Novel Strategies To Increase Read Length And Accuracy For DNA Sequencing By Synthesis

Lin Yu

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2010

Illumina Ex. 1093
IPR Petition - USP 10,435,742

ABSTRACT

Novel Strategies To Increase Read Length And Accuracy For DNA Sequencing By Synthesis

Lin Yu

The completion of the Human Genome Project has increased the need for high-throughput DNA sequencing technologies aimed at uncovering the genomic contributions to diseases. The DNA sequencing by synthesis (SBS) approach has shown great promise as a new platform for decoding the genome. This thesis focuses on the development and improvement of a chip-based four-color DNA SBS platform using molecular engineering approaches. In this approach, four nucleotides (A, C, G, T) are modified as fluorescent nucleotide reversible terminators (CF-NRTs) by tethering a cleavable fluorophore to the base and capping the 3'-OH with a small chemically reversible moiety so that the nucleotide analogues are still recognizable as substrates by DNA polymerase. First, we explored the potential of using an azido modified group for nucleotide modification. Based on our established rationale for nucleotide reversible terminator (NRT) design, we synthesized a complete set of NRTs capped at the 3' position with an azidomethyl group (3'-O-N_3-dATP, 3'-O-N_3-dCTP, 3'-O-N_3-dGTP, 3'-O-N_3-dTTP). Through testing and optimization, it was apparent that these NRTs were good substrates of a DNA polymerase. Afterwards, we worked out an optimum chemical cleavage condition to remove the azidomethyl group capping the 3'-OH of the nucleotide analogues under conditions that were

compatible with DNA, allowing the next NRT to be incorporated in the subsequent polymerase reaction. We then designed and synthesized two sets of azido-modified CF-NRTs for applications in SBS. The four CF-NRTs of the first set (3'-N_3-O-dNTP-azidomethylbenzoyl-fluorophores) were capped at the 3'-OH with an azidomethyl group identical to the NRTs and contained a substituted 2-azidomethylbenzoyl linker to tether a fluorophore. These CF-NRTs were used to produce four-color *de novo* DNA sequencing data on a chip based our sequencing by synthesis approach. After each round of sequencing, both the fluorophores linked to the CF-NRTs and the 3'-azidomethyl group on the DNA extension products generated by incorporating 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores were removed using a TCEP [Tris(2-carboxyethyl)phosphine] cleavage solution. This one-step dual-cleavage process for reinitiating the polymerase reaction increased the overall SBS efficiency. After confirming the feasibility of implementing azido-modified CF-NRTs in SBS, we synthesized a second set of CF-NRTs (3'-O-N_3-dNTP-N_3-fluorophores) to further improve and optimize the sequencing process. During the incorporation stage of SBS, a mixture of CF-NRTs and NRTs was used to simultaneously extend the primer strand of various target DNA linear templates. This approach led to a more efficient DNA polymerase reaction since the smaller 3'-O-N_3-dNTPs were much easier to incorporate. Moreover primers extended with NRTs resembled nascent strands of DNA that had no traces of modification after cleavage of the 3'-azidomethyl capping group. After the incorporation reaction, two separate capping steps, first with 3'-O-N_3-dNTPs and then with ddNTPs, were performed to synchronize all the templates on the surface. Without these precautionary synchronization procedures, mixed fluorescent signals would prevent the identification of the correctly incorporated nucleotide. Hence, we have successfully addressed one of the key drawbacks of SBS, which was the miscalling of the base due to lagging signals. In addition, since both 3'-O-N_3-dNTP-N_3-fluorophores and 3'-O-N_3-dNTPs were reversible terminators, which allow the sequencing of each base in a serial manner, they could accurately determine the homopolymeric regions of DNA. Finally, we developed a novel template walking strategy to increase read length for DNA SBS. The template walking method involved resetting the sequencing start site by extending the sequencing primer with three natural nucleotides and one NRT so that the polymerase reaction was temporarily paused when the NRT was incorporated. Upon restoring the 3'-OH group of the NRT incorporated into the primer via cleavage, the next cycle of walking could be carried out until the entire preiously sequenced portion of the template was skipped. We have successfully demonstrated the integration of this template walking strategy into our four-color DNA SBS platform by performing one round of SBS, four cycles of template walking reactions, and then a second round of SBS. Through this effort, we were able to sequence a linear DNA template in its entirety, nearly doubling the read length of our previous sequencing results. We are also taking advantage of the massive throughput of a next generation sequencer that is based on our SBS technology to conduct digital gene expression study of *Aplysia* central nervous system in an ongoing project that explores the molecular mechanism of long-term memory formation.

