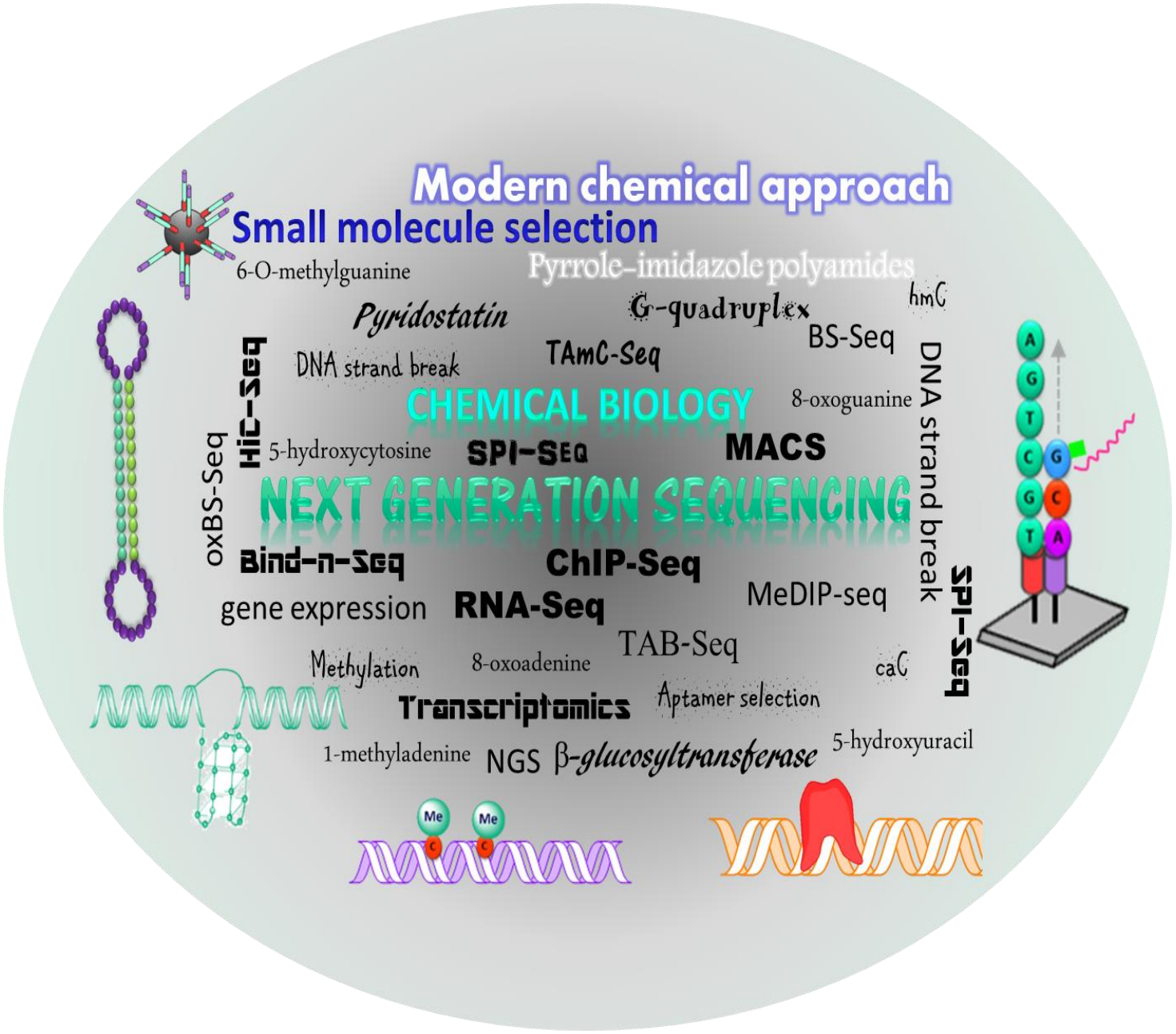
Advancing Small-Molecule-Based Chemical Biology With Next-Generation Sequencing Technologies

Chandran Anandhakumar,[a] Seiichiro Kizaki,[a] Toshikazu Bando,[a] Ganesh N. Pandian,[b] Hiroshi Sugiyama*[a,b]



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 prototype outputs 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 facilitates significant advances in diverse disciplines, ranging from discovery, design, and evaluation of many small molecules and relevant biological mechanisms to maturation of personalized therapies. Recent affordable NGS technologies 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 review, we focus on the overlooked implications of NGS technologies in chemical biology, with a special emphasis on small-molecule development and screening.

1. Introduction

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 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 machine 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 is 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] Next-generation sequences are also 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 (Figure 1).

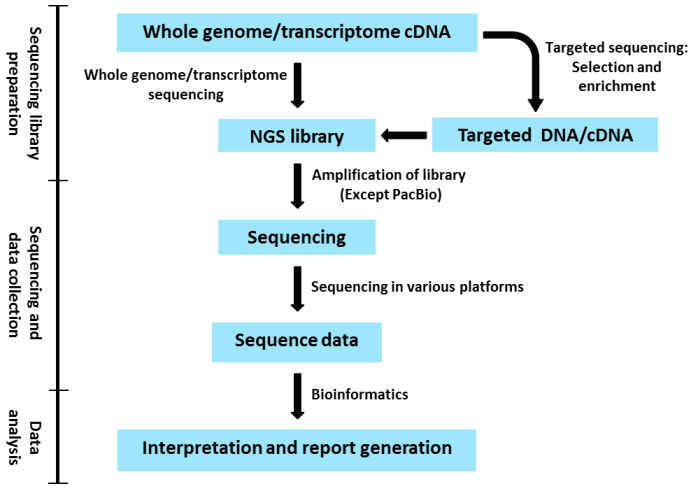


Figure 1. Schematic representation of general working pipeline for next-generation sequencing. This working protocol is typically used in all kind of sequencing approaches.

2. Next generation sequencing

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]

[a] C. Anandhakumar, S. Kizaki, Dr. T. Bando, Prof. Dr. H. Sugiyama
Department of Chemistry, Graduate School of Science
Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan
Fax: (+81) 75-753-3670
E-mail: hs@kuchem.kyoto-u.ac.jp
Homepage: http://kuchem.kyoto-u.ac.jp/chembio

[b] Dr. G. N. Pandian, Prof. Dr. H. Sugiyama
Institute for Integrated Cell-Material Sciences (iCeMS)
Kyoto University, Yoshida-ushinomiyacho, Sakyo-ku, Kyoto 606-8501, Japan

Chandran AnandhaKumar received his Masters in Marine Genomics from MaduraiKamaraj University, India. Then, he Joined as a Senior Research Fellow in Madras Veterinary College. Now, he joined Professor Sugiyama's Chemical Biology group of Kyoto University as a Doctoral student and studying on global gene expression profililing in small molecule-mediated cellular reprogramming.

Seiichiro Kizaki received his master degree from Kyoto University in 2014 and is currently Doctoral student at Kyoto University. He works under the guidance of Professor Sugiyama, and is working on DNA genome wide DNA epigenetic modifications and DNA charge transfer.

Toshikazu Bando received his Ph.D. in 1998 with Kozo Shishido at Tokushima University. After Postdoctoral studies at the Scripps Institute with K. C. Nicolaou, he joined at the Institute of Biomaterials and Bioengineering at Tokyo Medical and Dental University as an Assistant Professor with Hiroshi Sugiyama. In 2004, he moved to Kyoto University as an Assistant Professor. Since 2005, he has been an Associate Professor of Chemical Biology at Kyoto University. His area of specialization is to design and synthesize DNA sequence-specific functional molecules.

Ganesh N. Pandian received his PhD in Biotechnology in 2009 from Niigata University with Monbukagakusho (Japanese Government) scholarship under the guidance of Prof. Hidetaka Hori. He continued his research with Prof. Hori as Assistant Professor (Research). Currently, he has been working at Institute for Integrated Cell-Material Sciences (WPI-iCeMS) Kyoto University as Assistant Professor with Prof. Hiroshi Sugiyama to develop artificial genetic switches for stem cell control.

Hiroshi Sugiyama received his Ph.D. in 1984 with Teruo Matuura at Kyoto University. After Postdoctoral studies at the University of Virginia with Sidney M. Hecht, he returned to Kyoto University in 1986 as an Assistant Professor and became an Associate Professor in 1993. In 1996, he joined the Institute of Biomaterials and Bioengineering at Tokyo Medical and Dental University. He has been a Professor of Chemical Biology at Kyoto University since 2003. Among the honours he has received are the Nippon IBM Award and the Chemical Society of Japan Award for Creative Work.

2.1. 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.

2.1.1. 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 (Figure 2b).[10a,10b,10c,20] 3′-Blocking terminators such as 3′-O-azidomethyl-dNTPs (Figure 2e)[10d] 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.

