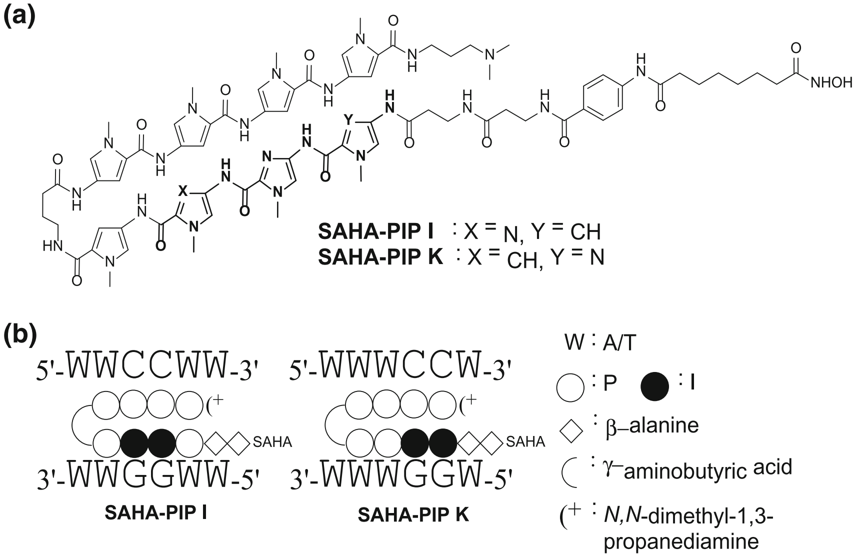
© Springer Nature Singapore Pte Ltd. 2018

A. Chandran, *Advancing Development of Synthetic Gene Regulators*,
Springer Theses, https://doi.org/10.1007/978-981-10-6547-7_3

modifications are reversible by enzymes [5]. Synthetic biology approaches aim at controlling these modifications by using sequence-specific artificial small molecules and is one of the key objectives in developing personalized nucleic acid targeted therapy.

The use of the versatile *N*-methylpyrrole (P) *N*-methylimidazole (I) synthetic polyamides (PIPs) in regulating sequence specific gene expression has been successful and these molecules are efficient enough to penetrate cell membranes even at nanomolar concentrations. PIPs belong to the category of DNA minor-groove binders that can recognize each of the Watson–Crick base pairs with a programmable canonical DNA binding rule. The predetermined DNA binding rule of PIP is that an antiparallel arrangement of I opposite P (I–P) recognizes a G–C base pair, P–I recognizes a C–G base pair, and P–P recognizes either T–A or A–T base pairs [6, 7].

Previously, we have reported the design and application of SAHA-PIPs as epigenetically active artificial transcriptional regulators. SAHA-PIP was synthesized by conjugating a histone deacetylase inhibitor SAHA (suberoyl anilide hydroxamic acid) with PIPs. These studies showed that SAHA-PIPs can specifically trigger efficient transcriptional activation of an epigenetically silent gene network in human dermal fibroblasts [8–11]. However, the actual binding sites of SAHA-PIPs in a broad context were unclear. Therefore, previously we used Bind-n-Seq methods to understand the sequence recognition property of the SAHA-PIPs [12]. Our investigation on the global gene expression of a chemical library containing 32 SAHA-PIPs in human dermal fibroblast cells showed that each of the conjugates can regulate a unique gene set, including some therapeutically important genes [13]. The underlying factors and characteristics of SAHA-PIPs that lead to regulation of a unique gene set remain unresolved. Previous studies with alkylating PIP suggested that the genomic regions occupied by histones can influence the binding and gene regulating effect of PIPs [14]. Beforehand, "Cross-linking of small molecules for isolation of chromatin" (COSMIC) was used to determine the high-affinity binding of PIP in the nucleus by conjugating a photo-cross-linker with PIP [15]. Our recent report on the use of alkylating-PIP was further extended to examine the binding conformation [16]. These studies established a platform upon which to investigate the PIP conjugate binding mechanism in chromatinized human genome. It remained an unresolved challenge to understand the binding behavior of SAHA-PIPs without any covalent cross-linking to DNA. In this report, we have developed a method to understand the binding behavior of SAHA-PIPs by combining micrococcal nuclease (MNase) digestion, affinity purification, and high-throughput sequencing. Our recent studies showed that SAHA-PIP I (Scheme 3.1) could precisely trigger the expression of essential pluripotency genes such as *OCT-3/4*, *NANOG*, and *DPPA4*, on the otherhand its structural counterpart, SAHA-PIP K (Scheme 3.1), activated a completely different set of genes related to germ-cell such as *PIWIL2*, *PIWIL4*, and *MOV10L1* in human dermal fibroblast cells [17, 18]. These observations prompted us to study the genomic binding occupancy of SAHA-PIPs I and K on the chromatinized human genome. The results of this study reveal the high-affinity binding sites and binding preferences of SAHA-PIPs across the complex human genome.



Scheme 3.1 **a** Structure of SAHA-PIP I and SAHA-PIP K. **b** SAHA-PIP I and SAHA-PIP K binding sequences based on PIP-DNA binding rule

3.2 Results and Discussion

As an artificial transcription activators, the biological effects of SAHA-PIP I and K mainly rely on their high-affinity binding preferences in the chromatinized genome. To characterize this phenomenon, we used our recent report [16] as a platform and developed a method shown in Fig. 3.1, with non-covalently binding small molecules in the extracted nucleus of live cells. For this study, we used SAHA-PIP I [SAHA-β-β-Py-Im-Im-Py-γ-Py-Py-Py-Py-β-(+)PEG12-Biotin; **3**] and SAHA-PIP K [SAHA-β-β-Im-Im-Py-Py-γ-Py-Py-Py-Py-β-(+)PEG12-Biotin; **4**], from our previously reported library of biotin-conjugated SAHA-PIPs [12] (Scheme 3.2). The compounds were synthesized as described in the previous report [12]. Our results unambiguously showed that neither the modification in the SAHA moiety nor attachment of biotin in SAHA-PIPs affected their binding specificity [12]. In breif, the nucleus was extracted from live human fibroblast BJ cells [neonatal foreskin (ATCC, USA)] and incubated with **3** and **4** (400 nM) separately, control experiments were performed without SAHA-PIPs. To avoid the dissociation of non-covalently bound **3** and **4**, and to obtain the PIP bound target DNA fragment, MNase digestion was performed [19] (see the materials and methods). After DNA fragmentation and nuclear protein digestion by proteinase K, PIP bound DNA was enriched and purified by using biotin-streptavidin chemistry based affinity purification. The extracted genomic DNA from the biological triplicates were used to

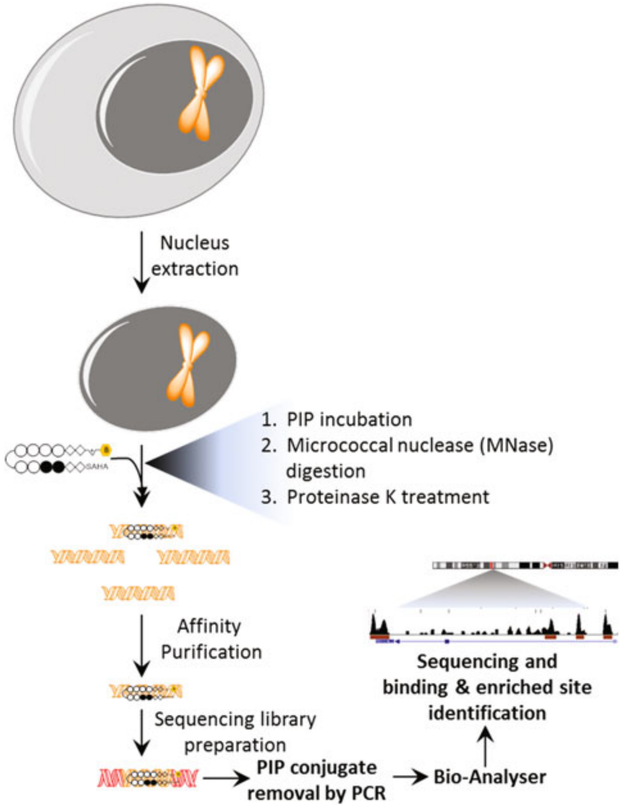
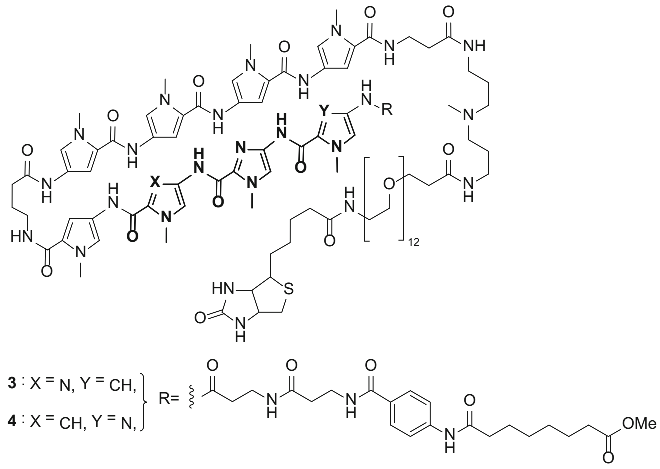


Fig. 3.1 MNase digestion and affinity purification-based high-throughput sequencing method pipeline using noncovalently binding PIP conjugates in the human nucleus from live cells



Scheme 3.2 Chemical structures of **3** and **4**

construct sequencing libraries and subjected to high-throughput sequencing, using standard sequencing methods for Ion PGM^{TM}/Proton^{TM}.

In breif, the nucleus was extracted from live human fibroblast BJ cells [neonatal foreskin (ATCC, USA)] and incubated with **3** and **4** (400 nM) separately, control experiments were performed without SAHA-PIPs. To avoid the dissociation of non-covalently bound **3** and **4**, and to obtain the PIP bound target DNA fragment, MNase digestion was performed [19] (see the materials and methods). After DNA fragmentation and nuclear protein digestion by proteinase K, PIP bound DNA was enriched and purified by using biotin-streptavidin chemistry based affinity purification. The extracted genomic DNA from the biological triplicates were used to construct sequencing libraries and subjected to high-throughput sequencing, using standard sequencing methods for Ion PGM^{TM}/Proton^{TM}.

High-quality sequencing reads were mapped with the human genome. To find the high-affinity binding sites of **3** and **4** over the control sequence reads (without any PIP treatment), uniquely mapped sequence data were randomly extracted by using subsampling Perl code and analyzed with a Bind-n-Seq pipeline and motif calling [20–22]. The identified high-affinity binding motifs for **3** and **4** are shown in Fig. 3.2 and the sequence details are in Tables 3.1, 3.2, 3.3, 3.4, 3.5 and 3.6. One of the highly enriched motifs shown in Fig. 3.2a for **3** obeyed well its canonical binding rule (5′-WWCCWW-3′), but the palindromic nature of the sequence could not allow the easy determination of its binding orientation. Figure 3.2b shows the high affinity binding sites of PIP conjugate **4** with significant enrichment scores; the obtained data revealed that **4** can bind to DNA in both the forward (N-terminal to C-terminal of PIP recognize 5′ to 3′ of DNA) and reverse orientation (N-terminal to C-terminal of PIP recognize 3′ to 5′ of DNA) with an enrichment ratio of 13.91 and 14.63, respectively. This type of binding was also observed in our previous report [12]. Comparative

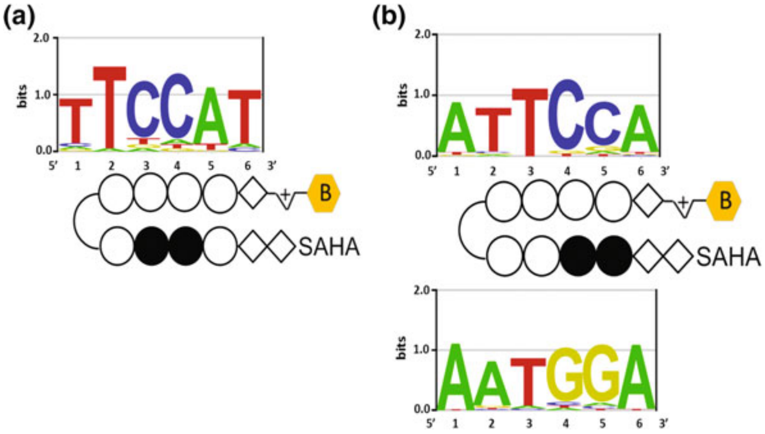


Fig. 3.2 Identified high-affinity binding motif of **a** **3** and **b** **4** (forward and reverse binding) in the human genomic enriched sequence

Table 3.1 Base composition for **3** enriched high-affinity binding motif (Fig. 3.2a)

Base/Position	1	2	3	4	5	6
A	101,542	74,463	95,607	121,875	1795,588	135,205
C	133,568	0	1660,957	1729,644	0	128,146
G	82,898	0	110,992	80,128	44,857	0
T	1723,257	1966,802	173,709	109,618	200,820	1777,914

Table 3.2 Base composition in percentage (%) for **3** enriched high-affinity binding motif (Fig. 3.2a)

Base/Position	1	2	3	4	5	6
A	4.97	3.64	4.68	5.97	87.96	6.62
C	6.54	0	81.37	84.73	0	6.27
G	4.06	0	5.43	3.92	2.19	0
T	84.42	96.35	8.50	5.37	9.84	87.10

Table 3.3 Base composition for **4** enriched high-affinity forward binding motif (Fig. 3.2b)

Base/Position	1	2	3	4	5	6
A	11,484,904	377,658	154,445	0	596,910	11,407,282
C	0	449,598	0	11,554,127	10,168,801	222,343
G	330,252	400,567	0	476,424	1186,363	325,987
T	619,941	11207,274	12,280,652	404,546	483,023	479,485

Table 3.4 Base composition in percentage (%) for **4** enriched high-affinity forward binding motif (Fig. 3.2b)

Base/Position	1	2	3	4	5	6
A	92.36	3.037	1.24	0	4.80	91.73
C	0	3.61	0	92.92	81.78	1.78
G	2.65	3.22	0	3.83	9.54	2.62
T	4.98	90.13	98.76	3.25	3.88	3.855

Table 3.5 Base composition for **4** enriched high-affinity reverse binding motif (Fig. 3.2b)

Base/Position	1	2	3	4	5	6
A	10,409,874	9781,071	646,082	392,937	521,859	10,299,390
C	0	0	340,220	659,086	435,721	0
G	0	481,514	0	9062,792	9343,911	0
T	148,100	295,389	9571,672	443,159	256,483	258,584

Table 3.6 Base composition in percentage (%) for **4** enriched high-affinity reverse binding motif (Fig. 3.2b)

Base/Position	1	2	3	4	5	6
A	98.60	92.64	6.11	3.72	4.94	97.55
C	0	0	3.22	6.24	4.13	0
G	0	4.56	0	85.84	88.50	0
T	1.40	2.80	90.66	4.20	2.430	2.45

assessments of SAHA-PIPs binding on chromatinized genomic DNA with a protein-free synthetic DNA library [12] showed that both experimentally derived high-affinity binding motifs follow the PIP binding rule. Although the side-by-side arrangement of P–I/I–P in PIP recognizes a unique base pair (CG/GC), the P–P arrangement shows a slight variation in recognizing A/T (W). This variation may be due to the primary and secondary preferential nucleotide distribution in the given sequence context. Interestingly their base recognition(W) strictly follows the canonical binding rule.