Table of Contents

List of Figures ix

Acknowledgements xvii

Abbreviations and Symbols xix

Chapter 1: Introduction to DNA Sequencing Technologies 1

1.1 Introduction 1

1.2 Background and Significance 2

1.2.1. Sanger dideoxynucleotide sequencing 6

1.2.2. MALDI-TOF MS based DNA sequencing 9

1.2.3. Pyrosequencing 13

1.2.4. DNA sequencing by ligation 15

1.2.5. DNA sequencing by engineered nanopores 17

1.3 Conclusion 19

1.4 References 20

Chapter 2: Overview of DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators 25

2.1 Introduction 25

2.2 General Methodology for DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators 26

2.3 Four color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators 29

2.3.1. Overview 29

2.3.2. Design, synthesis, and characterization of cleavable fluorescent nucleotide reversible terminators 32

2.3.3. DNA chip construction 35

2.3.4. Four color sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators 37

2.4 References 40

Chapter 3: Exploration of a New Chemical Moiety for Nucleotide Reversible Terminator Modification in DNA Sequencing by Synthesis 43

3.1 Introduction 43

3.2 Experimental Rationale and Overview 44

3.3 Results and Discussion 46

3.3.1. Design and synthesis of 3'-O-azidomethyl-dUTP-NH_2, a model NRT compound 46

3.3.2. Polymerase reaction using 3'-O-azidomethyl-dUTP-NH_2 and characterization by MALDI-TOF MS 47

3.3.3. Cleavage reaction to restore 3’-OH of DNA extension product and its optimization 50

3.3.4. Design, synthesis, and evaluation of a complete nucleotide reversible terminator set: 3’-O-N_3-dNTPs 52

3.3.5. Continuous polymerase extension using 3’-O-modified NRTs and characterization by MALDI-TOF mass spectrometry 54

3.4 Materials and Methods 56

3.4.1. Synthesis of 3’-O-azidomethyl-dUTP-NH_2, a model NRT compound.. 57

3.4.2. Polymerase reaction using 3’-O-azidomethyl-dUTP and characterization by MALDI-TOF MS 57

3.4.3. Cleavage reaction to restore 3’-OH of DNA extension product and its optimization 58

3.4.4. Design, synthesis, and evaluation of a complete nucleotide reversible terminator set: 3’-O-N_3-dNTPs 58

3.4.5. Continuous polymerase extension using 3’-O-modified NRTs and characterization by MALDI-TOF mass spectrometry 60

3.5 Conclusion 61

3.6 References 61

Chapter 4: Design, Synthesis, and Evaluation of a Novel Class of Cleavable Fluorescent Nucleotide Reversible Terminators Containing Substituted 2-Azidomethyl Benzoic Acid Linker for DNA Sequencing by Synthesis 64

4.1 Introduction 64

4.2 Experimental Rationale and Overview 65

4.3 Results and Discussion 69

4.3.1. Synthesis of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores 69

4.3.2. Polymerase single base extension and subsequent cleavage reactions of 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2 in solution and characterization by MALDI-TOF MS 69

4.3.3. Polymerase extension of the complete set of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores as reversible fluorescent nucleotide terminators in solution and characterization by MALDI-TOF MS 73

4.3.4. Four-color DNA sequencing on a chip using 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (CF-NRTs) and unlabeled 3′-*O*-N_3-dNTPs (NRTs) 75