2.1.2. Roche 454

The Roche 454 sequencing platform utilizes a type of SBS method called pyrosequencing to determine the nucleotide incorporation during sequencing.[11] In pyrosequencing (Figure 2a), 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 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.[12,13] 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.[14]

2.1.3. Ion Torrent

Ion semiconductor sequencing[15] 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

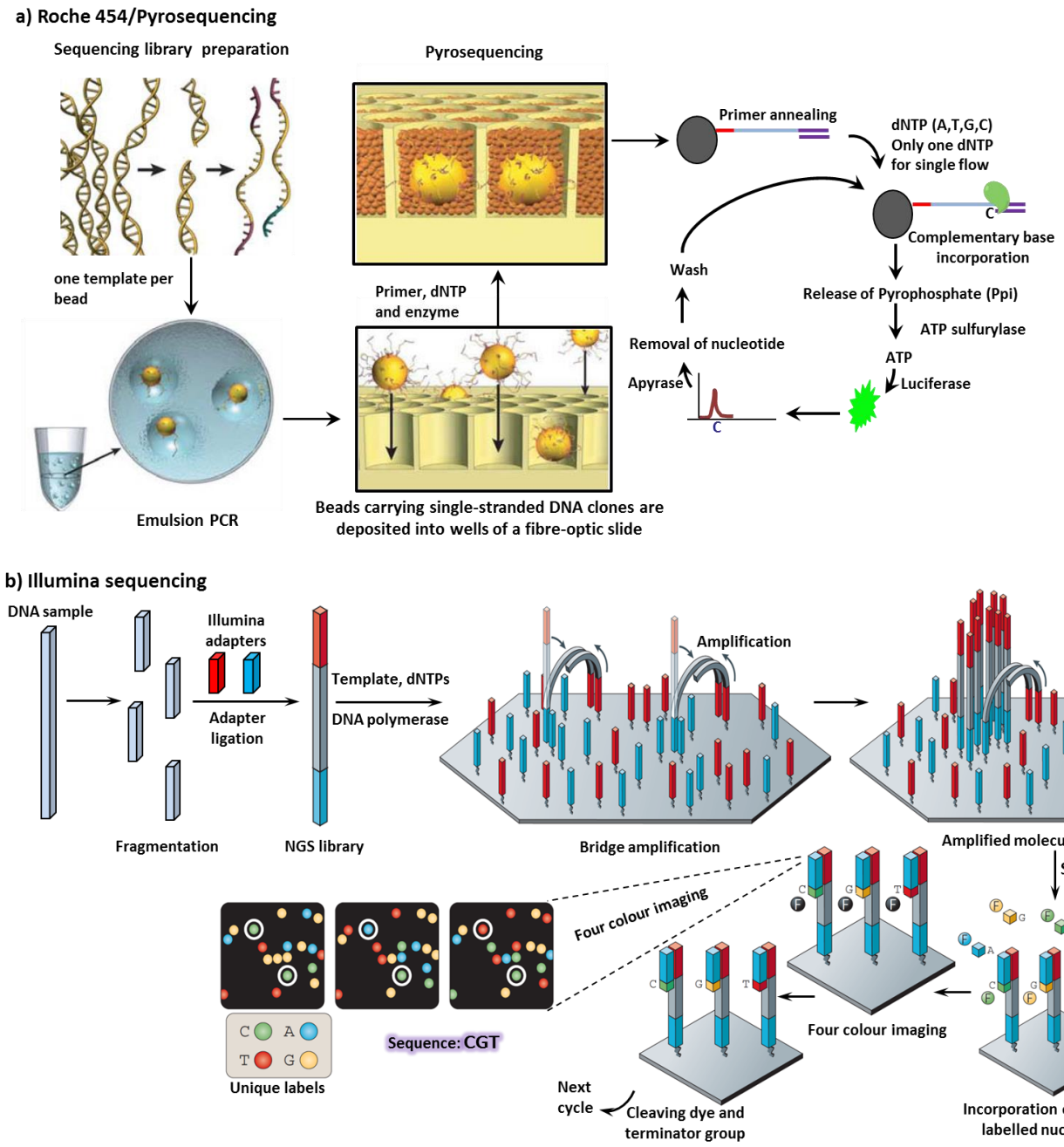
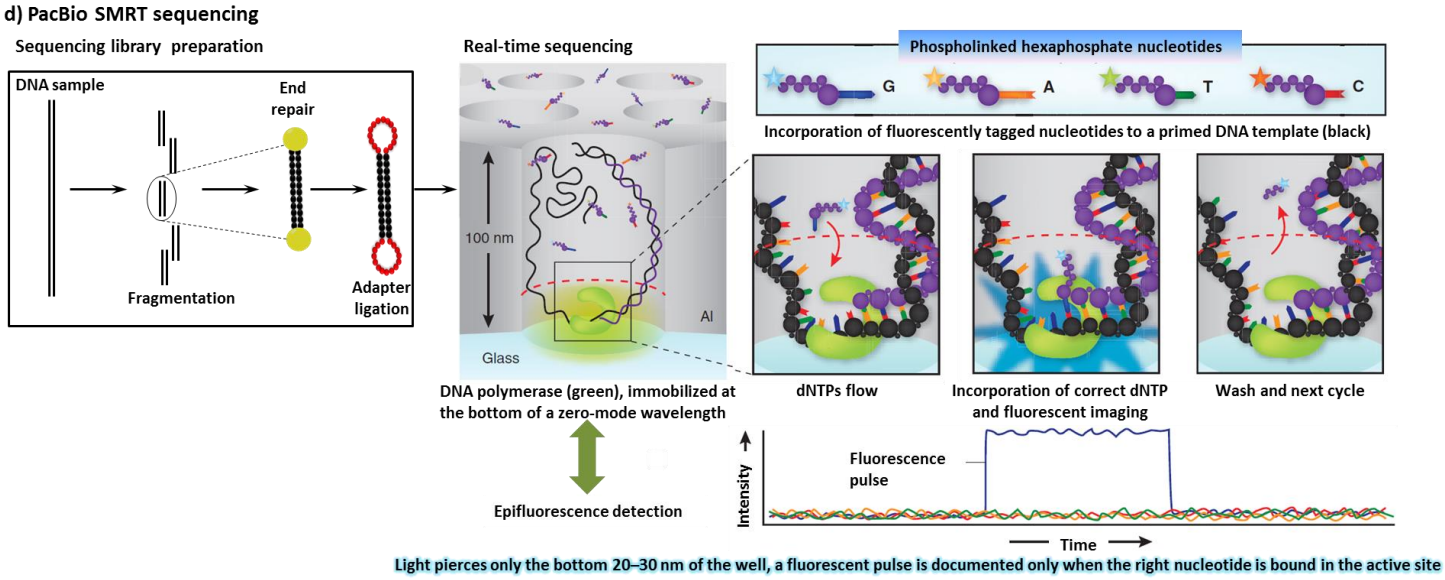
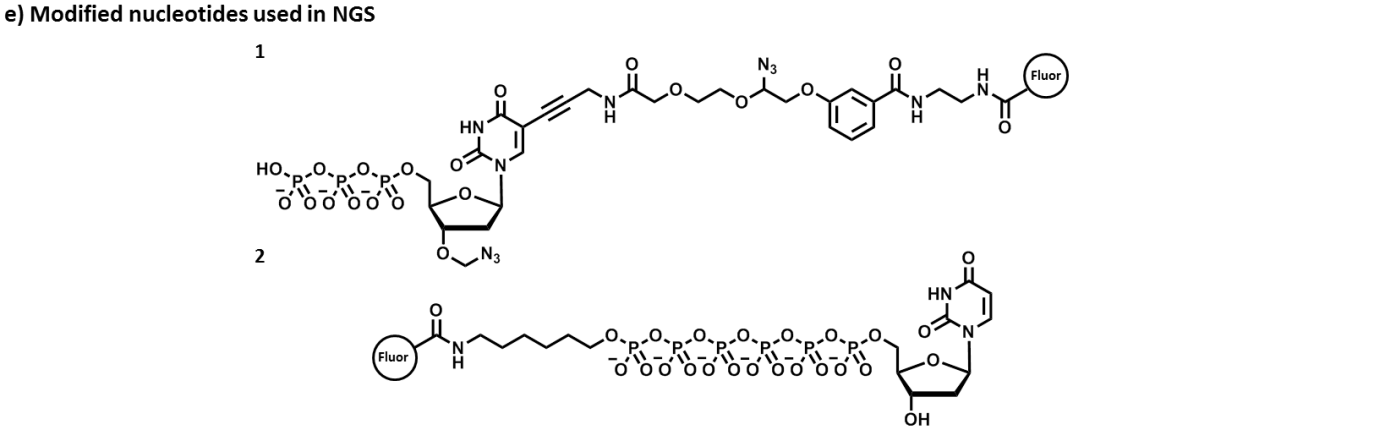


Figure 2. Major sequencing platform's sequencing chemistry. a) Roche 454: Emulsion PCR amplification followed by pyro sequencing[11] b) Illumina sequencing: Bridge amplification followed by sequencing by synthesis.[10a,10b,10c,20] c) Ion torrent: Same as B till the emulsion PCR next the enriched libraries sequenced by polymerization[16] d) PacBIO: Single-molecule real-time[18a,18b,20] (SMRT) sequencing. e) 1. Modified nucleotides used in Illumina[10a,10b,10c,20] and 2. terminally labelled polyphosphate used in PacBio sequencing,[18a] these nucleotides are added specifically to the template DNA, after imaging cleaved efficiently, and stretched as modified or natural nucleotide in next cycles.





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 sensor, which can detect the nucleotide by converting the electric signal into base sequences (Figure 2c). 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 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.

2.2. 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[16]. 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.

2.2.1. Helicos

This was the first commercially available SMS system in the NGS market.[17] 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.

2.2.2. Pacific BioSciences

Pacific BioSciences released a new single-molecule real-time sequencing (SMRT) technology (Figure 2d).[18a,18b] 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)[19] with a single DNA template. By supplying uniquely fluorescently labeled nucleotides (A, T, G, and C), 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.[20] The preparation of a circularized template can enable repeated sequencing of the template and an increased base accuracy.

2.3. 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.[21] 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.

2.3.1. Polonator and SOLiD (support oligonucleotide ligation detection)

SOLiD system[22] is based on the polonator technology[23] (http://www.azcobiotech.com/instruments/polonator.php). 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 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. The sequencing chemistries of these four platforms are explained and illustrated in Figure 2.

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. In this review, some of the key methodologies and analytical strategies are discussed.