3.2.1 Revealing the Unique Gene Set Activation Mechanism of PIP Conjugates in the Human Genome

The epigenetically active form of **3** (SAHA-PIP I) (Scheme 3.1) was reported to activate an epigenetically silent pluripotency gene set in human fibroblast cells [17], whereas its structural counterpart epigenetically active form of **4** (SAHA-PIP K) (Scheme 3.1) triggered the activation of meiosis controlling PIWI pathway genes in somatic cells, which is involved in germ-cell generation [18]. When the obtained **3** and **4** enriched sequence read signal and enrichment peaks (peaks were identified with respect to control data) [23] were mapped, the results were noteworthy. The data for PIP conjugate **3** showed enrichment peaks at the promoter regions of SAHA-PIP I upregulated genes associated with pluripotency: *OCT3/4* (also known as *POU5F1*: POU domain, class 5, transcription factor 1) (Fig. 3.3), *DPPA4* (developmental pluripotency associated 4) (Extended Fig. 3.1a), and *EPCAM* (Extended Fig. 3.1b). In contrast, **4** did not show any significant level of enrichment at the SAHA-PIP I induced pluripotent genes. In line with this, enrichment peaks for **4** were observed in the promoter region and gene body of SAHA-PIP K upregulated genes involved in germ-cell generation such as *PIWIL4*, *PIWIL2*, and *TDRD9* (Fig. 3.3b, Extended Fig. 3.1c, d). Careful observation of *TDRD9* shown in Extended Fig. 3.1d demonstrated the enrichment of peaks in the gene body but not around the promoter, explaining the reason behind the mild effect of SAHA-PIP K on *TDRD9* promoter acetylation, and the minor change in mRNA expression noted in our previous report [18]. Interestingly, compound **3** did not show any specific significant enrichment on either *PIWIL4* or *TDRD9*; eventhough it displayed a mild

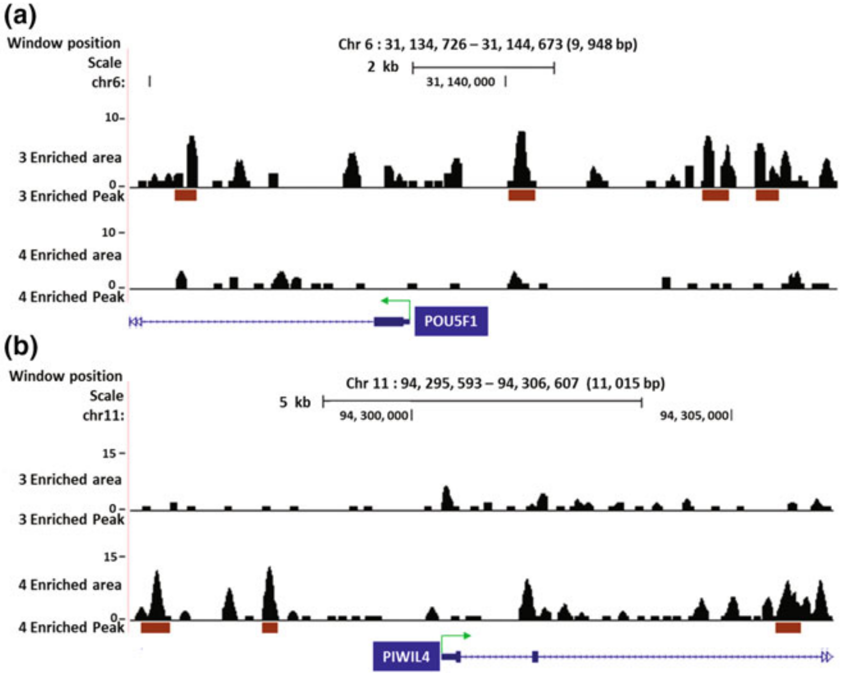


Fig. 3.3 Identified genomic regions of PIP conjugates **3** and **4** binding and enrichment in **a** *OCT3/4* (*POU5F1*) **b** *PIWIL4*

enhancement on *PIWIL2*, but it could not be comparable with the effective significant enrichment by conjugate **4**.

In contrast, neither **3** nor **4** showed any clearly enriched region on the housekeeping gene *GAPDH* (Extended Fig. 3.1e). Although the active forms of SAHA-PIP I and SAHA-PIP K have closely related recognition sites, their transcriptional activation network is distinctive to each other. Taken all together, our high-throughput sequencing analysis of **3** and **4** enriched genomic regions in chromatinized human genome allowed us to provide; the direct evidence of DNA minor-groove binding SAHA-PIP's differential gene activation mechanism in eukaryotic cells.

3.2.2 Identification of the Preferential Binding Region of PIP Conjugates in the Tightly Packed Heterochromatin Region

As noted in the previous report, histone-occupied chromatin regions play an important role in PIP binding preferences in genomic DNA [14]. To study the binding inclinations of SAHA-PIPs, we compared the genomic regions that were enriched with **3** and **4** with the MNase-seq data of nucleosomal positioning from ENCODE at UCSC [24–28]. Heterochromatin play a critical role in gene silencing because of the tighter packaging of DNA [29].

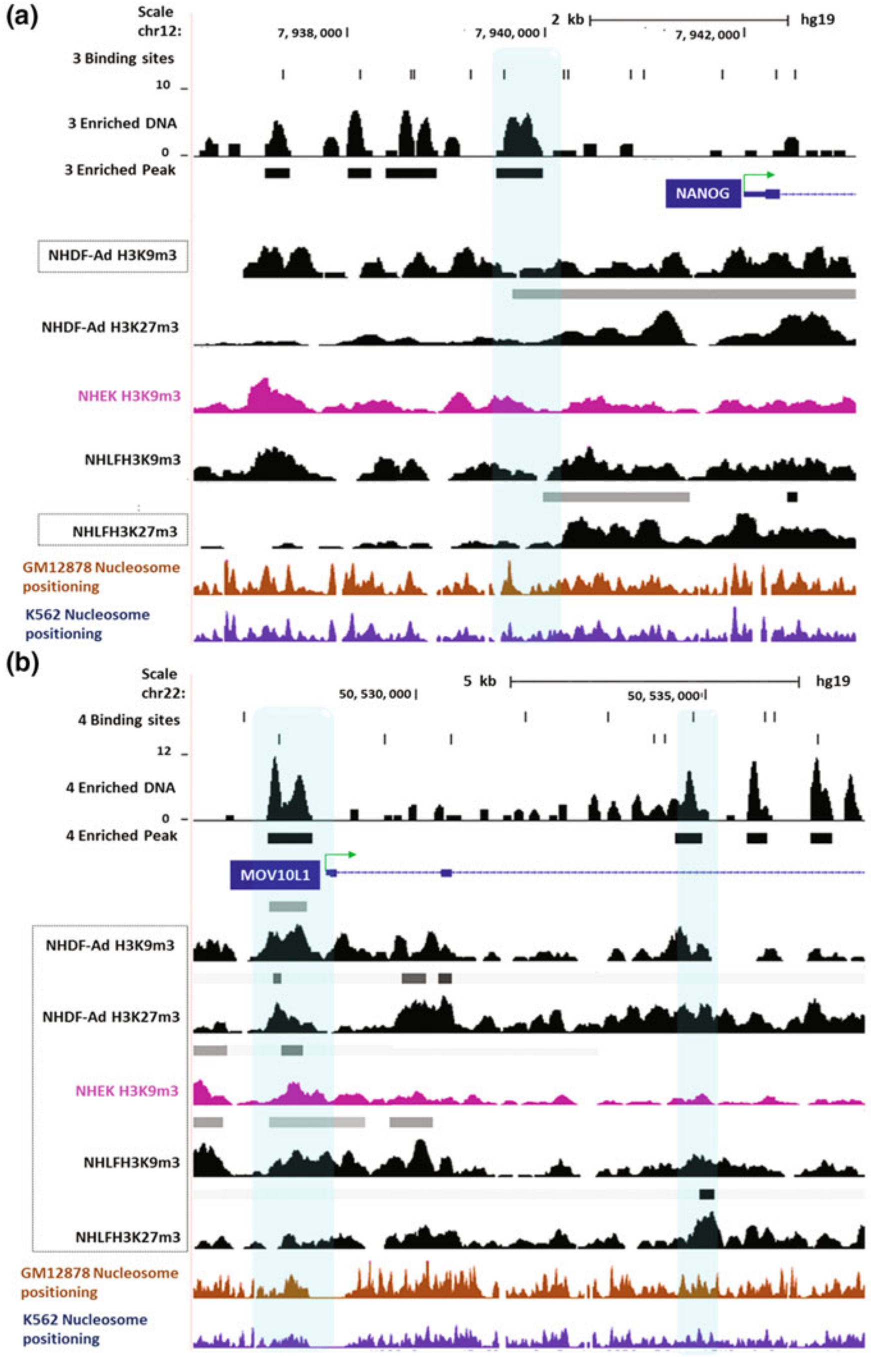
Covalent modifications of eukaryotic histone tails such as histone methylation are vital for heterochromatin formation; [29, 30] these modifications occur mainly through trimethylation at histone 3 lysine 9 (H3K9) [29, 31] and lysine 27 (H3K27) [31, 32]. We therefore analyzed and compared the H3K9me3 and H3K27me3 regions that were identified on the skin and lung fibroblast cell lines [NHEK (skin epidermal keratinocytes), NHDF-Ad (adult dermal fibroblasts skin), and NHLF (lung fibroblasts)] [27, 28], along with the nucleosomal positioning data.

Our comparative analysis showed that the **3** enriched genomic region with high-affinity binding site around the SAHA-PIP I upregulated *NANOG* [17] promoter is H3K27 trimethylated (Shown in Fig. 3.4a with blue shade) in the reported fibroblast cell lines and also falls under the nucleosomal region (MNase-seq data) (Fig. 3.4a, last two tracks). Similarly, SAHA-PIP K highly upregulated *MOV10L1* [17] promoter region and gene body is enriched in **4**, (also harboring its binding site) are trimethylated at both H3K9 and H3K27 (Fig. 3.4b, blue shade) in the reported fibroblast cell lines; these regions are also packed into nucleosomal units (MNase-seq data) (Fig. 3.4b, last two tracks).

These results clearly demonstrate that both **3** and **4** can also bind efficiently to the nucleosome within the possible heterochromatin-forming regions. This provides evidence on silent gene activation principle of SAHA-PIPs.

3.2.3 Discussion

The application of high-throughput sequencing in chemical biology has produced some notable outcomes [33]. Furthermore, the identification of genomic targets by using small molecules help us in understanding the drug effects and optimise the drug design [34]. The combination of ChIP-seq and Chem-seq approaches [35, 36] have been reported to characterise the protein-binding small molecules. Here, we report a method involving small molecules that form noncovalent interaction with DNA. We studied the high-affinity binding motif of such molecules and the corresponding enriched genomic regions. Annotation of **3** and **4** enriched region showed similar pattern of genomic region distribution (slightly varied in some genomic region) with high correlation among the experimentally identified peaks, actual binding sites and the deduced binding sites based on predetermined binding rule shown in Extended Fig. 3.2a, b. Comparative analysis of **3** and **4** provided evidence for the differential activation of the gene network by epigenetically active forms such as SAHA-PIP I and SAHA-PIP K. These results also suggest a binding mechanism for SAHA-PIPs in actual chromatinized genome. The following points can be noted in particular. (i) Small variations in the P–I arrangement in PIP can result in a large difference in genome-wide binding site recognition. (ii) The result


(a)
Scale
chr12:
2 kb
hg19
7, 938, 000
7, 940, 000
7, 942, 000
3 Binding sites
10
3 Enriched DNA
0
3 Enriched Peak
NANOG
NHDF-Ad H3K9m3
NHDF-Ad H3K27m3
NHEK H3K9m3
NHLFH3K9m3
NHLFH3K27m3
GM12878 Nucleosome positioning
K562 Nucleosome positioning
(b)
Scale
chr22:
5 kb
hg19
50, 530, 000
50, 535, 000
4 Binding sites
12
4 Enriched DNA
0
4 Enriched Peak
MOV10L1
NHDF-Ad H3K9m3
NHDF-Ad H3K27m3
NHEK H3K9m3
NHLFH3K9m3
NHLFH3K27m3
GM12878 Nucleosome positioning
K562 Nucleosome positioning

◂**Fig. 3.4** Binding and enrichment of PIP conjugates **3** and **4** in possible heterochromatin-forming regions of **a** *NANOG* promoter and **b** *MOV10L1* promoter. Possible heterochromatin regions were identified by mapping the MNase-Seq data of nucleosomal positioning with histone trimethylated regions of histone 3 lysine 9 and lysine 27 (Modified histone regions with PIP binding and enrichment site is shown in box)

that **3** and **4** bind on the nucleosomal region with histone marks may explain the epigenetic activation by SAHA-PIP of the silenced gene network. Given that this method used non-cross-linking or noncovalent binding-based affinity purification, the approach is expected to be widely applicable to the study of genomic effects of other DNA-binding small molecules.

3.3 Materials and Methods

3.3.1 Nucleus Extraction and PIP Conjugate Incubation

75–80% confluent human fibroblast BJ cells (neonatal foreskin (ATCC, USA)) were used to extract nuclei [19]. In brief, P6 of BJ-HFF cells were grown in Dulbecco's modified eagle medium (DMEM, Nacalai Tesque, Japan) supplemented with 10% FBS (FBS, Japan Serum) in a humidified atmosphere of 5% CO_2 at 37 °C. At 75–80% confluency, the cells were isolated by PBS washing followed by 3–4 min tripsinization. Isolated 2×10^6 cells were washed twice with ice cold PBS. The cell membrane was digested with 5 ml of ice cold NP-40 lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 0.5% Nonidet P-40, 0.15 mM spermine, 0.5 mM spermidine and 0.1× protease inhibitor cocktail) for 5 min on ice. Loosen pellet of cell nuclei were obtained by centrifugation and dissolved in binding and resuspension buffer [10 mM Tris-Cl (pH 8.0), 5 mM $MgCl_2$, 1 mM DTT, 0.3 M KCl, 0.3× protease inhibitor cocktail and 10% glycerol]. **3** and **4** were dissolved in DMSO and incubated (400 nM of **3** and **4,** 0.1% final concentration of DMSO) separately with the isolated nuclei about 16–18 h at 4 °C. For the control experiments only DMSO (PIPs were dissolved in DMSO) was used without PIP conjugates. Concentrations were used according to the previous reports [15, 16].

3.3.2 MNase Digestion

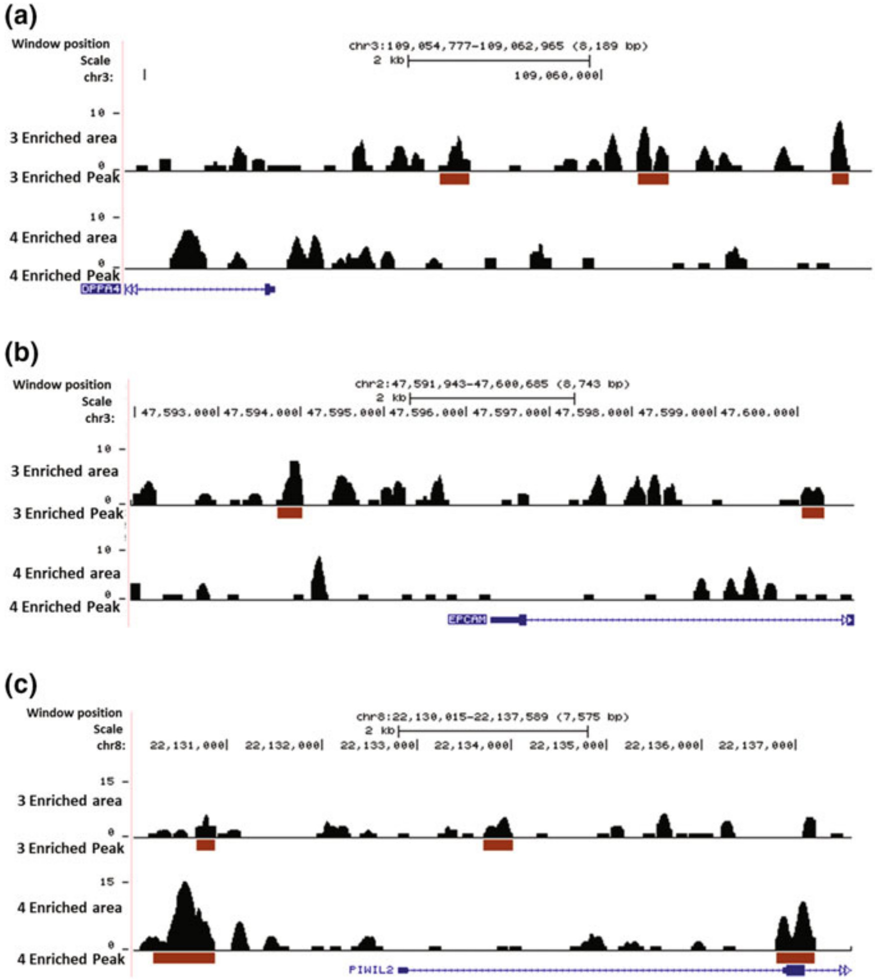
After PIP derivatives incubation, micrococcal nuclease (MNase) buffer [19] [10 mM Tris-HCl (pH 7.4), 15 mM NaCl, 60 mM KCl, 0.15 mM spermine, 0.5 mM spermidine and 0.1× protease inhibitor cocktail] pre-washed nuclei was subjected to MNase (Takara, Japan) digestion at 37 °C for 30 min. Sonication shearing may affect the non-covalently bound PIP, so we used MNase digestion. To avoid protein hindrance during affinity purification, the MNase digested nucleosomes were treated with proteinase K.