4.4 Materials and Methods 80

4.4.1. Synthesis of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores 80

4.4.2. Polymerase single base extension and subsequent cleavage reactions of 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2 in solution and characterization by MALDI-TOF MS 81

4.4.3. Polymerase extension of the complete set of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores as reversible fluorescent nucleotide terminators in solution and characterization by MALDI-TOF MS 82

4.4.4. Four-color DNA sequencing on a chip using 3′-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (CF-NRTs) and unlabelled 3′-O-N_3-dNTPs (NRTs) 83

4.5 Conclusion 84

4.6 References 86

Chapter 5: Four-color DNA Sequencing by Synthesis (SBS) Improvements using Cleavable Fluorescent Nucleotide Reversible Terminators 88

5.1 Introduction 88

5.2 Experimental Rationale and Overview 91

5.3 Results and Discussion 93

5.3.1. Design and synthesis of cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs for SBS 93

5.3.2. Polymerase extension using cleavable fluorescent nucleotide reversible terminators in solution and their characterization by MALDI-TOF MS. 98

5.3.3. Four-color DNA sequencing by synthesis on a chip using cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs 99

5.4 Materials and Methods 104

5.4.1. Design and synthesis of cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs for SBS 104

5.4.2. Polymerase extension using cleavable fluorescent nucleotide reversible terminators in solution and their characterization by MALDI-TOF MS. 105

5.4.3. Construction of a DNA Immobilized Chip of Multiple Linear Templates 106

5.4.4. Four-color DNA sequencing by synthesis on a chip using cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs 107

5.5 Conclusion 108

5.6 References 113

Chapter 6: Exploration of Novel Primer Resetting Strategies to Extend Read-length for DNA Sequencing by Synthesis 117

6.1 Introduction 117

6.2 Experimental Rationale and Overview 118

6.2.1. Strategy 1: template “walking” by unlabeled nucleotides 119

6.2.2. Strategy 2: template “walking” with universal bases 126

6.2.3. Strategy 3: multiple primer hybridization 127

6.3 Results and Discussion 130

6.3.1. Template walking using three natural nucleotides and MALDI-TOF MS characterization of walking products 131

6.3.2. Template walking using three natural nucleotides and one NRT for four-color DNA Sequencing by Synthesis 136

6.4 Materials and Methods 142

6.4.1. Template walking using three natural nucleotides and MALDI-TOF MS characterization of walking products..143

6.4.2. Template walking using three natural nucleotides and one NRT for four-color DNA Sequencing by Synthesis..145

6.5 Conclusion...147

6.6 References...149

Chapter 7: Massively Parallel Monitoring of Gene Expression in *Aplysia* Central Nervous System (CNS) using Four-color DNA Sequencing by Synthesis..............150

7.1 Introduction 150

7.2 Experimental Rationale and Overview..153

7.3 Results and Discussion 155

7.4 Materials and Methods 159

7.4.1. Gene transcript analysis in *Aplysia* neuronal cells using Illumina Genome Analyzer...160

7.5 Conclusion..160

7.6 References..161

Chapter 8: Summary and Future Outlook...163

8.1 Exploration of a New Chemical Moiety for Nucleotide Reversible Terminator Modification in DNA Sequencing by Synthesis [1] 163

8.2 Design, Synthesis, and Evaluation of a Novel Class of Cleavable Fluorescent Nucleotide Reversible Terminators Containing Substituted 2-Azidomethyl Benzoic Acid Linker for DNA Sequencing by Synthesis [2]........164

8.3 Four-color DNA Sequencing by Synthesis (SBS) Improvements Using Cleavable Fluorescent Nucleotide Reversible Terminators [3]............................165

8.4 Exploration of Novel Primer Resetting Strategies to Extend Read Length for DNA Sequencing by Synthesis [4]..166

8.5 Future Outlook for 4-color DNA Sequencing by Synthesis using CF-NRTs [5, 6] 166

8.6 References..168

List of Figures

Fig. 1.1. Chemical structures of 2'-deoxyribonucleotide triphosphates. Each nucleotide is composed of a base (adenine, guanine, cytosine, or thymine), a sugar, and a phosphate group. 3