3. 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, left-handed Z-DNA, cruciform, and others.[24,25] The G-quadruplex, formed by Hoogsteen hydrogen bonds, is one of the most significant DNA

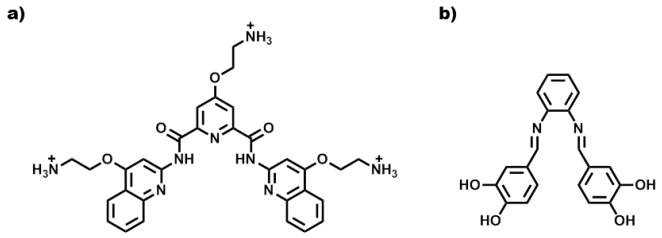


Figure 3. Small molecules used to identify non B-DNA: a) pyridostatin (PDS). b) Schiff-base catechol derivatives.[34,37]

structures, and is always formed in a G-rich region in the presence of some monovalent cations.[26-28] 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.[29]

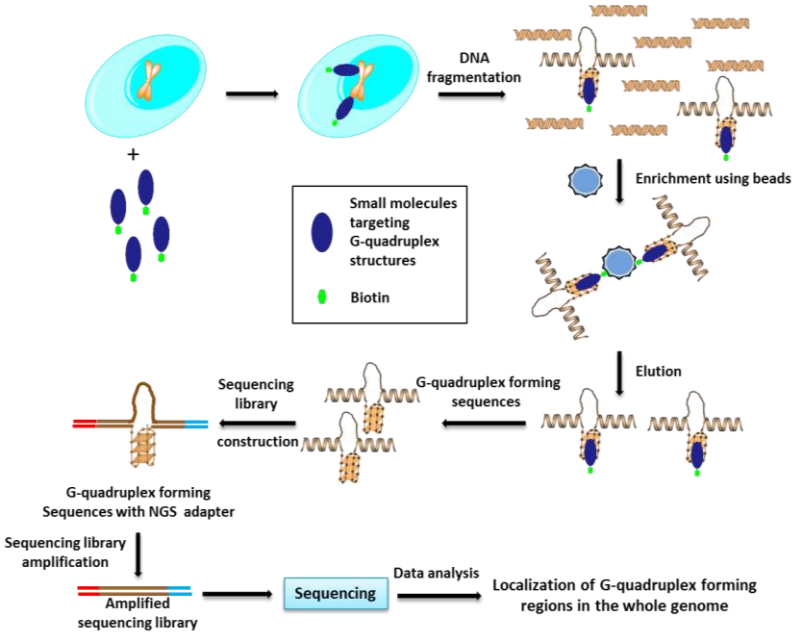


Figure 4. Experimental procedure for detecting G-quadruplex forming region and sequencing.[37]

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.[30,31] G-quadruplex-associated small molecules like pyridostatin have been shown to localize the G-quadruplex in cells[32] and have been used to enrich human telomeric DNA (Figure 3a).[33] we have reported it in realtime observation.[34] 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*.[35] 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.[36] 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 results in weakened efficacy. Yuan *et al*.[37] 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 (Figure 3b) 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.[34,38] Figure 4 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.

4. Analysis of gene expression induced by small molecules

4.1. 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.[40-42] 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 (Figure 5). 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),[43] hypoxia inducible factor 1 alpha (HIF-1α), [44] the glucocorticoid

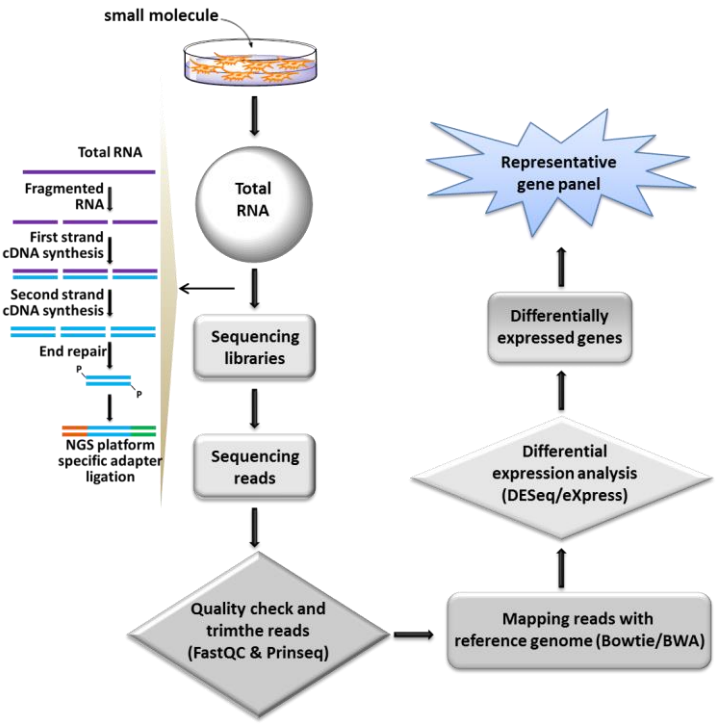


Figure 5. Schematic representation of the pipeline for RNA-Seq analysis of small molecule-regulated RNA.

receptor (GR),[45] and nuclear factor kappa B (NF-κB)[46] 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*.[47] 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.[44] These transcriptome studies demonstrate the crucial importance of deep-sequencing strategies for the rapid validation of small-molecule potency.

4.2. 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.[49-52] 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[53] reviews current studies using the transcription-factor-binding ChIP together with high-throughput sequencing and its full downstream analysis pipeline. Figure 6 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 (e.g., 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[54,33] known to inhibit the growth of human cancer cells by inducing replication/transcription dependent on DNA damage. Rodriguez *et al.*[55] used ChIP-Seq methods to analyze genome-wide pyridostatin-induced DNA damage with the DNA damage-marker protein