3.3.3 *Affinity Purification*

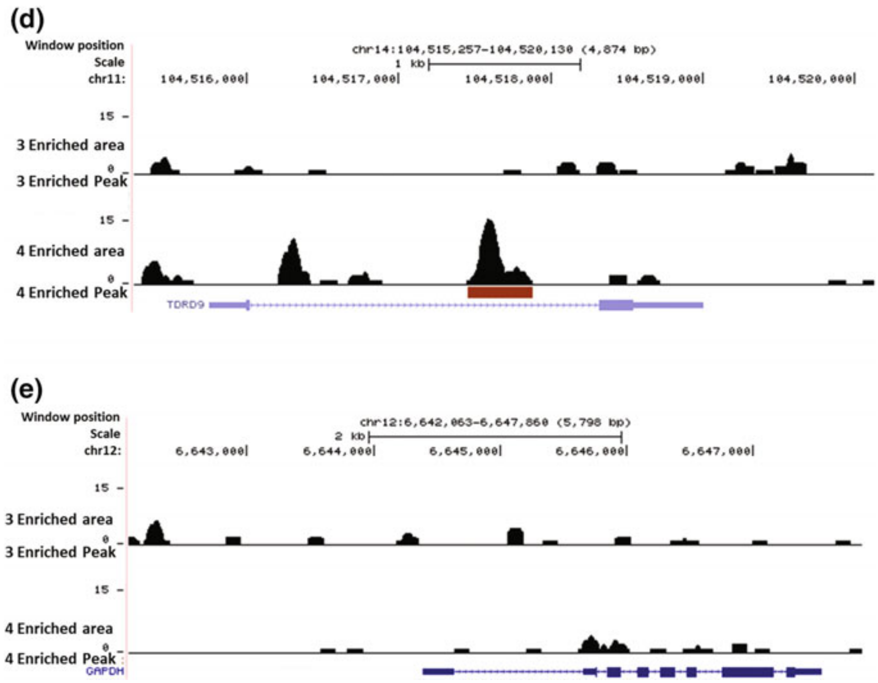
Protein digested PIP bound DNA containing suspension was mixed with modified COSMIC buffer [15] [20 mM Tris-Cl (pH 8.1), 2 mM EDTA, 150 mM NaCl, 0.1× protease inhibitor cocktail, 1% Triton-X100, and 0.1% SDS]. 10% of the sample was saved as input DNA. Dynabeads MyOne C1 preparation: Streptavidin-coated magnetic beads (Dynabeads MyOne C1, Life Technologies, USA) was washed twice with modified COSMIC buffer after removing the suspension solution. 0.5 mg per sample of beads was incubated with samples at 4 °C for 16 h by way of rotor mixing. After incubation, samples were subjected to sequential of washing with buffer 1, washing Buffer 2 [10 mM Tris-Cl (pH 8.0), 250 mM LiCl, 1 mM EDTA, 0.5% NP40], washing Buffer 3 [10 mM Tris-Cl (pH 7.5), 1 mM EDTA, 0.1% NP-40], and finally with TE. Washed samples were resuspended in elution buffer-I [10 mM Tris-HCl (pH 7.6), 0.4 mM EDTA and 100 mM KOH] and DNA was eluted by heating at 90 °C for 30 min. The unrecovered DNA with the beads were subjected to second elution using elution buffer-II (2% SDS, 100 mM $NaHCO_3$ and 3 mM biotin) with the heating of 65 °C for 8–12 h. The eluted DNA samples were purified with the QIAquick PCR purification Kit (Qiagen, CA, USA) and quantified using Nanodrop.

3.3.4 *High-Throughput Sequencing Library Construction and Sequencing*

Optimum amount of purified DNA was pooled (To get sufficient amount of DNA) from the biological triplicates and sequencing libraries were prepared by using Ion Xpress^{TM} Plus gDNA Fragment Library preparation reagents and protocols (Life techologies, USA) as per the instruction. Adapter ligated DNA was amplified and purified. The purified libraries were analyzed with Agilent DNA High sensitivity BioAnalyzer kit (Agilent technologies, USA). The sequencing was carried out, as (1) template preparation using Ion PGM^{TM} template OT2 200 v2 kit and Ion PI^{TM} template OT2 200 kit in Ion one touch2 system. (2) The templates enrichment on Ion one touch ES. (3) Sequencing the enriched libraries with Ion PGM^{TM}/Ion Proton^{TM} sequencer using Ion PGM^{TM} sequencing 200 kit v2/Ion PI^{TM} Sequencing 200 kit v3 and 318 v2 chip/Ion PI chip according to the manufacturer's guidelines. Single read sequencing was performed with 260–300 flow, 28–31 million post filtered reads per library was produced. Ion torrent suit was used for the preliminary data analysis. TMAP 4.4.2 was used for aligning the good quality reads with human genome, To identify high-affinity binding motif a random selected 10–15% of uniquely mapped reads were used (Normalization performed with control data). We followed our previous analysis pipeline for motif calling [12, 16, 20–22]. The aligned data was further analysed for enriched peak calling using standard ChIP-seq analysis program MACS 1.4.2. [23] (SAHA PIP enriched reads as a treat and control (without SAHA-PIP) reads as control) MACS enriched signals and peak



Extended Fig. 3.1 Identified genomic regions of PIP conjugates **3** and **4** binding and enrichment in **a** *DPPA4* **b** *EPCAM* **c** *PIWIL2* **d** *TDRD9* **e** *GAPDH*



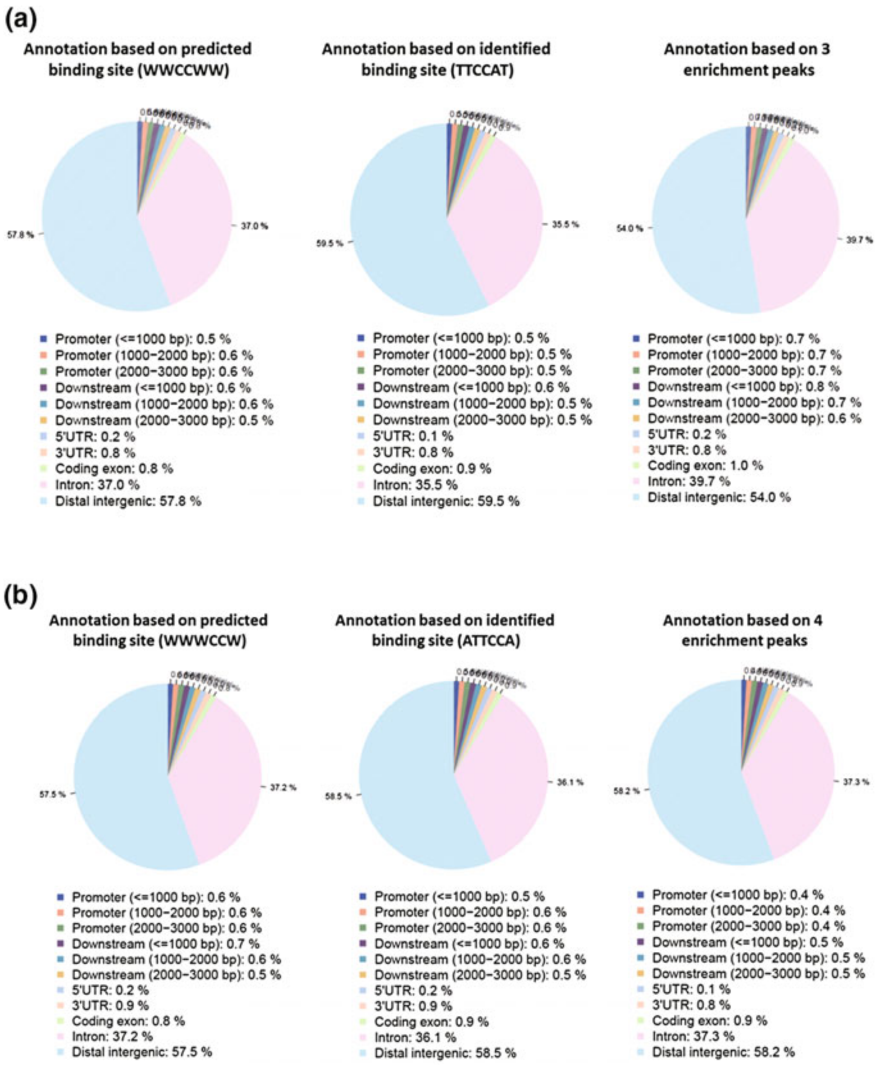
Extended Fig. 3.1 (continued)

regions were visualized on UCSC genome browser to identify uniquely enriched regions between **3** and **4** [24–26]. To analyse the influence of chromatin bound DNA on PIP conjugates binding, we compared various histone marks and nucleosomal positioning ENCODE data with our obtained data [27, 28].

Appendix

Extended Fig. 3.1.

Extended Fig. 3.2.



Extended Fig. 3.2 Comparison of genomic binding and enriched region with binding rule based binding site. **a** For compound **3**. **b** For compound **4**

References

1. Takahashi K, Tanabe K, Ohnuki M et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 107:861–872. doi:10.1016/j.cell.2007.11.019
2. Mohn F, Schübeler D (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends Genet 25:129–136
3. Pachaiyappan B, Woster PM (2014) Design of small molecule epigenetic modulators. Bioorg Med Chem Lett 24:21–32

4. Arrowsmith CH, Bountra C, Fish PV et al (2012) Epigenetic protein families: a new frontier for drug discovery. Nat Rev Drug Discov 11:384–400. doi:10.1038/nrd3674
5. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705
6. Carlson CD, Warren CL, Hauschild KE et al (2010) Specificity landscapes of DNA binding molecules elucidate biological function. Proc Natl Acad Sci U S A 107:4544–4549. doi:10.1073/pnas.0914023107
7. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. Curr Opin Struct Biol 13:284–299
8. Ohtsuki A, Kimura MT, Minoshima M et al (2009) Synthesis and properties of PI polyamide-SAHA conjugate. Tetrahedron Lett 50:7288–7292. doi:10.1016/j.tetlet.2009.10.034
9. Pandian GN, Shinohara KI, Ohtsuki A et al (2011) Synthetic small molecules for epigenetic activation of pluripotency genes in mouse embryonic fibroblasts. ChemBioChem 12:2822–2828. doi:10.1002/cbic.201100597
10. Pandian GN, Ohtsuki A, Bando T et al (2012) Development of programmable small DNA-binding molecules with epigenetic activity for induction of core pluripotency genes. Bioorg Med Chem 20:2656–2660. doi:10.1016/j.bmc.2012.02.032
11. Pandian GN, Nakano Y, Sato S et al (2012) A synthetic small molecule for rapid induction of multiple pluripotency genes in mouse embryonic fibroblasts. Sci Rep 2:1–8. doi:10.1038/srep00544
12. Anandhakumar C, Li Y, Kizaki S et al (2014) Next-generation sequencing studies guide the design of pyrrole-imidazole polyamides with improved binding specificity by the addition of β-alanine. ChemBioChem 15:2647–2651. doi:10.1002/cbic.201402497
13. Pandian GN, Taniguchi J, Junetha S et al (2014) Distinct DNA-based epigenetic switches trigger transcriptional activation of silent genes in human dermal fibroblasts. Sci Rep 4:3843. doi:10.1038/srep03843
14. Jespersen C, Soragni E, James Chou C et al (2012) Chromatin structure determines accessibility of a hairpin polyamide-chlorambucil conjugate at histone H4 genes in pancreatic cancer cells. Bioorg Med Chem Lett 22:4068–4071. doi:10.1016/j.bmcl.2012.04.090
15. Erwin GS, Bhimsaria D, Eguchi A, Ansari AZ (2014) Mapping polyamide-DNA interactions in human cells reveals a new design strategy for effective targeting of genomic sites. Angew Chem Int Ed 53:10124–10128. doi:10.1002/anie.201405497
16. Chandran A, Syed J, Taylor RD et al (2016) Deciphering the genomic targets of alkylating polyamide conjugates using high-throughput sequencing. Nucleic Acids Res 44:4014–4024. doi:10.1093/nar/gkw283
17. Pandian GN, Sato S, Anandhakumar C et al (2014) Identification of a small molecule that turn 'ON' the pluripotency gene circuitry in human fibroblasts. ACS Chem Biol 141024155048006. doi:10.1021/cb500724t
18. Han L, Pandian GN, Junetha S et al (2013) A synthetic small molecule for targeted transcriptional activation of germ cell genes in a human somatic cell. Angew Chem Int Ed 52:13410–13413
19. Published in association with Cold Spring Harbor Laboratory Press (2005) Micrococcal nuclease–Southern blot assay. Nat Methods 2:719–720
20. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res 37:e151. doi:10.1093/nar/gkp802
21. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659. doi:10.1093/bioinformatics/btr261
22. Workman CT, Yin Y, Corcoran DL et al (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. doi:10.1093/nar/gki439
23. Feng J, Liu T, Qin B et al (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728–1740. doi:10.1038/nprot.2012.101
24. Kent WJ, Sugnet CW, Furey TS et al (2002) The human genome browser at UCSC. Genome Res 12:996–1006. doi:10.1101/gr.229102

25. Karolchik D, Barber GP, Casper J et al (2014) The UCSC genome browser database: 2014 update. Nucleic Acids Res. doi:10.1093/nar/gkt1168
26. Rosenbloom KR, Armstrong J, Barber GP et al (2015) The UCSC genome browser database: 2015 update. Nucleic Acids Res 43:D670–D681. doi:10.1093/nar/gku1177
27. Consortium EP, Dunham I, Kundaje A et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247
28. Ernst J, Kheradpour P, Mikkelsen TS et al (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49. doi:10.1038/nature09906
29. Grewal SIS, Rice JC (2004) Regulation of heterochromatin by histone methylation and small RNAs. Curr Opin Cell Biol 16:230–238
30. Martin C, Zhang Y (2005) The diverse functions of histone lysine methylation. Nat Rev Mol Cell Biol 6:838–849. doi:10.1038/nrm1761
31. Peters AHFM, Kubicek S, Mechtler K et al (2003) Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell 12:1577–1589. doi:10.1016/S1097-2765(03)00477-5
32. Santenard A, Ziegler-Birling C, Koch M et al (2010) Heterochromatin formation in the mouse embryo requires critical residues of the histone variant H3.3. Nat Cell Biol 12:853–862. doi:10.1038/ncb2089
33. Anandhakumar C, Kizaki S, Bando T et al (2015) Advancing small-molecule-based chemical biology with next-generation sequencing technologies. ChemBioChem 16:20–38
34. Rodriguez R, Miller KM (2014) Unravelling the genomic targets of small molecules using high-throughput sequencing. Nat Rev Genet 15:783–796
35. Anders L, Guenther MG, Qi J et al (2014) Genome-wide localization of small molecules. Nat Biotechnol 32:92–96. doi:10.1038/nbt.2776
36. Jin C, Yang L, Xie M et al (2014) Chem-seq permits identification of genomic targets of drugs against androgen receptor regulation selected by functional phenotypic screens. Proc Natl Acad Sci U S A 111:9235–9240. doi:10.1073/pnas.1404303111

Chapter 4
Deciphering the Genomic Targets of Alkylating Polyamide Conjugates Using High-Throughput Sequencing

Abstract Chemically engineered small molecules targeting specific genomic sequences play an important role in drug development research. Pyrrole-imidazole polyamides (PIPs) are a group of molecules that can bind to the DNA minor-groove and can be engineered to target specific sequences. Their biological effects rely primarily on their selective DNA binding. However, the binding mechanism of PIPs at the chromatinized genome level is poorly understood. Herein, we report a method using high-throughput sequencing to identify the DNA-alkylating sites of PIP-indole-seco-CBI conjugates. High-throughput sequencing analysis of conjugate **2** showed highly similar DNA-alkylating sites on synthetic oligos (histone-free DNA) and on human genomes (chromatinized DNA context). To our knowledge, this is the first report identifying alkylation sites across genomic DNA by alkylating PIP conjugates using high-throughput sequencing.