Fig. 1.2. (A) A cartoon illustrating the double helical structure of DNA. Each strand is supported by the sugar-phosphate backbone. They are held together via hydrogen bonds in anti-parallel fashion (the 5' end of one strand aligns with the 3' end of the other one). (B) A figure depicting two DNA strands held together by hydrogen bonds between paired bases. (C) More detailed chemical structure showing the hydrogen bonding between bases. 4

Fig. 1.3. Scheme of DNA polymerase reaction. DNA synthesis takes place via the addition of a nucleotide to the 3'-OH end of a DNA primer strand. The base-pairing between the incoming nucleotide and the DNA template strand dictates which nucleotide is added. DNA polymerase facilitates the addition of the incoming nucleotide by catalyzing the formation of a phosphodiester bond between the terminal 3'-OH group of the primer strand and the alpha phosphorus atom of the nucleotide. A pyrophosphate (PPi) group is released as a by-product. 6

Fig. 1.4. Chemical structures of 3'-deoxyribonucleotide (dNTP) and 2', 3'-dideoxyribonucleotide (ddNTP). Since ddNTPs do not have the 3'-OH group, which is necessary for DNA synthesis, they terminate further extension of the DNA strand once incorporated. 7

Fig. 1.5. Sanger dideoxy chain-termination sequencing method. DNA fragments are generated by extending the primer with a mixture of dNTPs and ddNTPs. Upon incorporation of a ddNTP, the DNA strand ceases to participate in polymerase reaction due to the lack of the 3'-OH group. Thus, a mixture of DNA strands with different length complementary to the template DNA is produced. To determine the sequence of the template, these DNA fragments are separated based on size by electrophoresis, and the resulting bands of DNA are detected by their fluorescent signals. 8

Fig. 1.6. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). A mixture of analyte (e.g. DNA sequencing fragments) and matrix molecules (blue) are spotted on the sample plate and allowed to co-crystallize prior to loading into the vacuum chamber. After UV laser irradiation, the desorbed and ionized analyte and matrix molecules are accelerated under a constant electric voltage, causing them to fly towards the detector. The charged molecules arrive at the detector at different times based on their masses. Therefore, the masses of the charged particles can be determined from their time-of-flight. 11

Fig. 1.7. DNA sequencing using MALDI-TOF MS. (A) Sanger sequencing fragments generated using biotin-labeled ddNTPs. (B) Example of mass sequencing spectrum using biotin-labeled ddNTPs. 13

Fig. 1.8. General scheme of pyrosequencing. As the polymerase catalyzes the incorporation of nucleotide(s) into a growing strand of DNA, PPi molecules are released and then converted to ATPs by ATP-sulfurylase. The ATPs participate in the luciferase reaction in which a luciferin molecule is oxidized to produce oxyluciferin and light. The resulting luminescence can be registered using a photon detector. 15

Fig. 1.9. Scheme of DNA sequencing by ligation using degenerate nonamers. 16

Fig. 1.10. 3-D rendering of α-hemolysin structure. 17

Fig. 1.11. Scheme for DNA sequencing using nanopore and theoretical plot of translocation time versus blockade current elicited by DNA strand. 18

Fig. 2.1. In the SBS approach, a chip is constructed with immobilized DNA templates that are able to self-prime for initiating the polymerase reaction. Four nucleotide analogues are designed such that each is labeled with a unique fluorescent dye on the specific location of the base, and a small chemical group (R) to cap the 3'-OH group. Upon adding the four nucleotide analogues and DNA polymerase, only the nucleotide analogue complementary to the next nucleotide on the template is incorporated by polymerase on each spot of the chip (step 1). After removing the excess reagents and washing away any unincorporated nucleotide analogues, a 4-color fluorescence scanner is used to image the surface of the chip, and the unique fluorescence emission from the specific dye on the nucleotide analogues on each spot of the chip will yield the identity of the nucleotide (step 2). After imaging, the small amount of unreacted 3'-OH group on the self-primed template moiety will be capped by excess 3'-O-modified nucleotide reversible terminators and DNA polymerase to avoid interference with the next round of synthesis for synchronization (step 3). The dye moiety and the R protecting group will be removed to generate a free 3'-OH group with high yield (step 4). The self-primed DNA moiety on the chip at this stage is ready for the next cycle of the reaction by repetition of steps 1, 2, 3, and 4 to identify the next nucleotide in the sequence of the template DNA. 28