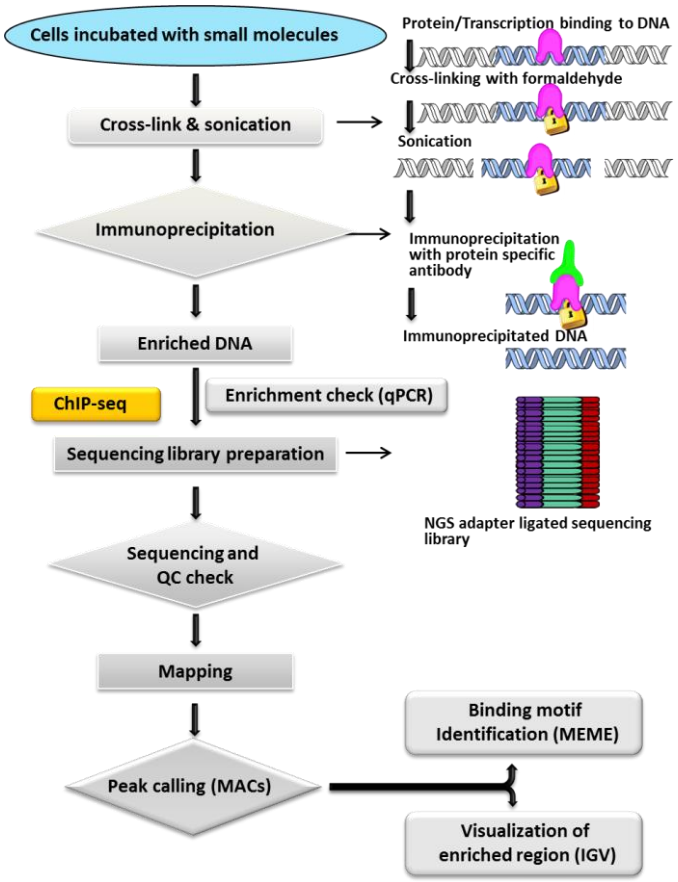


Figure 6. Complete work flow of ChIP-Seq and its downstream analysis pipeline

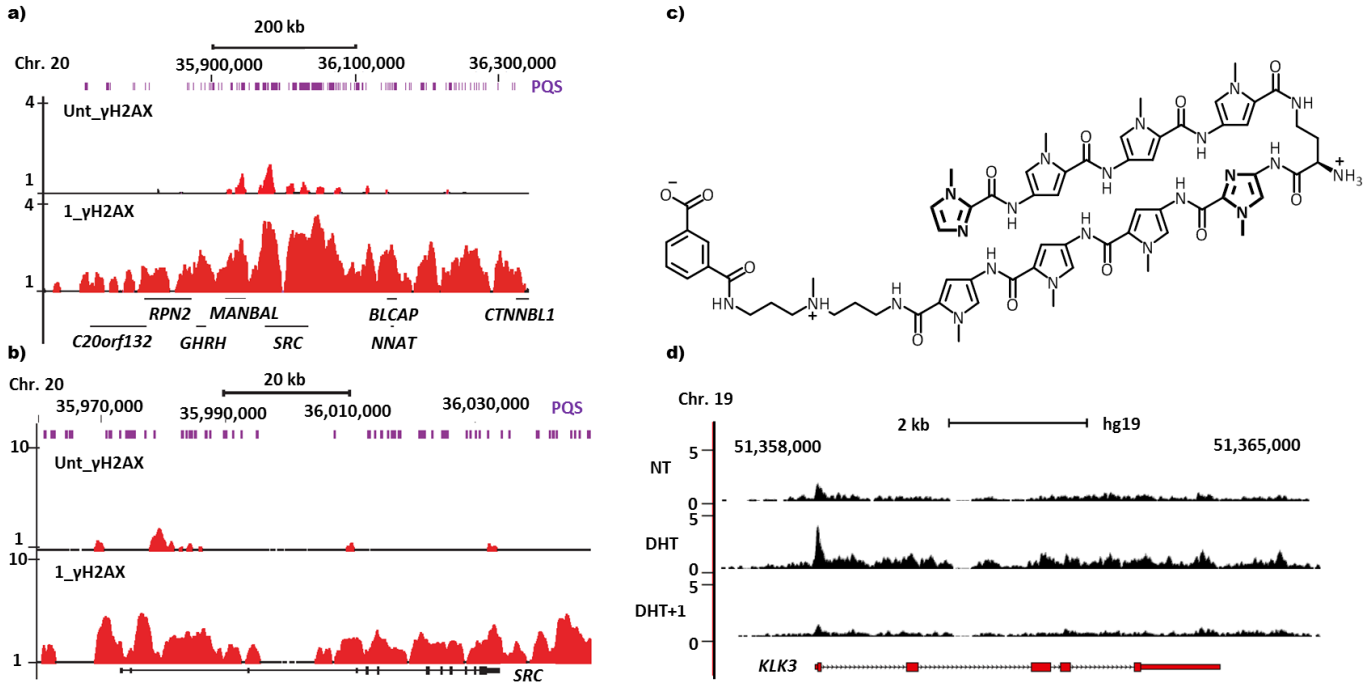


Figure 7. Genome wide mapping of small molecule induced gene. a) ChIP-Seq recognized γH2AX regions containing putative G-quadruplex–forming sequence (PQS) clusters in oncogenes and tumor suppressor genes. MRC MRC-5-SV40 cells treated with 2 μM pyridostatin (1_ γH2AX) compared to untreated MRC MRC-5-SV40 cells (Unt_ γH2AX). b) A zoomed window size showing the enriched sequencing read distribution in the chromosomal region containing *SRC* gene. 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*. 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. [55,58]

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 Figure 7a. This indicates the impact of pyridostatin on gene expression including that of the *SRC* gene (proto-oncogene) (Figure 7b). 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 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.[56,57] Yang *et al*.[58] investigated the effect of PIPs targeted to the RNAP2 transcription machinery. ChIP-Seq was used to map 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 (Figure 7c, 7d). 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 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[59] 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.

5. 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.[60-63] They have been synthesized against a variety of molecular targets including proteins, small molecules, and cell-surface markers.[64-69] Recently, aptamers have found uses in a wide range of applications including diagnostics, molecular imaging, therapeutics, gene delivery, and drug delivery.[70-76]

The commonly used systematic evolution of ligands by exponential enrichment (SELEX) method has been used to prepare aptamers targeting proteins and small molecules.[60,77,78] It takes months to assess and optimize just a handful of aptamers. NGS can greatly speed up this process. Studies[79,80] 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.*[81] developed 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)).

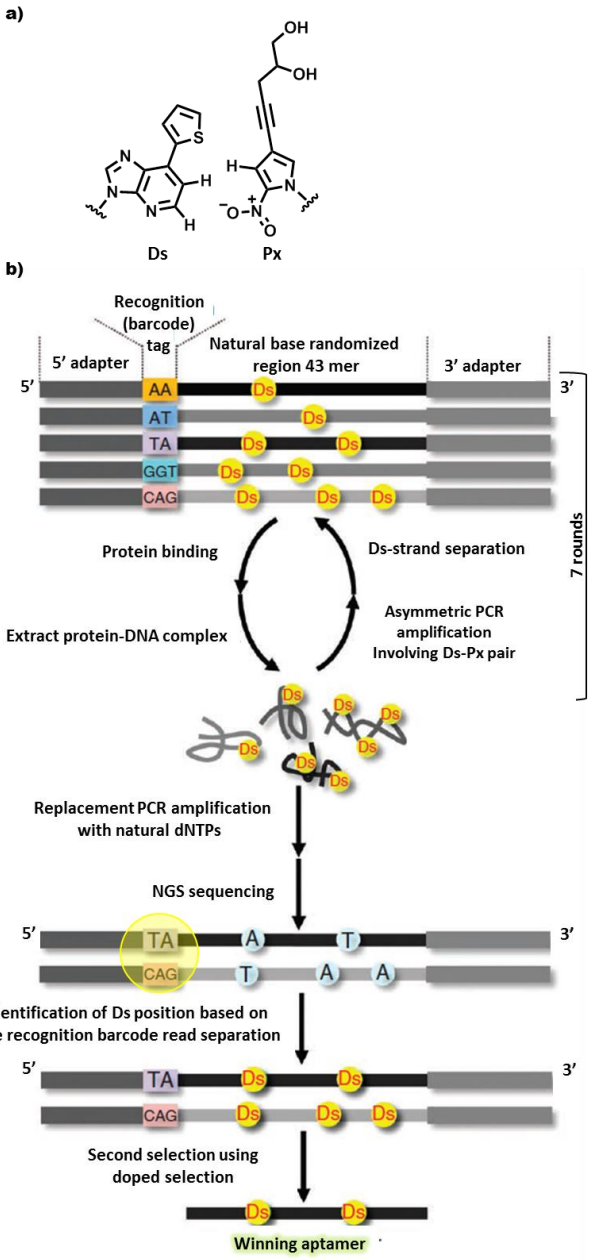


Figure 8. High-affinity DNA aptamers selection using an additional genetic alphabet. a) Structure of an unnatural nucleotide with the hydrophobic base 7-(2-thienyl) imidazo [4,5-b] pyridine (Ds) and its base pairing diol-modified 2-nitro-4-propynylpyrrole (Px). b) design of oligomer library used in the 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.[84]

In some cases, modified natural nucleotides have been incorporated into aptamers to increase their efficiency.[82,83] 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.*[84] used an unnatural nucleotide (Figure 8) 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.

6. Applications of high-throughput sequencing technologies in base modification

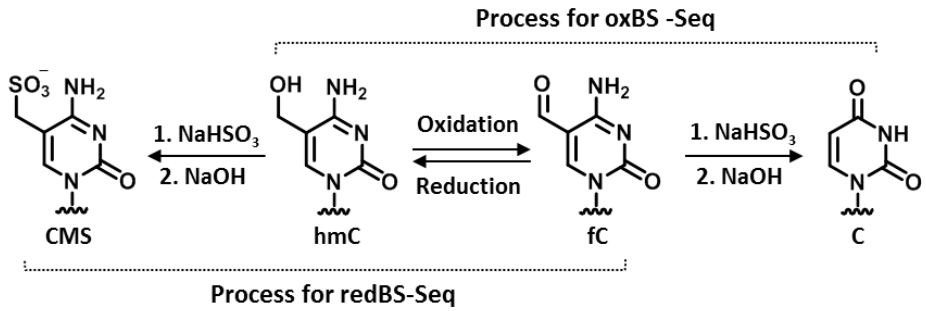
Cancer is known to be a disease process where somatic mutations drive the evolution of more virulent phenotypes. Nonetheless, high-throughput sequencing has unveiled a surprising degree of genetic alteration or base modification.[5] Several studies that compare diseased genomes with healthy ones uncovered 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 to 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.

6.1. Studies targets epigenetically modified bases

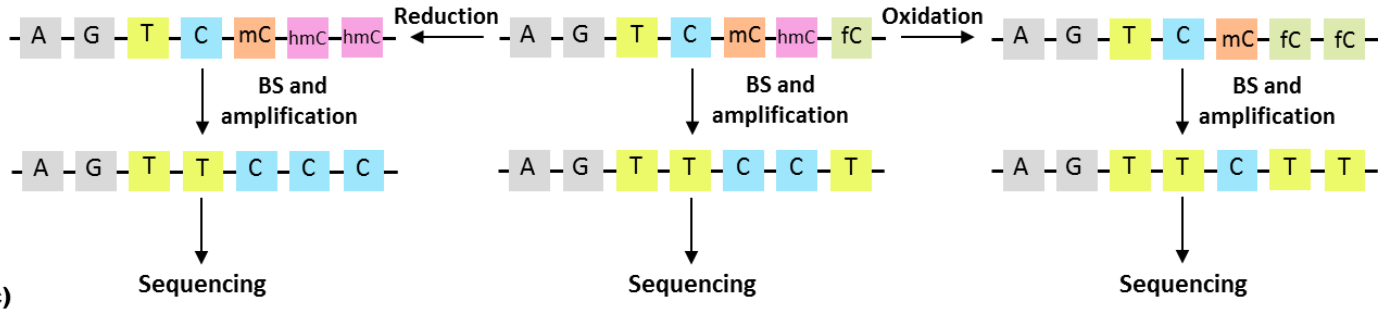
In mammals, genomic DNA 5-methylcytosine (5mC) plays a vital role in variety of biological process through epigenetic gene regulation. It is an epigenetic modification caused by the action of DNMTs. In the progression of several diseases like cancer, the CpG islands of gene promoters become abnormally 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.[85] 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-

a)



b)



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

Figure 9. 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.[86,100]

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).[86-89] 5hmC was discovered in 2009, as another relatively abundant form of cytosine modification.[90,91] It may be an intermediate in active DNA demethylation, but it can also identify an epigenetic mark.[92]

In 2012, a new strategy, "oxidative bisulfite" sequencing (oxBS-Seq), (Figure 9) 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 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.[86]

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.[93-95] 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.[96] 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 (Figure 10). 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).[34]

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.[97]

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:[98] 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.[99] 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) (Figure 9)[100] 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[86] and 5hmC is modified to a cytosine-5- methylsulfonate (CMS) adduct to be detected as C.[89]