Keywords DNA binder · Affinity purification · High-throughput sequencing · ERBB2 · Alkylation

4.1 Introduction

N-Methylpyrrole (P)—*N*-methylimidazole (I) polyamides (PIPs) are a class of programmable minor-groove binders that follow a canonical DNA recognition rule. The recognition rules are that an antiparallel arrangement of P opposite I (P-I) recognizes a C-G base pair; I-P recognizes a G-C base pair, and P-P recognizes T-A or A-T base pairs [1]. Several studies have investigated the binding specificity of these programmable PIPs [2–7]. However, there is a bias toward PIPs binding chromatinized DNA. PIPs have the ability to penetrate the cell membrane and show an excellent DNA-binding efficiency, even in nanomolar concentrations [8], thus hindering the binding of transcription factors to their respective DNA sequences.

The normal transcriptional machinery sometimes becomes dysfunctional because of alteration of DNA bases causing variation in gene expression and development of disease. Such fluctuations in gene regulation, largely dictated by modifications such

© Springer Nature Singapore Pte Ltd. 2018

A. Chandran, *Advancing Development of Synthetic Gene Regulators*,
Springer Theses, https://doi.org/10.1007/978-981-10-6547-7_4

as DNA alkylation and methylation, are caused by various factors in day-to-day life [9, 10]. DNA damage induced by alkylating agents can modify the genetic code, resulting in faulty protein synthesis [9, 11] that can cause abrupt cell-cycle arrest or apoptosis [12]. This makes alkylating agents attractive as antitumor drugs. To date, many DNA-alkylating agents have been reported to exhibit anticancer activity toward a variety of leukemias and solid tumors [13]. One of the major disadvantages of these agents is their nonselective DNA alkylation. Driving the alkylating agents toward tumor-specific target sequences in the human genome is a promising approach to advancing their efficacy as anticancer agents. Collectively, we developed various sequence-specific alkylating agents by coupling sequence-specific PIPs with alkylating moieties [14]. Among the coupling linkers, the indole linker extends to 2 bases and its N-terminal sequence selectivity is achieved by the hydrogen bond of the amide group of indole with O_2 of C or T, or with N3 of adenine (A) [15, 16]. Conjugating the alkylating moiety *seco*-CBI to a PIP can produce a covalent adduct with N3 of A within a predetermined sequence (Fig. 4.1a, c). A PIP-indole-*seco*-CBI conjugate with unique sequence recognition was tested for antitumor activity by selective silencing of tumor-inducing genes [17]. In this study, we have synthesized two PIP-indole-*seco*-CBI conjugates [**1** with a symmetrical binding site of 5′-WGGCCA-3′ and **2** with asymmetrical binding site 5′-WGGWCA-3′ (Scheme 4.1)] to investigate their DNA-alkylating sites.

Previous studies have employed high-resolution denaturing polyacrylamide gel electrophoresis to test the sequence-specific DNA-alkylating efficiency of PIP-*seco*-CBI conjugates on pre-determined 200–300 bps template sequences. Under high-temperature conditions, alkylated sequences at the PIP-binding sites are cleaved, generating patterns of DNA fragments, which can be analyzed further [14–16]. However, the selective DNA-alkylating ability of PIP-*seco*-CBI conjugates in a broad genome space using high-throughput sequencing has not yet been explored. Here, we describe the experimental design and data analysis methods to address this investigation.

Although PIP conjugates have many predicted binding sites across the human genome, only a small number of these binding sites play a significant role in gene regulation. This phenomenon is clearly observed in our previous report, where the whole genome expression analysis of a library of SAHA-PIP conjugates showed that individual compounds can trigger a distinctive set of genes in human dermal fibroblast (HDF) cells [18]. The complex organization of chromatin packaging in the nucleus may be a critical factor for the PIP-binding preferences along the genome [19–21]. A recent report by the Ansari group [22] has initiated the preliminary effort to map the PIP-binding sites in the human genome by developing an approach called “crosslinking of small molecules for isolation of chromatin” or COSMIC, thereby directing the evolution in PIP design strategy for effectively targeted gene regulation. Data provided by ChIP-seq regarding the genome-wide mapping of key transcription factors and regulatory element-binding sites are helpful in the derivation of transcriptional regulation models that govern normal and diseased cell states [23, 24]. At this juncture, we have employed cost-efficient semiconductor-sequencing technology to study the affinity purification-based

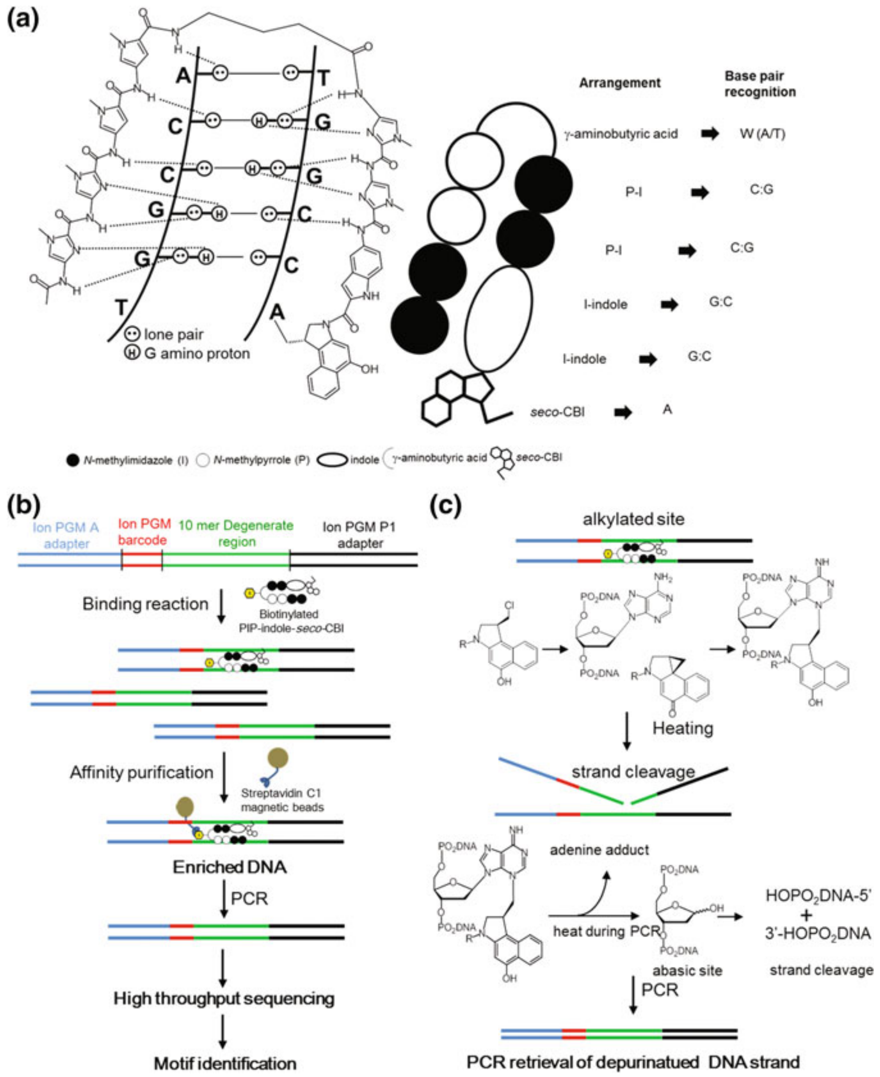
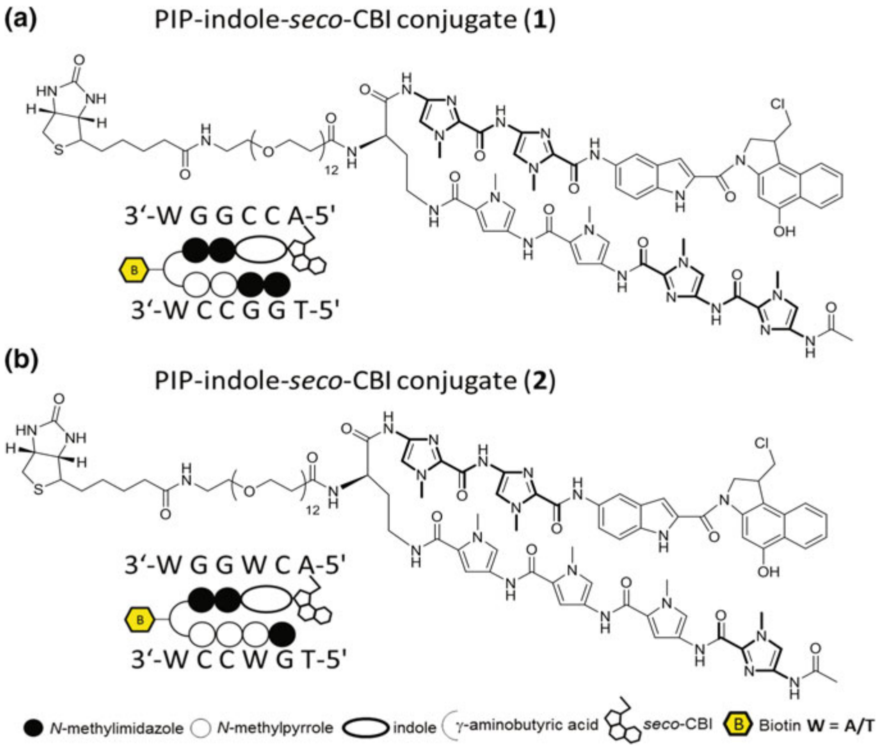


Fig. 4.1 PIP conjugate-binding mode and Bind-n-Seq. **a** Recognition of DNA minor groove by PIP-indole-*seco*-CBI conjugates. **b** Workflow of Bind-n-Seq analysis with PIP-indole-*seco*-CBI conjugate. **c** PCR amplification-based depurinated strand retrieval and chemical reaction of DNA N3 alkylation by *seco*-CBI through the formation of CBI and heat treatment

high-throughput sequencing of PIP-indole-*seco*-CBI conjugate-enriched human genomic regions. The present study will enable us to map the DNA-binding of small molecule DNA-alkylating sites all along the chromatin-packed genome utilizing high-throughput sequencing, which may provide a more detailed understanding of the mechanism of gene regulation by PIP conjugates.



Scheme 4.1 Chemical structures and representation of PIP-indole-*seco*-CBI conjugate. **a** **1** (Ac-I-I-P-P-*(R)*$^{\mathrm{NH-PEG12-Biotin}}$γ-I-I-Indole-*seco*-CBI). **b** **2** (Ac-I-P-P-P-*(R)*$^{\mathrm{NH-PEG12-Biotin}}$γ-I-I-Indole-*seco*-CBI). P = *N*-methylpyrrole and I = *N*-methylimidazole

4.2 Results and Discussion

*4.2.1 Bind-n-Seq with PIP-Indole-*Seco*-CBI Conjugate 1*

High-resolution denaturing polyacrylamide gel electrophoresis from our previous report showed multiple DNA-alkylating sites for the PIP-indole-CBI conjugate (B) [16]. These results are biased towards identifying primary DNA-alkylating sites. To address this kind of issue, we synthesized a biotin-conjugated PIP-indole-*seco*-CBI conjugate **1** (synthetic procedure is given in Materials and methods) and examined **1** for Bind-n-Seq (Fig. 4.1b, c) [25–27] (experimental procedure is given in the Materials and methods). The method is high-throughput sequencing-based, which may allow unbiased primary binding motif identification. *Seco*-CBI is readily converted to its cyclopropyl form CBI, and reacts with its DNA-reactive site [28]. The CBI could possibly depurinate the DNA alkylation site during heat elution of enriched DNA in Bind-n-Seq. Therefore, we performed a polymerase chain reaction (PCR) with sequencing adapter-specific primers to retrieve the damaged strand (Fig. 4.1c). The purified enriched DNA was subjected to high-throughput sequencing. High quality uniquely randomized sequence reads were analyzed using the Bind-n-Seq analysis method [25] to obtain the high-affinity DNA-alkylating site of **1**. The DNA showed a highly enriched “*k*-mer” ($k = 6$) binding site. The enriched motif with a 24.11-fold enrichment defined the DNA-alkylating sites of **1** and matches the PIP canonical binding rule. A graphical representation of the identified high-affinity motif for **1** is shown in Fig. 4.2a; and its corresponding highly enriched sequences are given in Extended Table 4.1. Extended Table 4.1 also shows the other potential binding sites of **1** at ranks 3 and 6 (highlighted in green). Although the recognition site of **1** followed the PIP canonical binding rule, the alkylation site of **1** (6th position from the 5′ end of the motif) was found to be W (A or T) instead of the expected base, A, because of the symmetrical nature of its binding. The results of the Bind-n-Seq analysis of **1** are vital because they show that this type of symmetrical PIP-indole-*seco*-CBI conjugate has the capability to bind with both strands of DNA in a forward-binding orientation (N-terminal to C-terminal of PIP binding with 5′ to 3′ of DNA) as shown in Fig. 4.2a, b.

When the binding of **1** is largely on symmetrical sequences, the acquired motif is acceptable with respect to the PIP-binding rule, but it is difficult to identify the precise alkylation site of CBI in **1** because of its symmetrical nature.

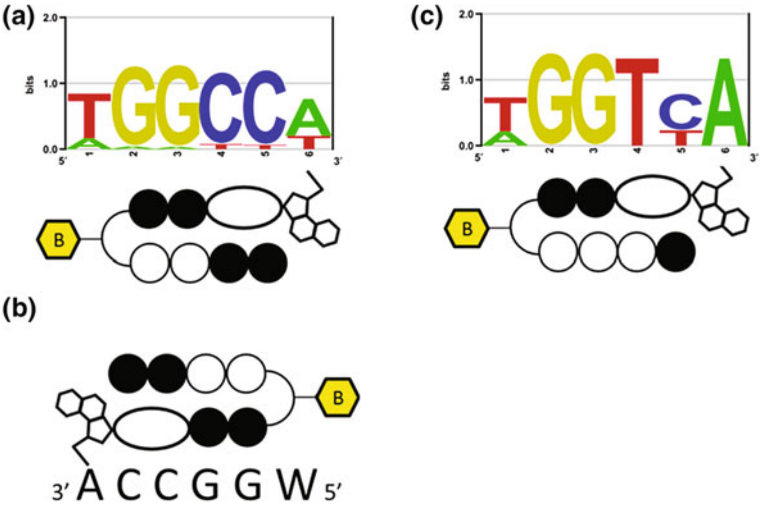


Fig. 4.2 PIP-indole-*seco*-CBI conjugate DNA-alkylating motifs. **a** Bind-n-Seq analysis identified high-affinity DNA-alkylating motif for 1 **b** 1 possible binding in its complementary recognition sequence. **c** Bind-n-Seq analysis identified a high-affinity DNA-alkylating motif for 2

4.2.2 *Bind-n-Seq with PIP-Indole-*Seco*-CBI Conjugate 2*

The investigation of Bind-n-Seq results for **1** showed the significance of PIP binding to its target site, but could not elucidate the CBI alkylation site. Taken together, our results for compound **1** motivated us to inspect the exact CBI alkylation site within a PIP conjugate. Therefore, we synthesized another asymmetrical PIP-indole-*seco*-CBI conjugate **2**. We designed **2** by replacing I with P in **1**, to obtain asymmetrical binding (synthetic procedure is given in the SI). Accordingly, **2** was subjected to Bind-n-Seq analysis. Analysis of the sequence reads obtained from **2** enriched DNA is shown in Extended Table 4.2. The high-affinity DNA-alkylating motif of **2** (Fig. 4.2c) was derived from a 7.99- and 7.36-fold enrichment (corresponding highly enriched sequences are given in Extended Table 4.2). The motif obtained follows the binding rule and specific A alkylation site for CBI in **2**. These findings support our earlier report of specific A alkylation by PIP-CBI conjugates within its recognition sequence [29].

4.2.3 *Bind-n-Seq with PIP-Conjugate 5*

To confirm the significance of the enriched DNA-alkylating motif of PIP-indole-*seco*-CBI conjugates and PCR repossession of the alkylated DNA strand, we performed a Bind-n-Seq experiment (21-mer randomized region) with our previously reported [30] biotinylated PIP-Conjugate **5** (designed to target 8 bp). Bind-n-Seq data was analyzed for the 8 and 9 bp motif windows ($k = 8$ and $k = 9$) to obtain the fold enrichment based on the control experiment (experiment without PIP-conjugate **5**). High-scoring motif hits (Extended Fig. 4.2) clearly demonstrated that there is no sequence selectivity at the 9th position of the motif (from the 5′ end of the motif). By contrast, PIP conjugate **2** showed a distinct (A) adenine-specific alkylation at the targeted 6th position from the 5′ end of the motif. This base specific alkylation site identification could be possible only when there is recovery of previously damaged alkylated DNA. These results demonstrated that successful retrieval of a damaged DNA strand could be possible using PCR reaction.