Fig. 2.2. DNA sequencing by synthesis based on modified nucleotide analogues. First, the primer is extended by the addition of four nucleotide analogues, each of which has a unique fluorescent label and can act as a reversible terminator in the DNA polymerase reaction. Upon detection of the incorporated nucleotide analogue by its unique fluorescent label, both the label and the 3' protecting group are removed to reinitiate the next DNA polymerase reaction cycle. 28

Fig. 2.3. The 3-D structure of the ternary complexes of a rat DNA polymerase, a DNA template-primer, and dideoxycytidine triphosphate (ddCTP). The left side of the illustration shows the mechanism for the addition of ddCTP and the right side shows the active site of the polymerase. It is important to notice that the 3' position of the dideoxyribose ring is very crowded, while ample space is available at the 5 position of the cytidine base. 30

Fig. 2.4. General structure of CF-NRT: the dual modified nucleotide analogue has a small reversible terminating moiety capping the 3' position and a cleavable linker tethering a unique fluorophore to the base of the nucleotide analgue. 31

Fig. 2.5. Structures of allyl modified CF-NRTs (3'-*O*-ally-dNTP-allyl-fluorophores), with the 4 fluorophores having distinct fluorescent emissions: 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510 [$\lambda_{abs(max)}$ = 502 nm; $\lambda_{em(max)}$ = 510 nm], 3'-*O*-allyl-dUTP-allyl-R6G [$\lambda_{abs(max)}$ = 525 nm; $\lambda_{em(max)}$ = 550 nm], 3'-*O*-allyl-dATP-allyl-ROX [$\lambda_{abs(max)}$ = 585 nm; $\lambda_{em(max)}$ = 602 nm], and 3'-*O*-allyl-dGTP-allyl-bodipy-650 [$\lambda_{abs(max)}$ = 649 nm; $\lambda_{em(max)}$ = 670 nm]............... 33

Fig. 2.6. The polymerase extension scheme (*Left*) and MALDI-TOF MS spectra of the four consecutive extension products and their deallylated products (*Right*). Primer extended with 3'-*O*-allyl-dUTP-allyl-R6G (1), and its deallylated product (2); Product (2) extended with 3'-*O*-allyl-dGTP-allyl-bodipy-650 (3), and its deallylated product (4); Product (4) extended with 3'-*O*-allyl-dATP-allyl-ROX (5), and its deallylated product (6); Product (6) extended with 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510 (7), and its deallylated product (8). After 30s of incubation with the palladium/TPPTS mixture at 70°C, deallylation is complete with both the fluorophores and the 3'-*O*-allyl groups cleaved from the extended DNA products. 35

Fig 2.7. DNA immobilization on a surface with click chemistry. An amino modified glass surface is reacted with a linker containing an NHS-ester on one end and an alkyne on the other to functionalize the surface (step 1). Azide labeled PCR product is then attached to the surface using click chemistry (step 2). The unattached DNA strand is removed under alkaline conditions (step 3). A loop DNA primer is ligated to the single-stranded DNA template (step 4). DNA extension reactions are carried out to identify the sequence of the template (step 5). The sequence of the loop primer is shown in (A).. 36