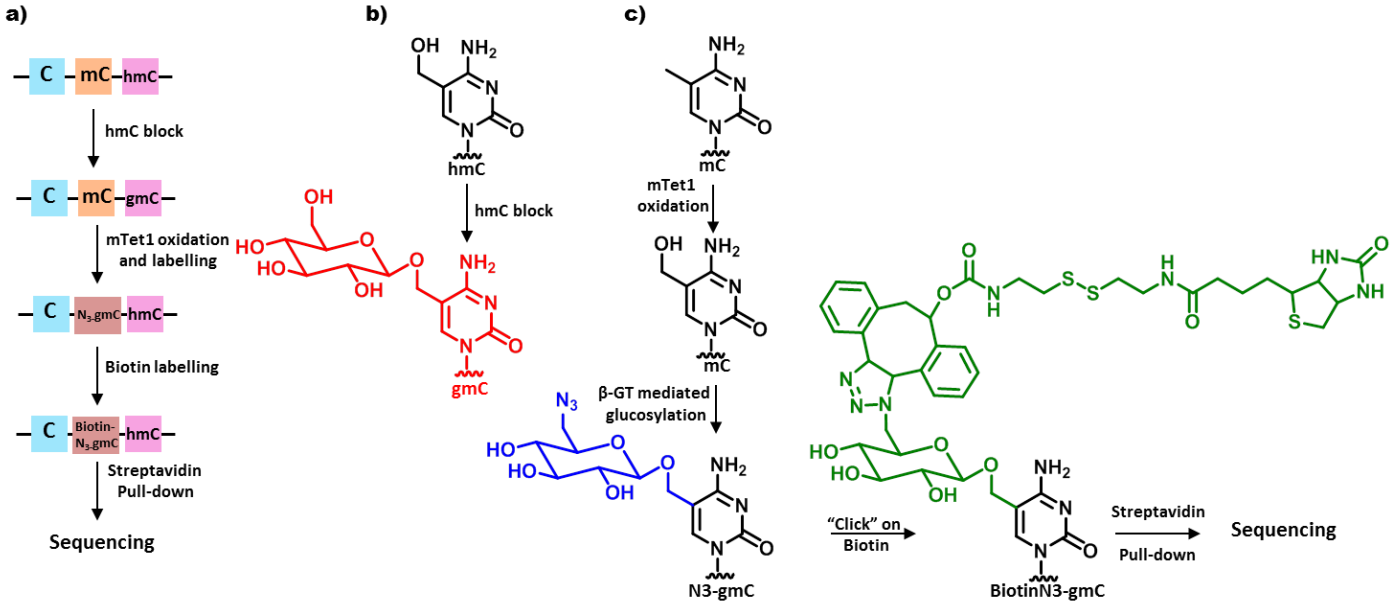


Figure 10. Selective labelling of 5mC in DNA and sequencing using a one-pot mTet1/b-GT protocol. a) Work flow of TAmC-Seq. b) The internal 5hmC in gDNA can be secured by glycosylation with regular glucose. c)The 5mC can be altered to 5hmC by mTet1-catalysed oxidation, and then 6-N3-b-glucosyl-5-hydroxymethyl-cytosine (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.[97]

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.

6.2. DNA strand breakage

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.[101-103] DNA damage happens at a frequency of 1,000 to 1,000,000 molecular cuts per cell per day.[104] The majority of DNA damage disturbs the structure of the Watson & 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*.[105] 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 (Figure 11).

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,[106] as they cannot identify new hotspots 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*.[107] 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).[108] 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.[109] 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.[110] 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.[111-113] 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.[114-116] Crosetto *et al*.[117] 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).

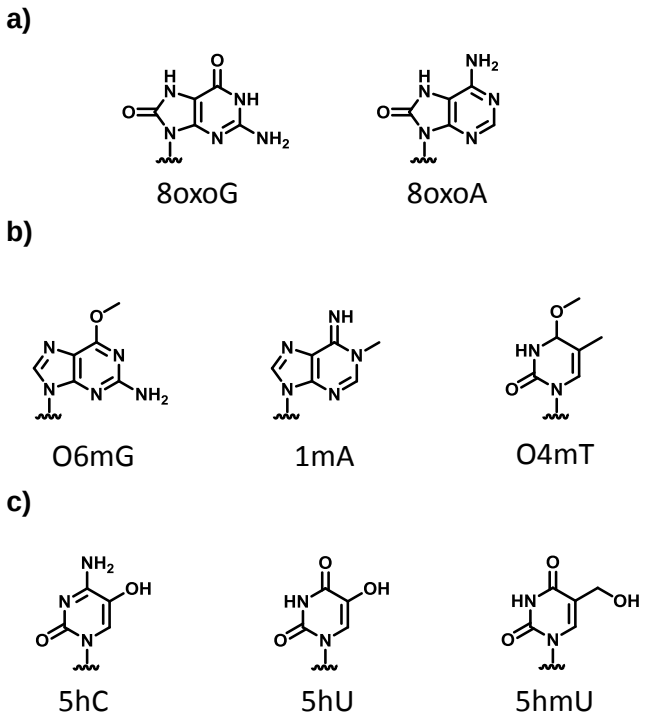


Figure 11. SMRT sequencing and direct identification of Damaged DNA Bases: Modified bases identified using SMART sequencing based on the Kinetic change of polymerase. a) Products of oxidative Damage. b) Products of alkylation DNA damage. c) Products of ionizing radiation DNA damage.[105]

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.*[118] 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.

7. Guiding the design and screening of small molecule using NGS

Small molecules are organic compounds with well-defined chemical structure less than 900 Daltons 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[119] 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.[120] 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.[121] These DNA-binding small molecules have been shown to have strong anti-tumor properties in mouse xenograft models.[58] (anti-tumor activities of PIP was also discussed in 4.1)

PIPs are modifiable synthetic oligomers that can bind to the DNA minor groove based on the DNA recognition rules.[122] This group of small molecules can act as an efficient synthetic gene regulator in either activation[123] or inhibition[124] 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.*[125a] 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 (Figure 12a, 12b). The study followed the methodologies of Bind-n-seq[126] 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.[127] 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)[128] and microarray-based binding site identification.[129-131] This unbiased technique revealed unanticipated binding/orientation sites of the PIPs, which was useful in improving the study of sequence-selective inhibition of CpG methylation.[125b] In this study, PIPs targeting the sequence 5′-CGCG-3′[125c] were used. Bind-n-seq analysis could discern structure–activity relationships and could guide the authors in designing improved CpG methylation antagonists (Figure 12b). Recent studies substantiate the potential of PIP conjugates in gene regulation.[132a, 132b]. 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.

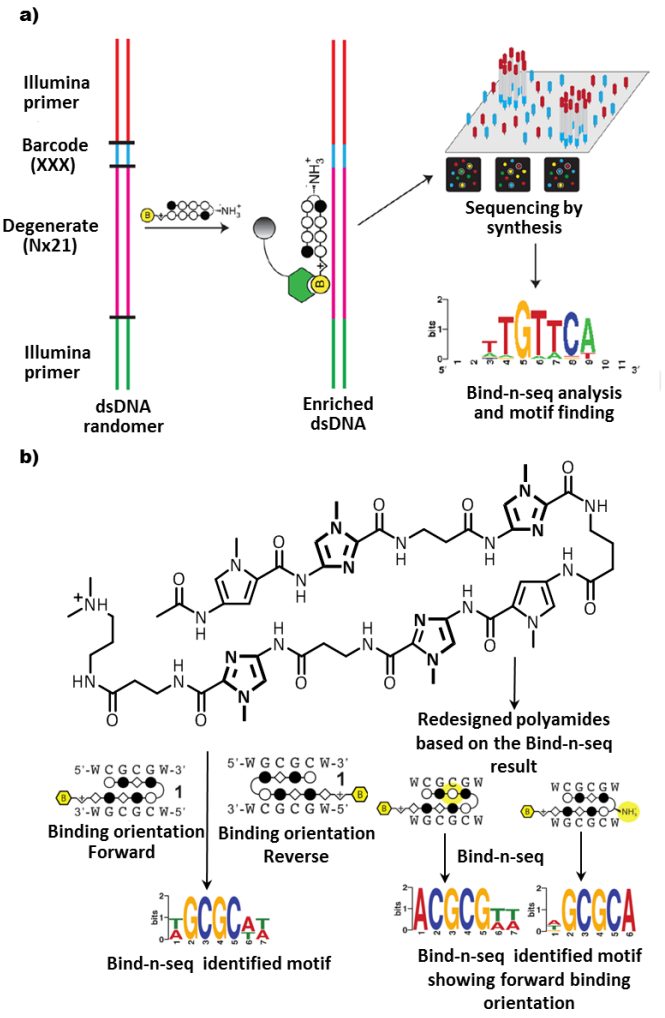


Figure 12. High-throughput sequencing guided designing and validation of DNA binding small molecule. a) Workflow for Bind-n-Seq analysis of Py-Im polyamides recognition motif identification. (21bp of degenerate dsDNA is enriched, affinity purified, and sequenced via high-throughput sequencing for binding site identification). b) Methylated 5-CGCG-3 targeted Py-Im polyamide structure (Polyamide could potentially bind in the forward orientation or the reverse orientation), Bind-n-seq analysis of polyamide showed the reverse orientation binding 5'-GCGC-3 and possible modification in the Py-Im polyamide modified Py-Im polyamide (Positions are highlighted in yellow) and Bind-n-seq analysis of redesigned polyamide showing forward orientation of binding 5'-CGCG-3'.[125a,125b]

8. Application of NGS in small-molecule chemical biology and its future prospects

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,[133] 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. As can be seen from the studies described above, the NGS technologies have extended to an impressive array of applications beyond just genomic sequencing and its large-scale performance in chemical biological research (Table 2).

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 daily.[134] 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 usage of NGS in small-molecule chemistry will be a great milestone in the field of chemical biology.