4.2.4 *Identification of PIP-Indole-*Seco*-CBI Conjugate 2 High Affinity DNA-Alkylating Site in Chromatinized Human Genome*

We then sought to extend the high-throughput sequencing approach to examine how chromatinized genomic architecture can impact the binding of **2** across the human genome in nuclei isolated from live cells. Consistent with this approach, we developed a method (Fig. 4.3) based on the COSMIC approach [22] that includes

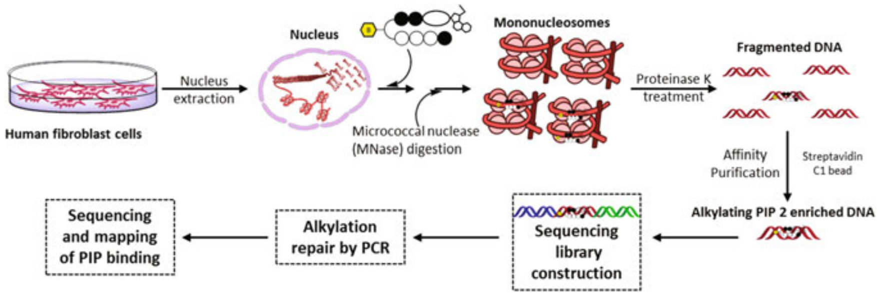
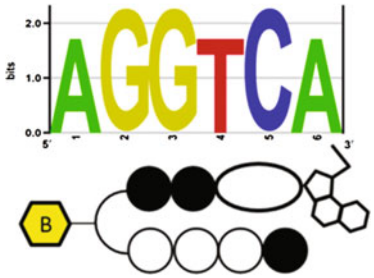


Fig. 4.3 Work flow for affinity purification-based high-throughput sequencing using PIP-indole-*seco*-CBI conjugates in the nucleus of human cells

affinity purification-based high-throughput sequencing of human genomic regions enriched with **2** without any photo-crosslinking procedure. We employed this method to elucidate the binding preferences of **2** in the biologically dynamic, histone-packed chromatinized surroundings of the nuclei (experimental procedure is described in the Materials and methods). The qualified sequencing libraries were subjected to sequencing. The processed sequence data was mapped along the human genome. To comprehensively determine the DNA-alkylating site (primary motif), we performed motif detection using a Bind-n-Seq data analysis pipeline. Among the enriched sequences (normalized with the control experimental data), the high-scoring motif hit (rank 1 in Extended Table 4.3) was unique DNA-alkylation site of **2** (Fig. 4.4) (enriched sequence details are given in Extended Table 4.3). Interestingly, the identified primary motifs of **2** propose that its sequence-specific DNA-alkylation of base A remains highly similar in both complex chromatinized human genomic DNA and randomized oligomer-based Bind-n-Seq analyses.

To compare the sensitivity of **2** enrichment, we have conducted MACS peak calling and the cross-correlation of identified DNA-alkylating motif with the obtained post-filtered genomic sequence data (25–30 million). Then we plotted (Fig. 4.5b) the identified motif (5′-AGGTCA-3′) on the genome-wide enriched peaks with the window of 300 bp [−150 to 150 bp from the center (0) of the peak]. This showed the greatest precision of distribution frequency with a **2** predicted

Fig. 4.4 Identified high-affinity DNA-alkylating motif of 2 in human genomic enriched sequence



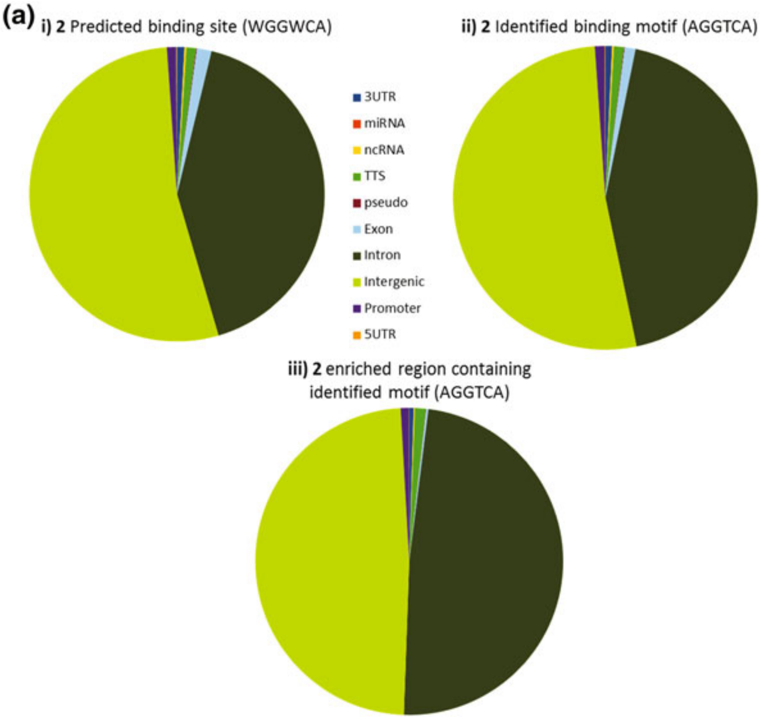


Fig. 4.5 Genomic enrichment and DNA-alkylating site distribution of PIP-indole-*seco*-CBI conjugate **2**: **a** (i) Predicted binding sites based on the canonical binding rule, (ii) experimentally identified high-affinity DNA-alkylating motif in human genome, (iii) affinity purification based sequencing enriched regions containing experimentally identified high-affinity DNA-alkylating motif. **b 2** related possible binding sites (predicted, experimentally identified, possible recognition of T by the γ-turn in the conjugate, one bp mismatch, two bp mismatch, and alkylation site mismatch) genome-wide distribution frequencies in the peak enriched region with the MACS peak window of 300 bp [−150 to 150 bp from the center (0) of the enriched peak]. **c 2** enriched region distribution on broad classes of chromatin states [six classes of chromatin states such as promoter (active, weak, and poised), enhancer (strong and weak), insulator, transcribed (strongly (Txn transition and Txn elongation) and weakly transcribed regions), repressed and inactive states (heterochromatin)] (chromatin stated were organized based on the ENCODE-ChromHMM data). **d** Nucleosomal occupancy of the genome-wide **2** enriched region containing identified binding site. The region inside the yellow box corresponding to the nucleosomal region of approximately 147 bps [nucleosomal positioning was measured based on the ENCODE nucleosomal positioning data (16c)]

recognition site (5′-WGGWCA-3′). Whereas, one bp mismatch (5′-AGGTWA-3′) and two bps mismatch (5′-AGGCWA-3′) sequence of **2** showed poor distribution in the enriched regions. In Fig. 4.5b, we have also illustrated the mismatch DNA-alkylating site (5′-AGGTCT-3′) that displayed very low frequency. These results confirm the efficiency of the genomic pull-down enrichment by **2**. Additionally, the γ-turn in **2** uniquely recognized the base A (adenine), which may

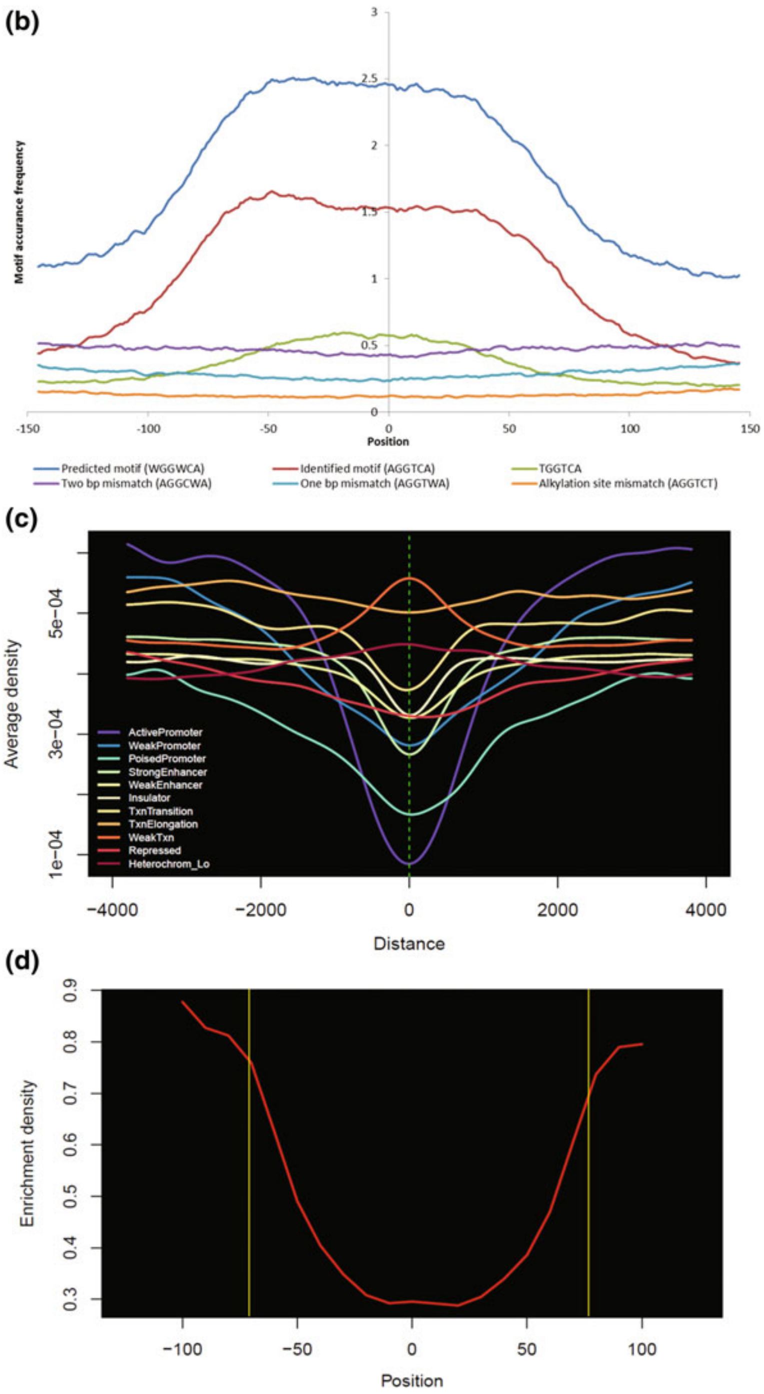


Fig. 4.5 (continued)

be the result of chromatinized DNA binding (Fig. 4.4). This exclusive detection was further verified with the results of poorly dispersed possible recognition of T (5′-TGGTCA-3′) by the γ-turn in the conjugate **2** (green line in Fig. 4.5b). Reproducibility of the experiment is confirmed with $R^2 = 0.92$ for two separate experimental enriched motifs (Extended Fig. 4.1) and MACS peak annotation distribution (Extended Table 4.4).

*4.2.5 Identification of PIP-Indole-*Seco*-CBI Conjugate 2 Representative Enriched Sites in Chromatinized Human Genome*

Targeted gene silencing can be achieved by PIPs, either by designing the PIPs to the transcription factor recognition site on the gene promoter [31] or by alkylating PIPs targeted to the coding region of the respective gene producing nonfunctional truncated mRNA [32]. To obtain deeper insight into the gene of interest targeted by PIP-indole-*seco*-CBI conjugate **2**, the enriched peak regions were mapped, and correlated with the DNA-alkylating site of **2** on the human genome. One of the significantly enriched genomic regions *ERBB2* is shown in Fig. 4.6 (enrichment details are given in Extended Table 4.5) with the DNA-alkylating site. Several oncogenes have been studied in human cancers, but only a few have been reported

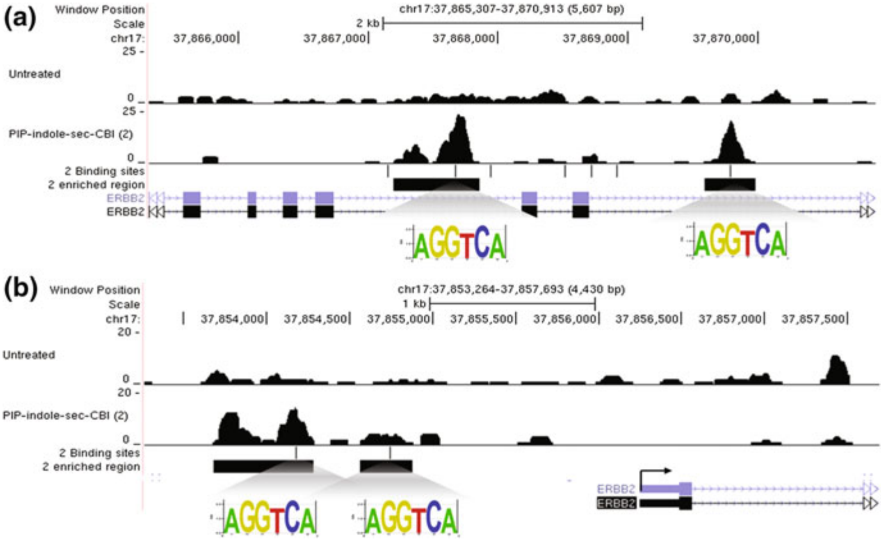


Fig. 4.6 Genome-wide mapping of PIP-indole-*seco*-CBI (2) **a** 2 binding and enriched genomic area in the *ERBB2* gene coding region. **b** 2 binding and enriched genomic area of the *ERBB2* promoter region

to play a critical role in the progression of breast cancer. *ERBB2* is one such oncogene overexpressed in many epithelial cancers and in about 20% of early-stage breast cancer patients with low survival rates, and confers chemo-resistance [33]. By contrast, the universally expressed reference genes, such as *ACTB* and *GAPDH* did not show any enriched regions (Extended Fig. 4.5). However, the quantification of *ERBB2* mRNA (using real-time PCR) in conjugate **2** treated human BJ skin fibroblast cells and SKBR3 breast adenocarcinoma cells showed inhibition of *ERBB2* mRNA expression with respect to reference genes in both cell types (Extended Fig. 4.6).

*4.2.6 Genome-Wide Analysis of PIP-Indole-*Seco*-CBI Conjugate 2 Enriched Sites Distribution*

To analyze the influence of complex eukaryotic chromatin conformation on the DNA-alkylation preferences of PIP conjugate, we performed a computational genome-wide distribution survey of **2** binding sites (predicted binding site 5′-WGGWCA-3′ based on the binding rule and genome-wide experimentally derived binding site 5′-AGGTCA-3′) in various annotated regions of the human genome (Extended Table 4.6 and Fig. 4.5a i and ii). The total numbers of experimentally identified binding sites are significantly less than the predicted binding sites. This clearly shows that chromatin conformation plays a critical role in polyamide binding. We have also annotated the **2** enriched genomic regions across the human genome (Extended Table 4.6 and Fig. 4.5a iii). The result showed high correlation with the genome-wide distribution pattern of DNA-alkylating sites.

We next sought to assess the annotated position of the predominant **2** enriched regions across the genome. We first compared the distribution of **2** identified motifs (DNA-alkylating site) with predicted binding sites (Extended Table 4.7 and Extended Fig. 4.3a) that retained comparable distribution of genomic positions (3′-UTR, miRNA, ncRNA, TTS, pseudo genes, exon, intron, intergenic, promoter, and 5′-UTR). We again compared **2** enriched regions with identified and predicted binding regions that showed the rate of enrichment to be high in TTS, intron, intergenic, and promoter regions when compared with the other genomic regions (Extended Table 4.7 and Extended Fig. 4.3b, c). To test this predominance, we generated an aggregation plot [34] with **2** enriched reads (Extended Fig. 4.4). The average enrichment read density in the upstream (−4 kb) of TSSs (proximal and distal promoter) possess convincing read density and gene body detects crimped average enrichment read density because of the existence of 5′-UTR, 3′-UTR, coding exons and introns in the gene body (Extended Fig. 4.4). These distribution marks are consistent with the enrichment distribution we determined earlier.

We next asked whether **2** enrichment profiling with various chromatin states could be used to infer a systematic means of perceiving PIP accessible regions;

because the chromatin framework of a genome plays a central role in controlling DNA access. We examined the **2** enriched sites distribution on ChromHMM segments and we present in Fig. 4.5c, on a broad scale; (i) **2** differs in accessing various promoter states and its low average enrichment density on active promoters may support the target specific gene suppression by PIP conjugates. (ii) The positional distribution along enhancer, insulator and transcribed regions contain an almost equivalent form of enrichment, so designing PIPs to target such genomic positions may provide a correlative genome-wide effect. (iii) **2** showed characteristic patterns of chromatin accessibility that have been observed at repressed and inactive states. By contrast, the access of nucleosomal DNA by PIP is limited [20]. In line with this, we sought to inspect **2** enrichment sites on nucleosomes at a fine scale of an approximately 147 bps window, we investigated the positioning of nucleosomes with the enriched regions using the MNase-Seq data [32]. In this model, we were able to estimate the **2** binding density, which appears to be higher at the ends of nucleosome than in the core middle region (Fig. 4.5d). Our data with greater precision at the genomic scale confirm the previously reported (studied at the defined nucleosomal core particle (NCP) context) limitation of PIP accessibility; in addition, we report that this limitation might be as a result of the central core of well-positioned nucleosomes. Overall, our genome-wide enrichment assessment results provide a deeper understanding of the PIP accessibility towards chromatinized DNA.

4.2.7 Discussion

New methods in chemical biology with deep sequencing applications and data analysis are constantly being developed [24]. In this study, we have made use of high-throughput sequencing technology to show the significant sequence-specific DNA-alkylation of PIP-indole-*seco*-CBI conjugates corresponding to the proposed DNA-binding rule. The binding specificity of our small molecule remains similar in a broad sequence context of free DNA and in complex chromatinized human genome. However, our DNA-alkylating site mapping on the nucleosome resulting in this sequence specific binding have limitations on the central core of chromatinized DNA. Progress in the field of sequencing technologies has allowed us to conduct this study cost-efficiently, and it may be a useful method for other DNA-binding small molecule design and redesign in the context of the complex genome space. Our results also indicate that the structural composition of PIP-indole-*seco*-CBI conjugates favorably alters the sequence specificity of PIPs. In future, this method may be an efficient tool for studying sequence-specific alkylation and in designing small molecules for targeted gene silencing compared with conventional PAGE analysis.

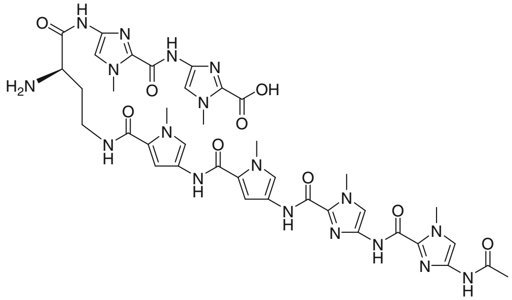
4.3 Materials and Methods

4.3.1 *Synthesis of Biotin-Conjugated Alkylating Polyamide Conjugates*

Reagents and solvents were purchased from standard providers and used without further refinement. The EZ-Link NHS-PEG$_{12}$-Biotin was obtained from Thermo Scientific, USA (No. 21312). Analytical HPLC was performed using a Cosmosil 5C$_{18}$-MS-II reversed phase column (4.6 mm × 150 mm, Nacalai Tesque) in 0.1% TFA in water with CH_3CN as eluent at 1.0 mL/min, and a linear gradient elution of 0–100% CH_3CN over 20 or 40 min with detection at 254 nm. The HPLC purification was performed with a Cosmosil 5C$_{18}$-MS-II reversed phase column (10 mm × 150 mm, Nacalai Tesque) in 0.1% TFA in water with CH_3CN as the eluent. The final products were analyzed by ESI-TOF-MS (Bruker).

1-i

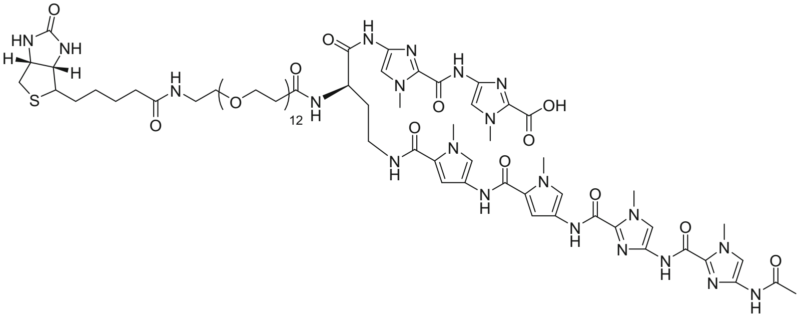
52 mg of CLEAR Acid resin (34.9 μmol/100 mg) was treated with 1 mL solution of 95% TFA, 2.5% water and 2.5% triisopropylsilane at room temperature for 30 min to obtain 8.2 mg of crude. This compound was purified by HPLC and analysed by HPLC, MS and NMR.



HPLC retention time: 11.0 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{38}H_{45}N_{18}O_9^+$ $[M + H]^+$ 897.3611, found 897.3615. ^1H NMR (600 MHz, DMSO-d_6): δ = 11.12 (s, 1H), 10.35 (s, 1H), 10.30 (s, 1H), 9.96 (s, 1H), 9.80 (s, 1H), 9.34 (s, 1H), 8.32 (brd, J = 4.1 Hz, 3H), 8.17 (brt, J = 6.2 Hz, 1H), 7.64 (s, 1H), 7.58 (s, 1H), 7.57 (s, 1H), 7.51 (s, 1H), 7.28 (d, J = 1.4 Hz, 1H), 7.19 (s, 1H), 7.17 (d, J = 1.4 Hz, 1H), 6.96 (d, J = 1.4 Hz, 1H), 4.03 (m, 1H), 4.01 (s, 3H), 3.98 (s, 6H), 3.93 (s, 3H), 3.85 (s, 3H), 3.81 (s, 3H), 3.31 (m, 2H), 2.04 (s, 3H), 2.00 (m, 2H).

1-bio

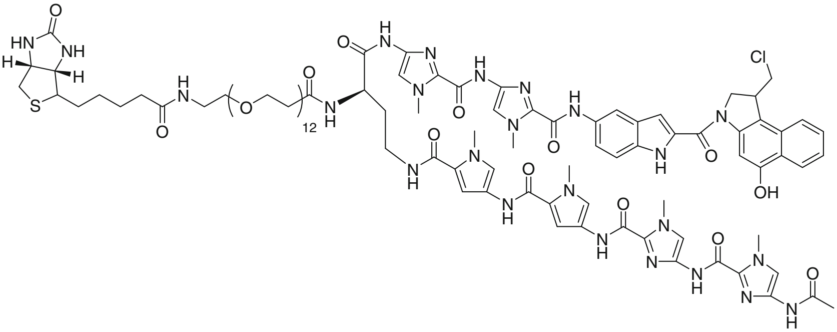
3.0 mg (3.3 μmol) of **1-i**, 31 μL (3.1 mg, 3.3 μmol, 1.0 eq) of DMF solution of NHS-PEG$_{12}$-Biotin and 3.4 μL (6.0 eq, 20 μmol) of DIEA were mixed and shaken at room temperature for 2 h. The products were triturated by Et_2O to obtain 6.2 mg of the crude.



HPLC retention time: 12.5 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{75}H_{113}N_{21}O_{24}S^{2+}$ [*M* + 2H]$^{2+}$ 861.8989, found 861.9007.

1

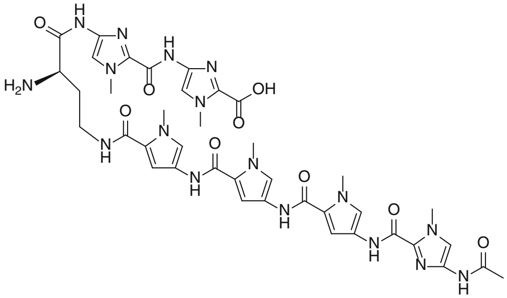
6.2 mg (3.6 µmol) of **1-bio**, 3.7 mg (2.0 eq, 7.2 µmol) of PyBOP were dissolved in 36 µL of DMF, and to that solution was added 3.8 µL (6.0 eq, 22 µmol) of DIEA. After shaking at room temperature for 1 h, 2.8 mg (2.0 eq, 7.2 µmol) of aminoindole-*seco*-CBI was added and shaken for another 1 h. The crude obtained by treatment with Et_2O was purified by HPLC, and 1.4 mg (0.67 µmol, 20% from **1-i**) was obtained as a reddish gray solid.



HPLC retention time: 13.6 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{97}H_{130}ClN_{24}O_{25}S^{3+}$ [*M* + 3H]$^{3+}$ 699.3011, found 699.3031.

2-i

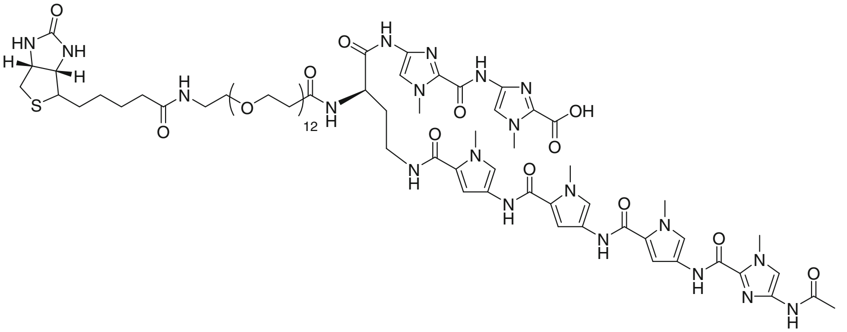
55 mg of CLEAR Acid resin (34.9 µmol/100 mg) was treated with 1 mL solution of 95% TFA, 2.5% water and 2.5% triisopropylsilane at room temperature for 30 min to obtain 15.6 mg of crude. This compound was purified by HPLC and analyzed by HPLC, MS and NMR.



HPLC retention time: 11.6 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{39}H_{46}N_{17}O_9^+$ $[M + H]^+$ 896.3659, found 896.3706. ^1H NMR (600 MHz, DMSO-d_6): δ = 11.11 (s, 1H), 10.23 (s, 1H), 9.97 (s, 1H), 9.95 (s, 1H), 9.93 (s, 1H), 9.80 (s, 1H), 8.32 (brd, J = 4.1 Hz, 3H), 8.17 (brt, J = 5.5 Hz, 1H), 7.64 (s, 1H), 7.57 (s, 1H), 7.43 (s, 1H), 7.27 (d, J = 1.4 Hz, 1H), 7.23 (d, J = 1.4 Hz, 1H), 7.19 (s, 1H), 7.16 (d, J = 1.3 Hz, 1H), 7.08 (d, J = 2.0 Hz, 1H), 6.96 (d, J = 1.4 Hz, 1H), 4.02 (m, 1H), 3.98 (s, 3H), 3.95 (s, 3H), 3.93 (s, 3H), 3.86 (s, 3H), 3.85 (s, 3H), 3.81 (s, 3H), 3.30 (m, 2H), 2.03 (s, 3H), 1.99 (m, 2H).

2-bio

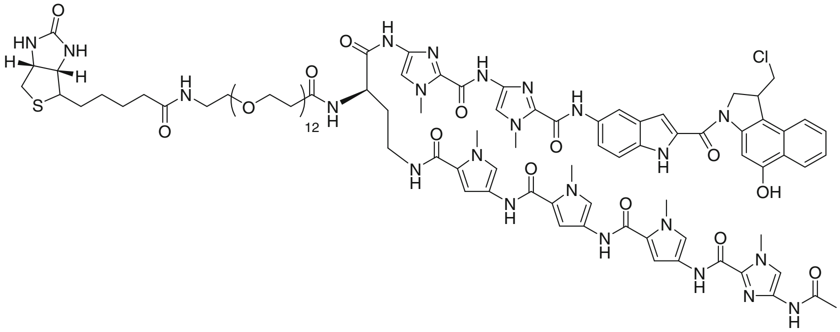
4.0 mg (4.5 μmol) of **2-i**, 42 μL (4.2 mg, 4.5 μmol, 1.0 eq) of DMF solution of NHS-PEG$_{12}$-Biotin and 4.6 μL (6.0 eq, 27 μmol) of DIEA were mixed and shaken at room temperature for 4 h. The products were triturated by Et_2O to obtain 8.3 mg of the crude.



HPLC retention time: 12.4 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{76}H_{114}N_{20}O_{24}S^{2+}$ $[M + 2H]^{2+}$ 861.4012, found 861.3976.

2

8.3 mg (4.8 μmol) of **2-bio**, 4.7 mg (1.9 eq, 9.0 μmol) of PyBOP were dissolved in 45 μL of DMF, and to that solution was added 4.6 μL (5.6 eq, 27 μmol) of DIEA. After shaking at room temperature for 2 h, 3.5 mg (1.9 eq, 9.0 μmol) of aminoindole-*seco*-CBI was added and shaken for another 1.5 h. 8.9 mg of the crude was obtained by treatment with Et_2O and CH_2Cl_2. After HPLC purification, 0.9 mg (0.4 μmol, 9% from **2-i**) was obtained as a reddish gray solid.



HPLC retention time: 13.5 min (0–100% over 20 min). ESI-TOFMS m/z calculated for $C_{98}H_{131}ClN_{23}O_{25}S^{3+}$ $[M + 3H]^{3+}$ 698.9693, found 698.9670.

PIP-conjugate **5**

We followed the synthetic procedure explained in our previous report [30].

4.3.2 Bind-n-Seq Experiment and High-Throughput Sequencing

Bind-n-Seq experiments and high-throughput sequencing were performed based on our previous report [25]. Broadly,

(1) Synthesis of biotinylated alkylating PIPs and PIP-Conjugate **5** [a PIP conjugate where the alkylating CBI moiety was substituted with 3-dimethylamino-propylamine (Dp)] [30], and a separate set of oligonucleotides with a 10- and 21-mer randomized region and ion torrent adapters. Oligonucleotides were duplexed by primer extension. Biotin conjugated alkylating PIPs and PIP-Conjugate **5** were allowed to bind and alkylate with their specific binding region of duplex randomized oligonucleotides separately at room temperature. Control experiments were performed without PIP-indole-*seco*-CBI conjugates/ PIP-Conjugate **5**, the data obtained were used for the normalization to obtain enrichment data. Biotin–streptavidin affinity-based purification was used to enrich the alkylating PIP attached DNA (washing steps were doubled for the PIP-indole-*seco*-CBI conjugates compared with the previously reported Bind-n-Seq to remove PIP simple binding).

(2) Enriched DNA was subjected to polymerase chain reaction to recover the alkylated DNA strand using sequencing library adapter specific primers. The purified sequencing libraries were quantified using a BioAnalyzer with an Agilent DNA High Sensitivity BioAnalyzer kit, Agilent technologies, USA. Sufficient sequencing libraries with various barcodes were pooled for template preparation (Ion PGM template OT2 200 kit) in an Ion OneTouch 2 system. The emulsion PCR amplified libraries were further enriched with Ion OneTouch ES. The enriched libraries were sequenced following the manufacturer's instructions with Ion PGM sequencer (Ion PGM sequencing 200 kit v2 and 318 chip V2 (Life Technologies, USA).

(3) The sequenced reads were analyzed for a primary motif calling based on our previous reports [25, 30, 35–37].