9. Summary and Outlook

In this review, we have made an effort to summarize and define the importance of multitasking NGS technologies in chemical biological applications as well as to describe the essential methodologies and bioinformatics analysis pipeline that make them possible. 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 chemical biology[59] (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. For example, ChIP-Seq is a commonly used technique to map protein-binding regions in the genome: S. Balasubramanian *et al.* group mapped G-quadruplex regions in the genome using advanced ChIP-Seq, which offers a protocol for G-quadruplex DNA sequencing.[35] Notably, the genomic targets of the G-quadruplex-associated small molecule pyridostatin were identified using ChIP-Seq.[55] In the case of gene regulation, RNA-Seq analysis of PIP-treated cells suggested the impact of PIP on transcription regulation.[47] It is also important to design PIPs in an established sequence-specific manner to precisely control gene regulation. The NGS platform provides a method, Bind-n-seq, that permits quick, quantitative detection of high-affinity PIP-binding regions.[125a] Using this unbiased technique, unanticipated PIP-binding modes were identified, so that NGS guided improved design of CpG methylation antagonists that showed high performance with specific binding in a targeted orientation.[125a] NGS utilization has reduced the long time scale required for aptamer screening and has also improved the selection method using unnatural bases.[84] Most importantly, effective approaches in epigenetic modification analysis have been developed with NGS. It is virtually impossible to assess DNA methylation and various moieties of cytosine on a genomic scale without high-throughput sequencing.[86,96-98,100] The available sequencing platforms are multipurpose, accurate, and capable of efficiently producing data. At the same time, novel methods that are evolving day-by-day suggest new sequencing pipelines that are built in different ways from a combination of the building blocks of existing strategies and protocols. Our future hope will be to continue this trend in new directions towards the development of rapid, cost-effective chemical biological applications through information generated by NGS. Thus, established chemical biology combined with NGS holds immense promise.

Received: ((will be filled in by the editorial staff))
Published online on ((will be filled in by the editorial staff))

Keywords:
Next-generation sequencing · Chemical biology · Small molecule · Bind-n-seq · Aptamer

[1] F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467.

[2] A. M. Maxam, W. Gilbert, *Proc.Natl. Acad. Sci. USA* **1977**, *74*, 560–564.

[3] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, et al., *Nature* **2008**, *452*, 872–876.

[4] International Human Genome Sequencing Consortium, *Nature* **2004**, *43*, 931–945.

[5] M. Meyerson, S. Gabriel, G. Getz, *Nat. Rev.Genet.* **2010**, *11*, 685–696.

[6] P. Hui, *Top. Curr. Chem.* **2012**, *336*, 1-18.

[7] P. J. Hurd, C. J. Nelson, *Briefings in functional genomics & proteomics* **2009**, *8*, 174–183.

[8] P. J. Park, *Nat. Rev.Genet.* **2009**, *10*, 669−680.

[9] A. N. Egan, J. Schlueter, D. M. Spooner, *American Journal of Botany* **2012**, *99*, 175–185.

[10] a) D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al., *Nature* **2008**, *456*, 53–59; b) C. Barnes, S. Balasubramanian, X. Liu, H. Swerdlow, J. Milton, Labelled nucleotides. US Patent **2002**, 7,057,026; c) M. L. Metzker, *Genome Res.* **2005**, *15*, 1767–1776; d) J. Guo, N. Xu, Z. Li, S. Zhang, J. Wu, D. H. Kim, M. Sano Marma, Q. Meng, H. Cao, X. Li, et al., *Proc.Natl. Acad. Sci. USA* **2008**, *105*, 9145–9150.

[11] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al., *Nature* **2005**, *437*, 376–380.

[12] M. Ronaghi, M. Uhlén, P. Nyrén, *Science (New York, N.Y.)* **1998**, *281*, 363, 365.

[13] A. Ahmadian, M. Ehn, S. Hober, *Clinica Chimica Acta* **2006**, *363*, 83–94.

[14] A. M. Smith, L. E. Heisler, R. P. St Onge, E. Farias-Hesson, I. M. Wallace, J. Bodeau, A. N. Harris, K. M. Perry, G. Giaever, N. Pourmand, et al., *Nucleic Acids Res.* **2010**, *38*, e142.

[15] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, et al., *Nature* **2011**, *475*, 348–352.

[16] F. Ozsolak, A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, P. M. Milos, *Nature* **2009**, *461*, 814–818.

[17] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J. W. Efcavitch, et al., *Science (New York, N.Y.)* **2008**, *320*, 106–109.

[18] a) J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank,P. Baybayan, B. Bettman, et al., *Science* **2009**, *323*, 133-138; b) M. L. Metzker, *Nature NEWS and VIEWS* **2009**, *27*, 150-151.

[19] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, W. W. Webb, *Science (New York, N.Y.)* **2003**, *299*, 682–686.

[20] M. L. Metzker, *Nat. Rev.Genet.* **2010**, *11*, 31–46.

[21] U. Landegren, R. Kaiser, J. Sanders, L. Hood, *Science (New York, N.Y.)* **1988**, *241*, 1077–1080.

[22] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, et al., *Genome Res.* **2008**, *18*, 1051–1063.

[23] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, G. M. Church, *Science (New York, N.Y.)* **2005**, *309*, 1728–1732.

[24] S. M. Mirkin, *Front. Biosci*. **2008**, *13*, 1064–1071.

[25] R. D. Wells, *J. Biol. Chem.* **1988**, *263*, 1095–1098. A. M. Zahler, J. R. Williamson, T. R. Cech, D. M. Prescott, *Nature* **1991**, *350*, 718–720.

[26] A. M. Zahler, J. R. Williamson, T. R. Cech, D. M. Prescott, *Nature* **1991**, *350*, 718–720.

[27] R. Rodriguez, G. D. Pantos, D. P. N. Goncalves, J. K. M. Sanders, S. Balasubramanian, *Angew. Chem. Int. Ed.* **2007**, *46*, 5405–5407.

[28] X. L. Wang, J. Huang, Y. Y. Zhou, S. Y. Yan, X. C. Weng, X. J. Wu, M. G. Deng, X. Zhou, *Angew. Chem. Int. Ed.* **2010**, *49*, 5305–5309.

[29] S. Balasubramanian, S. Neidle, *Curr. Opin. Chem. Biol.* **2009**, *13*, 345–353.

[30] A. Bugaut, K. Jantos, J. L. Wietor, R. Rodriguez, J. K. M. Sanders, S. Balasubramanian, *Angew.Chem. Int. Ed.* **2008**, *47*, 2677–2680.

[31] S. Balasubramanian, L. H. Hurley, S. Neidle, *Nature Rev. Drug Discov.* **2011**, *10*, 261–275.

[32] C. Granotier, G. Pennarun, L. Riou, F. Hoffschir, L. R. Gauthier, A. De Cian, D. Gomez, E. Mandine, J.-F. Riou, J.-L. Mergny, et al., *Nucleic Acids Res.* **2005**, *33*, 4182–4190..

[33] S. Mu¨ller, S. Kumari, R. Rodriguez, S. Balasubramanian, *Nat. Chem.* **2010**, *2*, 1095–1098.

[34] A. Rajendran, M. Endo, K. Hidaka, P. L. Thao Tran, M.-P. Teulade-Fichou, J.-L. Mergny, H. Sugiyama, *RSC Adv.* **2014**, *4*, 6346.

[35] E. Y. N. Lam, D. Beraldi, D. Tannahill, S. Balasubramanian, *Nat. Commun.* **2013**, *4*, 1796.

[36] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, et al., *Genome biol.* **2008**, *9*, R137.

[37] L. Yuan, T. Tian, Y. Chen, S. Yan, X. Xing, Z. Zhang, Q. Zhai, L. Xu, S. Wang, X. Weng, B. Yuan, Y. Feng, X. Zhou, *Sci. Rep.* **2013**, *3*, 1811.

[38] C.-X. Song, K. E. Szulwach, Y. Fu, Q. Dai, C. Yi, X. Li, Y. Li, C.-H. Chen, W. Zhang, X. Jian, et al., *Nat. Biotechnol.* **2011**, *29*, 68–72.

[39] A. Jacquier, *Nat. Rev.Genet.* **2009**, *10*, 833–844.

[40] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, J. Bähler, *Nature* **2008**, *453*, 1239–1243.

[41] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat. Methods* **2008**, *5*, 621–628.

[42] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. M. Jones, A. Marco, *Biotechniques* **2008**, *45*, 81–94.

[43] N. G. Nickols, P. B. Dervan, *Proc.Natl. Acad. Sci. USA* **2007**, *104*, 10418–10423.

[44] B. Z. Olenyuk, G. J. Zhang, J. M. Klco, N. G. Nickols, W. G. Kaelin, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16768–16773.

[45] K. A. Muzikar, N. G. Nickols, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 16598–16603.

[46] J. A. Raskatov, J. L. Meier, J. W. Puckett, F. T. Yang, P. Ramakrishnan, P.B. Dervan, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1023–1028.

[47] J. A. Raskatov, N. G. Nickols, A. E. Hargrove, G. K. Marinov, B. Wold, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16041–16045.

[48] N. G. Nickols, J. O. Szablowski, A. E. Hargrove, B. C. Li, J. A. Raskatov, P. B. Dervan, *Mol. Cancer Ther.* **2013**, *12*, 675–684.

[49] J. Han, P. Yuan, H. Yang, J. Zhang, B. S. Soh, P. Li, S. L. Lim, S. Cao, J. Tay, Y. L. Orlov, et al., *Nature* **2010**, *463*, 1096–1100.

[50] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, et al., *Cell* **2008**, *133*, 1106–1117.

[51] J. C. D. Heng, B. Feng, J. Han, J. Jiang, P. Kraus, J. H. Ng, Y. L. Orlov, M. Huss, L. Yang, T. Lufkin, et al., *Cell Stem Cell* **2010**, *6*, 167–174.

[52] P. Yuan, J. Han, G. Guo, Y. L. Orlov, M. Huss, Y. H. Loh, L. P. Yaw, P. Robson, B. Lim and H. H. Ng, *Genes and Dev.* **2009**, *23*, 2507-2520.

[53] T. S. Furey, *Nat. Rev.Genet.* **2012**, *13*, 840–852.

[54] R. Rodriguez, S. Müller, J. A. Yeoman, C. Trentesaux, J.-F. Riou, S. Balasubramanian, *J. Am. Chem. Soc.* **2008**, *130*, 15758–15759.

[55] R. Rodriguez, K. M. Miller, J. V Forment, C. R. Bradshaw, M. Nikan, S. Britton, T. Oelschlaegel, B. Xhemalce, S. Balasubramanian, S. P. Jackson, *Nat. Che. Biol.* **2012**, *8*, 301–310.

[56] D. M. Chenoweth, P. B. Dervan, *J. Am. Chem. Soc.* **2010**, *132*,14521–14529.

[57] D. M. Chenoweth, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2009**, *106*,13175–13179.

[58] F. Yang, N. G. Nickols, B. C. Li, G. K. Marinov, J. W. Said, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1863–1868

[59] L. Han, G. N. Pandian, S. Junetha, S. Sato, C. Anandhakumar, J. Taniguchi, A. Saha, T. Bando, H. Nagase, H. Sugiyama, *Angew. Chem. Int. Ed.* **2013**, 52, 13410-13413.

[60] C. Tuerk, L. Gold, *Science* **1990**, *249*, 505–510.

[61] T. Mairal, V. C. Ozalp, P. Lozano Sánchez, M. Mir, I. Katakis, C. K. O'Sullivan, *Anal. Bioanal. Chem.* **2008**, *390*, 989–1007.

[62] A. D. Ellington, J. W. Szostak, *Nature* **1990**, *346*, 818–822.

[63] S. D. Jayasena, *Clin. Chem.* **1999**, *45*, 1628–1650.

[64] S. E. Lupold, B. J. Hicke, Y. Lin, D. S. Coffey, *Cancer Res.* **2002**, *62*, 4029–4033.

[65] D. E. Huizenga, J. W. Szostak, *Biochemistry* **1995**, *34*, 656–665.

[66] L. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermaas, J. J. Toole, *Nature* **1992**, *355*, 564–566.

[67] P. Mallikaratchy, Z. Tang, S. Kwame, L. Meng, D. Shangguan, W. Tan, *Mol. Cell. Proteomics* **2007**, *6*, 2230–2238.

[68] L. S. Green, D. Jellinek, R. Jenison, A. Ostman, C. H. Heldin, N. Janjic, *Biochemistry* **1996**, *35*, 14413–14424.

[69] D. Mann, C. Reinemann, R. Stoltenburg, B. Strehlitz, *Biochem. Biophys. Res. Commun.* **2005**, *338*, 1928–1934.

[70] J. Zhou, H. Li, S. Li, J. Zaia, J. J. Rossi, *Mol. Ther.* **2008**, *16*, 1481–1489.

[71] J. S. Swensen, Y. Xiao, B. S. Ferguson, A. A. Lubin, R. Y. Lai, A. J. Heeger, K. W. Plaxco, H. T. Soh, *J. Am. Chem. Soc.* **2009**, *131*, 4262–4266.

[72] G. J. Tong, S. C. Hsiao, Z. M. Carrico, M. B. Francis, *J. Am. Chem. Soc.* **2009**, *131*, 11174–11178.

[73] W. Li, X. Yang, K. Wang, W. Tan, Y. He, Q. Guo, H. Tang, J. Liu, *Anal. Chem.* **2008**, *80*, 5002–5008.

[74] H. W. Chen, C. D. Medley, K. Sefah, D. Shangguan, Z. Tang, L. Meng, J. E. Smith, W. Tan, *Chem. Med. Chem.* **2008**, *3*, 991–1001.

[75] Y. Xiao, A. A. Lubin, A. J. Heeger, K. W. Plaxco, *Angew. Chem. Int. Ed.* **2005**, *44*, 5456–5459.

[76] J. O. McNamara, D. Kolonias, F. Pastor, R. S. Mittler, L. Chen, P. H. Giangrande, B. Sullenger, E. Gilboa, *J. Clin. Invest.* **2008**, *118*, 376–386.

[77] S. M. Nimjee, C. P. Rusconi, B. A. Sullenger, *Annu. Rev. Med.* **2005**, *56*, 555–583.

[78] M. Famulok, *Curr. Opin. Struct. Biol.* **1999**, *9*, 324–329.

[79] M. Cho, Y. Xiao, J. Nie, R. Stewart, A. T. Csordas, S. S. Oh, J. A. Thomson, H. T. Soh, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 15373–15378.

[80] G. V Kupakuwana, J. E. Crill, M. P. McPike, P. N. Borer, *PLoS One* **2011**, *6*, e19395.

[81] S. Hoon, B. Zhou, K. D. Janda, S. Brenner, J. Scolnick, *Biotechniques* **2011**, *51*, 413–416.

[82] L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, J. Carter, A. B. Dalby, B. E. Eaton, T. Fitzwater, et al., *PLoS One* **2010**, *5*, e15004.

[83] J. A. Latham, R. Johnson, J. J. Toole, *Nucleic Acids Res.* **1994**, *22*, 2817–2822.

[84] M. Kimoto, R. Yamashige, K. Matsunaga, S. Yokoyama, I. Hirao, *Nat. Biotechnol.* **2013**, *31*, 453–457.

[85] S. Beck, V. K. Rakyan, *Trends Genet.* **2008**, *24*, 231–237.

[86] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, *Science* **2012**, *336*, 934–937.

[87] S. G. Jin, S. Kadam, G. P. Pfeifer, *Nucleic Acids Res.* **2010**, *38*, e125.

[88] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, J. R. Ecker, *Cell* **2008**, *133*, 523–536.

[89] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, A. Rao, *PLoS One* **2010**, *5*, e8888.

[90] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, et al., *Science* **2009**, *324*, 930–935.

[91] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930.

[92] M. R. Branco, G. Ficz, W. Reik, *Nature Reviews. Geneics* **2012**, *13*, 7-13.

[93] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300–1303.

[94] Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, et al., *Science* **2011**, *333*, 1303–1307.

[95] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. a. Deiml, C. Hagemeier, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 7008–7012.

[96] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, et al., *Cell* **2012**, *149*, 1368–1380.

[97] L. Zhang, K. E. Szulwach, G. C. Hon, C.-X. Song, B. Park, M. Yu, X. Lu, Q. Dai, X. Wang, C. R. Street, et al., *Nat. Commun.* **2013**, *4*, 1517.

[98] C. X. Song, K. E. Szulwach, Q. Dai, Y. Fu, S. Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, et al., *Cell* **2013**, *153*, 678–691.

[99] C. G. Spruijt, F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. T. C. Jansen, C. Bauer, M. Munzel, M. Wagner, M. Muller, F. Khan, et al., *Cell* **2013**, *152*, 1146–1159.

[100] M. J. Booth, G. Marsico, M. Bachman, D. Beraldi, S. Balasubramanian, *Nat. Chem.* **2014**, *advance on*, DOI 10.1038/nchem.1893.

[101] N. E. Geacintov, S. Broyde, *Chemical Biology of DNA Damage*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2010,** pp 3-126.

[102] B. D. Preston, T. M. Albertson, A. J. Herr, *Semin. Cancer Biol.* **2010**, *20*, 281–293.

[103] M. R. Kelley, *DNA Repair in Cancer Therapy*, Academic Press, Elsevier Science, **2012**, pp1-16.

[104] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, J. Darnell, *Molecular Biology of the Cell*, WH Freeman, New York, NY. **2004**, 5th ed. p963.

[105] T. A. Clark, K. E. Spittle, S. W. Turner, J. Korlach,*Genome Integr.* **2011,** *2*, 10.

[106] S. Kumari, R. P. Rastogi, K. L. Singh, S. P. Singh, R. P. Sinha in *DNA Damage detection strategies*, EXCLI J **2008** , pp44-62.

[107] F. Leduc, D. Faucher, G. Bikond Nkoma, M.-C. Grégoire, M. Arguin, R. J. Wellinger, G. Boissonneault, *PLoS One* **2011**, *6*, e17353.

[108] Z. X. Zhou, M.-J. Zhang, X. Peng, Y. Takayama, X.-Y. Xu, L.-Z. Huang, L.-L. Du, *Genome Res.* **2013**, *23*, 705–715.

[109] U. H. Mortensen, L. M. R. Rothstein, *Curr. Biol.* **2009**, 19, 676–677.

[110] S. Bekker-Jensen, C. Lukas, R. Kitagawa, F. Melander, M. B. Kastan, J. Bartek, J. Lukas, *J. Cell Biol.* **2006**, *173*, 195–206.

[111] R. K. Szilard, P.-E. Jacques, L. Laramée, B. Cheng, S. Galicia, A. R. Bataille, M. Yeung, M. Mendez, M. Bergeron, F. Robert, et al., *Nat. Struct. Mol. Biol.* **2010**, *17*, 299–305.

[112] J. Seo, S. C. Kim, H.-S. Lee, J. K. Kim, H. J. Shon, N. L. M. Salleh, K. V. Desai, J. H. Lee, E.-S. Kang, J. S. Kim, et al., *Nucleic Acids Res.* **2012**, *40*, 5965–5974.

[113] J. A. Harrigan, R. Belotserkovskaya, J. Coates, D. S. Dimitrova, S. E. Polo, C. R. Bradshaw, P. Fraser, S. P. Jackson, *J. Cell Biol.* **2011**, *193*, 97–108.

[114] T. M. Marti, E. Hefner, L. Feeney, V. Natale, J. E. Cleaver, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9891–9896.