4.3.3 *Affinity Purification-Based High-Throughput Sequencing of Human Genomic Regions Enriched with PIP-Indole-*Seco*-CBI Conjugate 2*

Human fibroblast BJ from neonatal foreskin (ATCC, USA), were maintained in 10% FBS (FBS, Japan Serum) supplemented with Dulbecco's modified eagle medium (DMEM, Nacalai Tesque, Japan), 10% HyClone fetal bovine serum (FBS), nonessential amino acids, 100 U/mL penicillin, 100 μg/mL streptomycin, and grown to 75–80% confluency in a humidified atmosphere of 5% CO_2 at 37 °C. Nuclei were isolated for alkylating PIP treatment [38–41]. In brief, 2×10^6 P6 cells were washed with PBS and isolated by 3 min trypsinization. The isolated cells were again washed 2× with ice-cold PBS. The cell pellet was suspended in 5 mL of ice cold NP-40 lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 0.5% Nonidet P-40, 0.15 mM spermine, 0.5 mM spermidine, and 0.1× protease inhibitor cocktail) and incubated on ice for 5 min. The nuclei were pelleted by centrifugation at 300 *g* for 10 min. The pellet of nuclei was carefully resuspended in modified binding buffer [22] [10 mM Tris-Cl (pH 8.0), 5 mM $MgCl_2$, 1 mM DTT, 0.3 M KCl, 0.3× protease inhibitor cocktail and 10% glycerol]. The nuclei were incubated with 400 nM of **2** (dissolved in DMSO, 0.1% final concentration) at 4 °C for 16 h. Control experiments were performed without PIP-indole-*seco*-CBI conjugates and with a 0.1% final concentration of DMSO. We used the PIP concentrations from the previous report [9] that were consistent with the PIP quantity measured in the nuclei of treated cells [42]. PIP-containing nuclei were washed with micrococcal nuclease (MNase) buffer [10 mM Tris-HCl (pH 7.4), 15 mM NaCl, 60 mM KCl, 0.15 mM spermine, 0.5 mM spermidine, and 0.1× protease inhibitor cocktail] [38]. Cell nuclei suspension was digested with MNase (TaKaRa, Japan) for 30 min at 37 °C to obtain mononucleosomes in optimized reaction conditions. The reactions contained MNase buffer, MNase [0.2 μL of MNase (20 Units/μL) for nucleus extracted from 2×10^6 cells], RNase A and protease inhibitor cocktail. After digestion, the histone protein was removed by proteinase K treatment. After MNase digestion and proteinase K treatment the suspension was mixed with an equal volume of modified COSMIC buffer [20 mM Tris-Cl (pH 8.1), 2 mM EDTA, 150 mM NaCl, 0.1× protease inhibitor cocktail, 1% Triton-X100, and 0.1% SDS] [9]. 10% of the sample was saved as input DNA. The assessment of the size distribution of DNA samples showed about 100–180 bp fragment distribution.

Preparation of magnetic beads. After removing the suspension solution from streptavidin-coated magnetic beads (Dynabeads MyOne C1, Life Technologies, USA). They were washed with 2× modified COSMIC buffer and resuspended in the same buffer. Resuspended streptavidin-coated magnetic beads (0.5 mg) were incubated with samples for 16 h at 4 °C. After the incubation period, bound and unbound DNA samples were separated using affinity purification. Briefly, samples were washed 5 min once with 0.5 mL of washing buffer 1 [10 mM Tris-Cl (pH 8.0), 1 mM EDTA, 3% SDS], once with altered washing buffer 2 [10 mM Tris-Cl (pH 8.0), 250 mM LiCl, 1 mM EDTA, 0.5% NP40], 2× with altered washing buffer 3 [10 mM Tris-Cl (pH 7.5), 1 mM EDTA, 0.1% NP-40], and 3× with TE. The samples were then resuspended in elution buffer [10 mM Tris-HCl (pH 7.6), 0.4 mM EDTA and 100 mM KOH] [22] and DNA was eluted from magnetic beads after heating at 90 °C for 30 min. The remaining DNA with the beads were eluted using elution buffer 2 (2% SDS, 100 mM $NaHCO_3$ and 3 mM biotin) with heating to 65 °C for 8–12 h. The detached Samples were purified with a QIAquick PCR purification Kit (Qiagen, CA, USA) and quantified.

Polyamide-based affinity purification sequencing libraries were prepared using standard Ion Xpress™ Plus gDNA Fragment Library Preparation reagents and protocols (Life technologies, USA). Sequencing adapter ligated enriched DNA was subjected to a polymerase chain reaction to recover the alkylated DNA strand using sequencing library adapter specific primers and purified. The purified libraries were subjected to quality and quantity checks with an Agilent DNA High sensitivity BioAnalyzer kit (Agilent technologies, USA). The qualified libraries were used for high-throughput sequencing. The sequencing was performed, starting with template preparation using Ion PGM template OT2 200 v2 kit and an Ion PI template OT2 200 kit using an Ion OneTouch 2 system. The templates were then enriched using Ion OneTouch ES. The enriched libraries were sequenced with 260–300 flow of a single read performed with an Ion PGM sequencer using an Ion PGM sequencing 200 kit v2/318 v2 chip and an Ion Proton Sequencer using an Ion PI Sequencing 200 kit v3/Ion PI chip following the manufacturer's guidelines, and we produced 25–30 million post filtered reads per library. The data were handled by employing standard program packages in the Ion torrent suite. A Torrent Mapping Alignment Program version 4.4.2 (TMAP) was used for aligning reads (mean read length 121 bp), ChIP-seq peaks were called using MACS version 1.4.2 [43] with the default parameters and a $p < 0.001$ (cutoff < 10^{-2}). In total, 721,617 peaks were considered significant with the total number of binding sites. Enriched peak and PIP-indole-*seco*-CBI conjugate binding site annotations were made using Homer [44]. To determine the high-affinity DNA-alkylating sites (motif) of **2** over the control sequence reads, a randomly sampled 10–15% of the uniquely mapped reads for each setting were used. The random sampling was performed using a Perl script (http://meme-suite.org/doc/fasta-subsample.html). We followed our previous analysis pipeline for motif calling [25, 30, 35–37].

We evaluated the genome-wide enrichment signature of **2** by calculating the cross-correlation of MACS peaks with the identified binding sites (DNA-alkylating motif). The peaks containing identified binding sites were considered as significant enrichment regions (total of 355,882 peaks). Analysis pipelines agplus [45], Position Weight Matrix model generation and evaluation-PWMScan (http://ccg.vital-it.ch/pwmscan) [46, 47], ChIP-Cor Analysis Module (http://ccg.vital-it.ch/chipseq/chip_cor.php) and various platforms of the Signal Search Analysis Server (http://ccg.vital-it.ch/ssa) were used to evaluate the spatial precision of **2** enrichment data.

4.3.4 Validating PIP-Indole-Seco-CBI Conjugate 2 Bound and Enriched Region in the Human Genome

Human fibroblast BJ from neonatal foreskin and SKBR3 (breast adenocarcinoma, human) cell lines were purchased from ATCC. Fibroblast cells were grown in DMEM supplemented with 10% HyClone fetal bovine serum (FBS), nonessential amino acids, 100 U/mL penicillin, 100 μg/mL streptomycin, at 37 °C in 5% CO_2. SKBR3 cells were grown in ATCC-formulated McCoy's 5a modified medium complemented with 10% FBS and were maintained under an atmosphere of 5% CO_2 at 37 °C.

The effect of alkylating PIP **2** on the expression of *ERBB2* mRNA was determined in both fibroblast and SKBR3 cell lines using real-time PCR. BJ skin fibroblast cells were seeded at a density of 5×10^4 cells/well of a 6-well plate and SKBR3 cells were plated into the 6-well plate at 4×10^5 cells/well. The cells were then treated with 50 and 100 nM of alkylating PIP **2** for 48 h with DMSO as a corresponding control sample. After 48 h, total RNA was isolated using RNEasy Kit (Qiagen) and cDNA was synthesized by ReverTra Ace qPCR RT Master mix with genomic DNA remover (Toyobo, Japan) following the manufacturer's instructions. The expression level of *ERBB2* was normalized using *β-actin*, as an internal control. The primers used in this study includes, *β-actin* sense, 5′-CAATGTGGCCGAGGACTTTG-3′ and antisense, 5′-CATTC TCCTTAGAGAG AAGTGG-3′. The sense primer of *ERBB2* is 5′-AGCCGCGAGCA CCCAAGT-3′ and antisense, 5′-TTGGTGGGCAGGTAGGTGAGTT-3′.

Appendix

See Extended Tables 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7.

See Extended Figs. 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6.

Extended Table 4.1 Bind-n-Seq relative enrichment ratio of "K"mer (K = 6 bases), using **1** binding and enrichment reactions

Rank	Sequence	Enrichment
1	*TGGCCA*	*24.111*
2	TTGGCC	20.186
3	*AGGCCA*	*18.619*
4	ATGGCC	18.58
5	GTGGCC	16.21
6	*TGGCCT*	*16.194*
7	CTGGCC	15.085
8	GGCCAT	14.379
9	GGCCAA	12.114
10	TAGGCC	12.105
11	AAGGCC	11.993
12	GAGGCC	11.343
13	GGCCAG	9.981
14	GGCCTT	7.41
15	CAGGCC	7.302
16	GGCCAC	7.271
17	GGCCTC	5.878
18	GGCCTG	5.735
19	TATGGC	5.462
20	CATGGC	5.36

Extended Table 4.2 Bind-n-Seq relative enrichment ratio of “K”mer (K = 6 bases), using **2** binding and enrichment reactions

Rank	Sequence	Enrichment
1	*AGGTCA*	*7.99*
2	*TGGTCA*	*7.36*
3	GGTCAT	4.922
4	TGACCT	4.777
5	ATGGTC	4.752
6	TTGACC	4.642
7	CTGACC	4.531
8	TGACCA	4.493
9	GGTCAA	4.085
10	TTGGTC	3.998
11	GGTCAG	3.836
12	ATGACC	3.835
13	TAGGTC	3.797
14	GAGGTC	3.757
15	AAGGTC	3.665
16	CAGGTC	3.648
17	TGGTTA	3.618
18	CTGGTC	3.57
19	GTGGTC	3.509
20	AGGTTA	3.505

Extended Table 4.3 Genomic relative enrichment ratio of “K”mer (K = 6 bases), using **2** affinity purification based genomic DNA sequencing data

Rank	Sequence	Genomic Enrichment ratio
1	AGGTCA	3.082
2	GACCTC	3.037
3	TCGAGA	2.95
4	CTGACC	2.933
5	AGTTCG	2.914
6	CCTGAC	2.794
7	GTTCGA	2.757
8	GATTAC	2.74
9	ACCTCG	2.731
10	AAGTGC	2.656
11	CGAGAC	2.629
12	ACGCCT	2.576
13	ATCCGC	2.571
14	CGGATC	2.565
15	GGATTA	2.562
16	ACGAGG	2.488
17	GGCCAA	2.465
18	AGCACT	2.463
19	TCACGA	2.46
20	CGTGAG	2.449

Extended Table 4.4 Relative enrichment of MACS peak distribution of two separate affinity purification based sequencing

Genomic region	Experiment 1 (Number of peaks)	Experiment 2 (Number of peaks)
3UTR	4963	3716
miRNA	10	4
ncRNA	1257	864
TTS	8457	7846
pseudo	393	279
Exon	8408	3619
Intron	356,856	338,909
Intergenic	401,384	359,624
Promoter	8026	6470
5UTR	543	286

Extended Table 4.5 MACS peak details of the *ERBB2* gene promoter and gene coding region

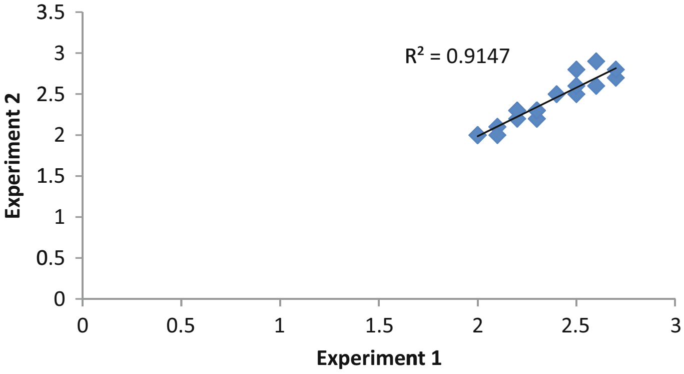
Chr	Start	End	Length	Summit	Tags	−10*log10 (*pvalue*)	Fold enrichment	FDR (%)
Chr17	37,853,681	37,854,283	603	497	33	116.67	9.69	1.85
Chr17	37,867,196	37,867,854	659	495	41	99.96	10.16	1.85
Chr17	37,869,590	37,869,982	393	209	22	89.59	14.3	1.89

Extended Table 4.6 Genomic distribution of alkylating PIP **2** predicted binding sites based on the canonical binding rule, experimentally identified high-affinity binding motif in human genome and affinity purification based sequencing enriched regions

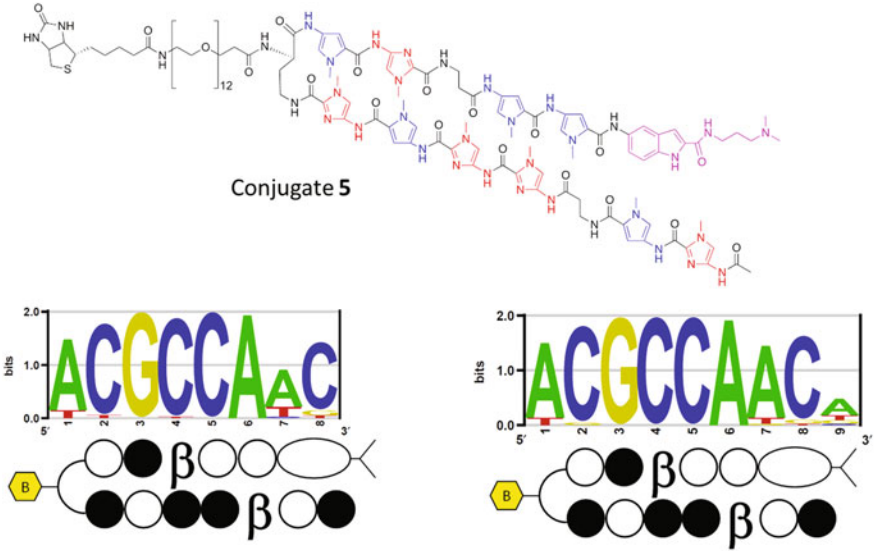
Alkylating PIP Binding	3UTR	miRNA	ncRNA	TTS	Pseudo	Exon	Intron	Intergenic	Promoter	5UTR
Predicted binding site (WGGWCA)	27,344	61	7563	36,025	2851	54,363	1,419,419	1,819,424	34,746	2654
Identified binding motif (AGGTCA)	6883	17	1861	10,477	683	11,211	424,333	508,885	10,131	638
Alkylating PIP 2 Enriched regions	1655	4	436	4109	125	770	172,826	172,791	3084	82

Extended Table 4.7 Comparative genomic distribution analysis of alkylating PIP **2**

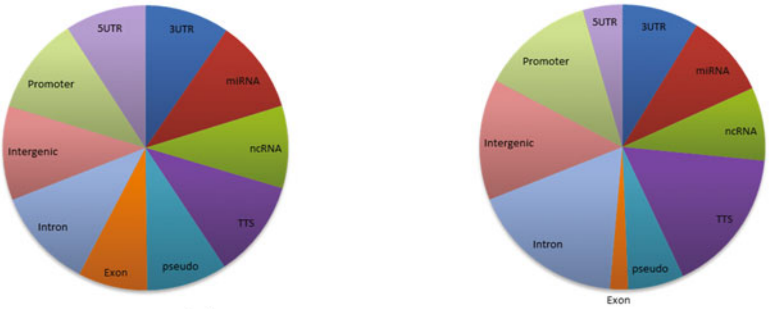
Comparative Percentage	3UTR (%)	miRNA (%)	ncRNA (%)	TTS (%)	Pseudo (%)	Exon (%)	Intron (%)	Intergenic (%)	Promoter (%)	5UTR (%)
Identified motif (AGGTCA) compared to predicted (WGGWCA)	25.17	27.87	24.61	29.08	23.96	20.62	29.89	27.97	29.16	24.04
Enriched region compared to predicted (WGGWCA)	6.05	6.56	5.76	11.40	4.38	1.42	12.17	9.50	8.86	3.08
Enriched region compared to identified motif (AGGTCA)	24.04	23.52	23.42	39.22	18.30	6.87	40.73	33.95	30.44	12.85



Extended Fig. 4.1 Graphical representation: Relative enrichment of two separate affinity purification based sequencing identified genomic binding sequence enrichment for **2**

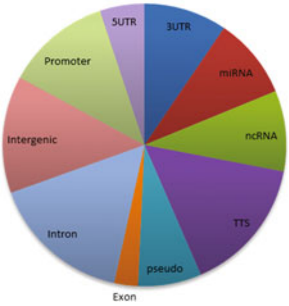

Conjugate **5**

Extended Fig. 4.2 Chemical structure of PIP-conjugate **5** and its Bind-n-Seq identified motifs with the kmer of 8 and 9 bp