[115] B. P. Chadwick, T. F. Lane, *Chromosoma.* **2005**, *114*, 432–439.

[116] S. Tuduri, L. Crabbé, C. Conti, H. Tourrière, H. Holtgreve-Grez, A. Jauch, V. Pantesco, J. De Vos, A. Thomas, C. Theillet, et al., *Nat. Cell Biol.* **2009**, *11*, 1315–1324.

[117] N. Crosetto, A. Mitra, M. J. Silva, M. Bienko, N. Dojer, Q. Wang, E. Karaca, R. Chiarle, M. Skrzypczak, K. Ginalski, et al., *Nat. Methods* **2013**, *10*, 361–365.

[118] Y. Zhang, R. P. McCord, Y.-J. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, J. Dekker, *Cell* **2012**, *148*, 908–921.

[119] P. J. Declerck, *Generics and Biosimilars Initiative Journal* **2012**, *1*, 13-16.

[120] U. Pindur, M. Jansen, T. Lemster, *Curr Med Chem.* **2005**, *12*, 2805–2847.

[121] J. W. Trauger, E. E. Baird, P. B. Dervan, *Nature* **1996**, *382*, 559–561.

[122] P. Dervan, A. Poulin-Kerstien, E. Fechter, *Top. Curr. Chem.* **2005**, *253*, 1–31.

[123] A. Groth, W. Rocha, A. Verreault, G. Almouzni, *Cell* **2007**, *128*, 721–733.

[124] J. M. Gottesfeld, L. Neely, J. W. Trauger, E. E. Baird, P. B. Dervan, *Nature* **1997**, *387*, 202–205.

[125] a) J. L. Meier, A. S. Yu, I. Korf, D. J. Segal, P. B. Dervan, *J. Am. Chem. Soc.* **2012**, *134*, 17814–17822; b) J. Kang, J. L. Meier, P. B. Dervan, *J. Am. Chem. Soc.* **2014**, *136*, 3687-3694; c) M. Minoshima, T. Bando, S. Sasaki, J. Fujimoto, H. Sugiyama, *Nucleic Acids Res.* **2008**, *36*, 2889-2894.

[126] A. Zykovich, I. Korf , D. J. Segal, *Nucleic Acids Res.* **2009**, *37*, e151.

[127] T. L. Bailey, *Bioinformatics* **2011**, *27*, 1653−1659.

[128] V. Y. N. Gopal, M. W. Van Dyke, *Biochemistry* **2003**, *42*, 6891−6903.

[129] C. L. Warren, N. C. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips, A. Z. Ansari, *Proc.Natl. Acad. Sci. USA* **2006**, *103*, 867−872.

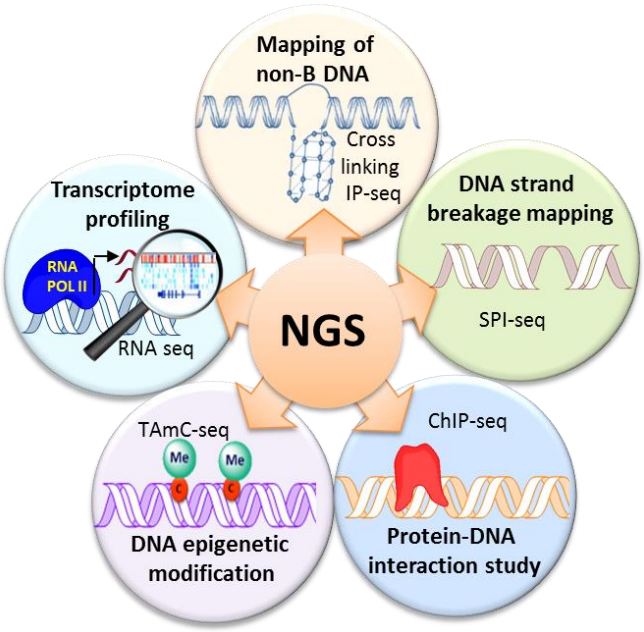
[130] J. W. Puckett, K. A. Muzikar, J. Tietjen, C. L. Warren, A. Z. Ansari, P. B. Dervan, *J. Am. Chem. Soc.* **2007**, *129*, 12310−12319.

[131] C. D. Carlson, C. L. Warren, K. E. Hauschild, M. S. Ozers, N. Qadir, D. Bhimsaria, Y. Lee, F. Cerrina, A. Z. Ansari, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4544−4549.

[132] a) G. N. Pandian, Y. Nakano, S. Sato, H. Morinaga, T. Bando, H. Nagase, H. Sugiyama, *Sci. Rep.* **2012**, *2*, 544; 19; b) G. N. Pandian, J. Taniguchi, S. Junetha, S. Sato, L. Han, A. Saha, C. AnandhaKumar, T. Bando, H. Nagase,T. Vaijayanthi, R. D. Taylor, H. Sugiyama, *Sci. Rep.* **2014**, *4*, 3843.

[133] P. Hou, Y. Li, X. Zhang, C. Liu, J. Guan, H. Li, T. Zhao, J. Ye, W. Yang, K. Liu, et al., *Science* **2013**, *341*, 651–654.

[134] J. Shendure, E. Lieberman Aiden, *Nat. Biotechnol.* **2012**, *30*, 1084–1094.



Next generation sequencing (NGS) is being utilized not only to study the genome assembly and to examine the human diseases in genetic level, but also to explore various kind of utility in small/single molecule chemical biological studies. In which, the applications of NGS is being effectively used in vast contexts, including small molecule/aptamer design and screening, hmC/mC analysis, DNA strand breakage and non-B DNA identification. In this review, we summarize the recent progress in the area of chemical biology using Next generation sequencing.

*Chandran Anandhakumar, Seiichiro Kizaki, Toshikazu Bando, Ganesh N. Pandian, Hiroshi Sugiyama**

Page No. - Page No.

Advancing Small-Molecule-Based Chemical Biology With Next-Generation Sequencing Technologies

Table1: 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	Maximum read length (bp)	Run time	Key features	Limitations
HiSeq 2500	Illumina	Sequencing by synthesis	120gb	100	27hrs	Larger read number	Higher cost per read
PGM/ Ion proton	Ion Torrent	Polymerization	~60gb	300	5-8hrs	Simple detection method	Low reads number per run
Genome Sequencer FLX system	Roche	Pyro sequencing	0.5gb	400	8hrs	Long read length	High cost per Mb
SOLiD	Life Technologies	Sequencing by ligation	50gb	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)	13gb	4,200 to 8,500 bp	-	Single molecule detection, long read length	High error rates and Low reads number per run
Polonator G.007	Dover	Ligation	12gb	26	5days	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

Table 2: List of studies used next-generation sequencing technologies beyond just genomic sequencing and its large scale operations in chemical biological research.

S.No	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 509.8M 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 51bp 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	1-1.5M reads per library

Abbreviation

1mA	1-methyladenine
5caC	5-carboxycytosine
5fC	5-formylcytosine
5hC	5-hydroxycytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxymethyluracil
5hU	5-hydroxyuracil
5mC	5-methylcytosine
8oxoA	8-oxoadenine
8oxoG	8-oxoguanine
AR	androgen receptor
AR	androgen receptor
ASKAS	aptamer selection by K-mer analysis of sequences
BLESS	direct in-situ DSB labeling, avidin enrichment and deep sequencing
BRCA2	breast cancer susceptibility gene 2
BS-Seq	bisulfite sequencing
CCL2	CC chemokine ligand 2
cDNA	complementary DNA
ChIP-Seq	combination of chromatin immune precipitation and high throughput sequencing
dDIP	damaged DNA immunoprecipitation
dDsTP	deoxyribonucleoside triphosphates containing Ds
DHT	dihydrotestosterone
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	deoxyribonucleoside triphosphates
dPxTP	deoxyribonucleoside triphosphates containing Px
DREME	Discriminative Regular Expression Motif Elicitation
Ds	7-(2-thienyl) imidazo[4,5-b]pyridine
DSB	double strand break
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GR	glucocorticoid receptor
H3Ac	histone H3 acetylation
HIF-1α	hypoxia inducible factor 1 alpha
IFN-γ	interferon-γ
KLK3	kallikrein-related peptidase 3
MACS	Model-based Analysis of ChIP-Seq
MBD-Seq	Methyl-CpG binding domain protein sequencing
MeDIP-Seq	methylated DNA immunoprecipitation
MMP28	matrix metalloproteinase 28
NF-κB	nuclear factor kappa B
NGS	next generation sequencing
NPTX1	neuronal pentraxin 1
O4mT	O4-metylthymine
O6mG	O6-methylguanine
oxBS-Seq	oxidative bisulfite sequencing
PCR	polymerase chain reaction
PIP	pyrrole imidazole polyamide
PIWIL1	piwi-like 1

Px	diol-modified 2-nitro-4-propynylpyrrole
qRT-PCR	quantitative reverse transcriptase PCR
redBS-Seq	reduced bisulfite sequencing
REPSA	restriction endonuclease protection, selection, and amplification
RNA	ribonucleic acid
RNAP2	RNA polymerase II
RNA-Seq	RNA high through put sequencing
RPB1	RNA polymerase II largest subunit
SAHA	suberoylanilide hydroxamic acid
SELEX	systematic evolution of ligands by exponential enrichment
SERPINE1	serpin peptidase inhibitor, clade E, member 1
SMRT	Single-Molecule Real Time
SPI-Seq	single-strand DNA (ssDNA)-associated protein immunoprecipitation followed by sequencing
TAB-Seq	TET-assisted bisulfite sequencing
TAmC-Seq	Tet-assisted 5mC sequencing
TET	ten eleven translocation
VEGF-165	vascular endothelial cell growth factor-165
β-GT	β-glucosyltransferase