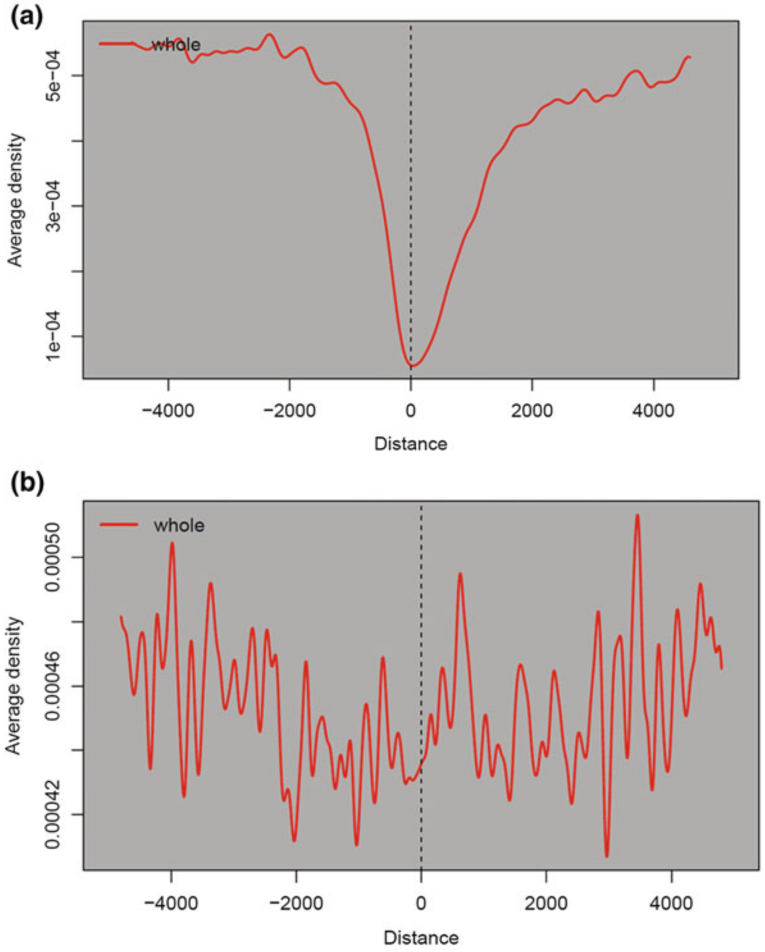

(a) 2 Identified motif (AGGTCA) compared to predicted (WGGWCA)

(b) 2 Enriched region compared to predicted (WGGWCA)

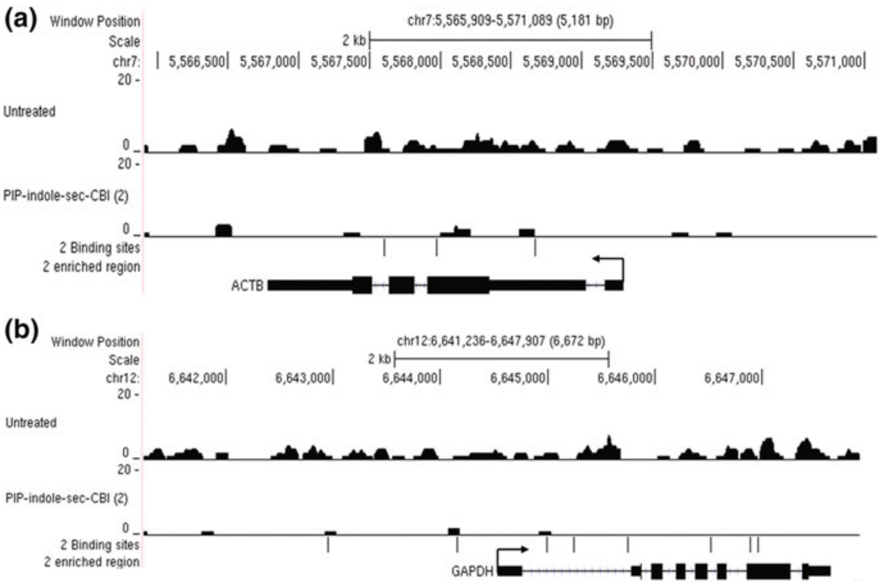
(c) 2 Enriched region containing to identified motif (AGGTCA)



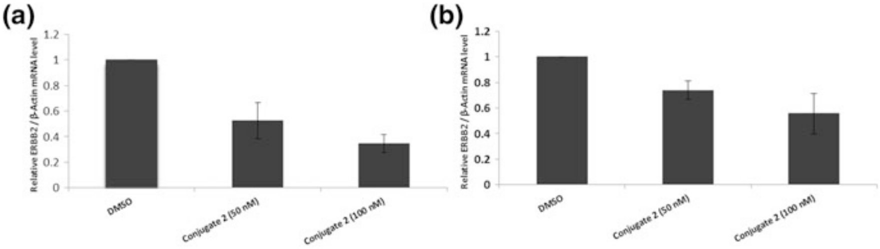
Extended Fig. 4.3 Graphical representation of comparative genomic region distribution of alkylating PIP **2** DNA-alkylating and enrichment regions



Extended Fig. 4.4 Genome-wide distribution of PIP-indole-*seco*-CBI **2** enriched region containing identified binding motif at the **a** TSSs (Transcription Start Sites) **b** gene body (Including 5′UTR, 3′UTR, coding exons and introns)



Extended Fig. 4.5 **a** Sequencing read distribution in ACTB and its promoter region. **b** Sequencing read distribution signal in GAPDH and its promoter region



Extended Fig. 4.6 Effects of PIP-indole-*seco*-CBI **2** on *ERBB2* mRNA inhibition. **a** Human BJ skin fibroblast cells **b** SKBR3 breast adenocarcinoma cells

References

1. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. Curr Opin Struct Biol 13:284–299
2. Warren CL, Kratochvil NCS, Hauschild KE et al (2006) Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A 103:867–872. doi:10.1073/pnas.0509843102
3. Puckett JW, Muzikar KA, Tietjen J et al (2007) Quantitative microarray profiling of DNA-binding molecules. J Am Chem Soc 129:12310–12319. doi:10.1021/ja0744899

4. Keleş S, Warren CL, Carlson CD, Ansari AZ (2008) CSI-Tree: a regression tree approach for modeling binding properties of DNA-binding molecules based on cognate site identification (CSI) data. Nucleic Acids Res 36:3171–3184. doi:10.1093/nar/gkn057
5. Moretti R, Donato LJ, Brezinski ML et al (2008) Targeted chemical wedges reveal the role of allosteric DNA modulation in protein–DNA assembly. ACS Chem Biol 3:220–229. doi:10.1021/cb700258r
6. Ozers MS, Warren CL, Ansari AZ (2009) Determining DNA sequence specificity of natural and artificial transcription factors by cognate site identifier analysis. Methods Mol Biol 544:637–653. doi:10.1007/978-1-59745-483-4_41
7. Tietjen JR, Donato LJ, Bhimisaria D, Ansari AZ (2011) Sequence-specificity and energy landscapes of DNA-binding molecules. Methods Enzymol 497:3–30. doi:10.1016/B978-0-12-385075-100001-9
8. Carlson CD, Warren CL, Hauschild KE et al (2010) Specificity landscapes of DNA binding molecules elucidate biological function. Proc Natl Acad Sci U S A 107:4544–4549. doi:10.1073/pnas.0914023107
9. Kondo N, Takahashi A, Ono K, Ohnishi T (2010) DNA damage induced by alkylating agents and repair pathways. J Nucleic Acids 2010:1–7. doi:10.4061/2010/543531
10. Lindahl T (1993) Instability and decay of the primary structure of DNA [see comments]. Nature 362:709–715
11. Rouse J, Jackson SP (2002) Interfaces between the detection, signaling, and repair of DNA damage. Science (80-) 297:547–551. doi:10.1126/science.1074740
12. Zhou BB, Elledge SJ (2000) The DNA damage response: putting checkpoints in perspective. Nature 408:433–439. doi:10.1038/35044005
13. Hurley LH (2002) DNA and its associated processes as targets for cancer therapy. Nat Rev Cancer 2:188–200
14. Bando T, Sugiyama H (2006) Synthesis and biological properties of sequence-specific DNA-alkylating pyrrole-imidazole polyamides. Acc Chem Res 39:935–944
15. Shinohara KI, Bando T, Sasaki S et al (2006) Antitumor activity of sequence-specific alkylating agents: pyrolle-imidazole CBI conjugates with indole linker. Cancer Sci 97:219–225. doi:10.1111/j.1349-7006.2006.00158.x
16. Shinohara KI, Sasaki S, Minoshima M et al (2006) Alkylation of template strand of coding region causes effective gene silencing. Nucleic Acids Res 34:1189–1195. doi:10.1093/nar/gkl005
17. Hiraoka K, Inoue T, Taylor RD et al (2015) Inhibition of KRAS codon 12 mutants using a novel DNA-alkylating pyrrole–imidazole polyamide conjugate. Nat Commun 6:6706. doi:10.1038/ncomms7706
18. Pandian GN, Taniguchi J, Junetha S et al (2014) Distinct DNA-based epigenetic switches trigger transcriptional activation of silent genes in human dermal fibroblasts. Sci Rep 4:3843. doi:10.1038/srep03843
19. Jespersen C, Soragni E, James Chou C et al (2012) Chromatin structure determines accessibility of a hairpin polyamide-chlorambucil conjugate at histone H4 genes in pancreatic cancer cells. Bioorg Med Chem Lett 22:4068–4071. doi:10.1016/j.bmcl.2012.04.090
20. Gottesfeld JM, Melander C, Suto RK et al (2001) Sequence-specific recognition of DNA in the nucleosome by pyrrole-imidazole polyamides. J Mol Biol 309:615–629. doi:10.1006/jmbi.2001.4694
21. Dudouet B, Burnett R, Dickinson LA et al (2003) Accessibility of nuclear chromatin by DNA binding polyamides. Chem Biol 10:859–867. doi:10.1016/j.chembiol.2003.09.001
22. Erwin GS, Bhimsaria D, Eguchi A, Ansari AZ (2014) Mapping polyamide-DNA interactions in human cells reveals a new design strategy for effective targeting of genomic sites. Angew Chem Int Ed 53:10124–10128. doi:10.1002/anie.201405497
23. Northrup DL, Zhao K (2011) Application of ChIP-Seq and related techniques to the study of immune function. Immunity 34:830–842
24. Anandhakumar C, Kizaki S, Bando T et al (2015) Advancing small-molecule-based chemical biology with next-generation sequencing technologies. ChemBioChem 16:20–38

25. Anandhakumar C, Li Y, Kizaki S et al (2014) Next-generation sequencing studies guide the design of pyrrole-imidazole polyamides with improved binding specificity by the addition of β-alanine. ChemBioChem 15:2647–2651. doi:10.1002/cbic.201402497
26. Meier JL, Yu AS, Korf I et al (2012) Guiding the design of synthetic DNA-binding molecules with massively parallel sequencing. J Am Chem Soc 134:17814–17822. doi:10.1021/ja308888c
27. Kang JS, Meier JL, Dervan PB (2014) Design of sequence-specific DNA binding molecules for DNA methyltransferase inhibition. J Am Chem Soc 136:3687–3694. doi:10.1021/ja500211z
28. Bando T, Sasaki S, Minoshima M et al (2006) Efficient DNA alkylation by a pyrrole-imidazole cbi conjugate with an indole linker: sequence-specific alkylation with nine-base-pair recognition. Bioconjug Chem 17:715–720. doi:10.1021/bc060022w
29. Bando T, Narita A, Sasaki S, Sugiyama H (2005) Specific adenine alkylation by pyrrole-imidazole CBI conjugates. J Am Chem Soc 127:13890–13895. doi:10.1021/ja052412j
30. Taylor RD, Chandran A, Kashiwazaki G et al (2015) Selective targeting of the KRAS codon 12 mutation sequence by pyrrole-imidazole polyamide seco -CBI conjugates. Chem Eur J 21:14996–15003. doi:10.1002/chem.201501870
31. Syed J, Pandian GN, Sato S et al (2014) Targeted suppression of EVI1 oncogene expression by sequence-specific pyrrole-imidazole polyamide. Chem Biol 21:1370–1380. doi:10.1016/j.chembiol.2014.07.019
32. Shinohara KI, Narita A, Oyoshi T et al (2004) Sequence-specific gene silencing in mammalian cells by alkylating pyrrole-imidazole polyamides. J Am Chem Soc 126:5113–5118. doi:10.1021/ja031673v
33. Arteaga CL, Sliwkowski MX, Osborne CK et al (2011) Treatment of HER2-positive breast cancer: current status and future perspectives. Nat Rev Clin Oncol 9:16–32. doi:10.1038/nrclinonc.2011.177
34. Ernst J, Kheradpour P, Mikkelsen TS et al (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49. doi:10.1038/nature09906
35. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res 37:e151. doi:10.1093/nar/gkp802
36. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659. doi:10.1093/bioinformatics/btr261
37. Workman CT, Yin Y, Corcoran DL et al (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. doi:10.1093/nar/gki439
38. Published in association with Cold Spring Harbor Laboratory Press (2005) Micrococcal nuclease–Southern blot assay. Nat Methods 2:719–720
39. Richard-Foy H, Hager GL (1987) Sequence-specific positioning of nucleosomes over the steroid-inducible MMTV promoter. EMBO J 6:2321–2328
40. Enver T, Brewer AC, Patient RK (1985) Simian virus 40-mediated cis induction of the Xenopus beta-globin DNase I hypersensitive site. Nature 318:680–683. doi:10.1038/318680a0
41. Gaffney DJ, McVicker G, Pai AA et al (2012) Controls of nucleosome positioning in the human genome. PLoS Genet. doi:10.1371/journal.pgen.1003036
42. Hsu CF, Dervan PB (2008) Quantitating the concentration of Py-Im polyamide-fluorescein conjugates in live cells. Bioorg Med Chem Lett 18:5851–5855. doi:10.1016/j.bmcl.2008.05.063
43. Feng J, Liu T, Qin B et al (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728–1740. doi:10.1038/nprot.2012.101
44. Heinz S, Benner C, Spann N et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589. doi:10.1016/j.molcel.2010.05.004

45. Maehara K, Ohkawa Y (2014) Agplus: a rapid and flexible tool for aggregation plots. Bioinformatics 31:3046–3047. doi:10.1093/bioinformatics/btv322
46. Iseli C, Ambrosini G, Bucher P, Jongeneel CV (2007) Indexing strategies for rapid searches of short words in genome sequences. PLoS ONE. doi:10.1371/journal.pone.0000579
47. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25

Curriculum Vitae

Dr. Anandhakumar Chandran
Department of Chemistry, Graduate School of Science
Kyoto University, Kyoto, Japan
Current address
Ludwig Cancer Research, NDM, University of Oxford
Oxford OX3 7DQ, UK.
Tel: +44-7789873325
Email: anandhakumar.chandran@ludwig.ox.ac.uk

Education

- Ph.D., in Chemical biology, Graduate school of science, Kyoto University, Japan (Oct' 2012–Sept' 2015).
  Supervisor: **Prof. Hiroshi Sugiyama**, Chemical Biology Laboratory (Sugiyama Lab),
  Dissertation Title: Advancing Synthetic Gene Regulators Development with High-throughput Sequencing Technologies.
- M.Sc. in Sub Aqua Marine Ecology & Toxicogenomics (Marine Genomics), Madurai Kamaraj University, India (Jun' 2006–May' 2008).
- B.Sc. in Plant Biology & Plant Biotechnology, St. Xavier's College, Manonmaniam Sundaranar University, India (Jun' 2003–Apr' 2006).

Awards and honours

- Awarded with iCeMS, Kyoto University, Japan overseas visit travel grant July 2014.
- Won the JEES Mitsubishi Corporation international scholarship from Apr' 2013–Sept' 2015.

© Springer Nature Singapore Pte Ltd. 2018

A. Chandran, *Advancing Development of Synthetic Gene Regulators*,
Springer Theses, https://doi.org/10.1007/978-981-10-6547-7

- Won the Third prize for the poster
  C. Anandhakumar, V. Lavanya, K.G. Tirumurugaan, A. Raja, G. Dhinakar Raj and K. Kumanan. "Analysis and functional annotation of expressed sequence tags from toll like receptor agonist induced and uninduced shark (Chiloscyllium griseum) spleen" in the DBT, New Delhi sponsored National workshop on Research Advances in Fish Vaccines and prophylactics at Fisheries College and Research Institution from 14 to 15 feb' 2011, Thoothukudi, India.
- Scored 283 the National level Graduate Aptitude Test in Engineering (GATE-2008) Qualified in GATE 2008.