Fig. 2.8. Four-color sequencing by synthesis data on a DNA chip. (*A*) Reaction scheme of SBS on a chip using four chemically cleavable fluorescent nucleotides. (*B*) The scanned four-color fluorescence images for each step of SBS on a chip: (1) incorporation of 3'-*O*-allyl-dGTP-allyl-Cy5; (2) cleavage of allyl-Cy5 and 3'-allyl group; (3) incorporation of 3'-*O*-allyl-dATP-allyl-ROX; (4) cleavage of allyl-ROX and 3'-allyl group; (5) incorporation of 3'-*O*-allyl-dUTP-allyl-R6G; (6) cleavage of allyl-R6G and 3'-allyl group; (7) incorporation of 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510; (8) cleavage of allyl-bodipy-FL-510 and 3'-allyl group; images 9–25 are similarly produced. (*C*) A plot (four-color sequencing data) of raw fluorescence emission intensity at the four designated emission wavelengths of the four chemically cleavable fluorescent nucleotides vs. the progress of sequencing extension. 38

Fig. 2.9. Comparison of four-color sequencing by synthesis and pyrosequencing data. (*A*) Four-color DNA sequencing raw data with our sequencing by synthesis chemistry using a template containing two homopolymeric regions. The individual bases (A, T, C, G), the 10 repeated A's, and the five repeated A's are clearly identified. The small groups of peaks between the identified bases are fluorescent background from the DNA chip, which does not build up as the cycle continues. (*B*) The pyrosequencing data of the same DNA template containing the homopolymeric regions (10 T's and five T's). The first four individual bases are clearly identified. The two homopolymeric regions (10 A's) and (five A's) produce two large peaks, from which it is very difficult to identify the exact sequence. 39

Fig. 3.1. Structure of the azido functional group 44

Fig. 3.2. Staudinger reduction of azido functional group using TCEP 45

Fig. 3.3. Staudinger reduction of 3'-O-N_3 group of an incorporated NRT to recover its 3'-OH group 46

Fig. 3.4. Structure of 3'-O-azidomethyl-dUTP 47

Fig. 3.5. Single base extension reaction scheme (top) and resulting MALDI-TOF MS spectrum (bottom) of 3'-O-azidomethyl-dUTP-NH_2. 48

Fig. 3.6. Single base extension reaction scheme (left) and MALDI-TOF MS spectra of 3'-O-azidomethyl-dUTP with varying reaction times of A) 20min, B) 10min, C) 5min, and D) 2min. 49

Fig. 3.7. MALDI-TOF MS spectra of cleavage products after treating the single base extension product using TCEP for varying times: A) 30min, B) 15min, C) 5min, and varying TCEP concentration: D) 25mM, E) 10mM, F) 5mM. In each case, the 3'-OH is recovered completely, showing a m/z of 8,308. 51

Fig. 3.8. Structures of the nucleotide reversible terminators (NRTs): 3'-*O*-N_3-dATP, 3'-*O*-N_3-dCTP, 3'-*O*-N_3-dGTP, 3'-*O*-N_3-dTTP. 52

Fig. 3.9. Scheme (left) and MALDI-TOF MS spectra (right) of single base extension reaction and 3'-O-azidomethyl cleavage reaction for each modified nucleotide reversible terminator (3'-O-N_3-dATP, 3'-O-N_3-dGTP, 3'-O-N_3-dTTP, and 3'-O-N_3-dCTP) 54

Fig. 3.10. The polymerase extension scheme (left) and MALDI-TOF MS spectra of the four consecutive extension products and their cleaveage products (right) using the nucleotide reversible terminators, 3'-O-N_3-dNTPs. Primer extended with 3'-O-N_3-dCTP (1) (right, A), and its cleavage product (2) (right, B); Product (2) extended with 3'-O-N_3-dGTP (3) (right, C), and its cleavage product (4) (right, D); Product (4) extended with 3'-O-N_3-dATP (5) (right, E), and its cleavage product (6) (right, F); Product (6) extended with 3'-O-N_3-dTTP (7) (right, G), and its cleavage product (8) (right, H). After brief incubation in a TCEP aqueous solution the azidomethyl moiety capping the 3'-OH group of the DNA extension products is completely removed to continue the polymerase reaction 56

Fig. 4.1. General Structure of Cleavable Fluorescent Nucleotide Reversible Terminators 64

Fig. 4.2. Structure of 2-azidomethylbenzoyl group 66

Fig. 4.3. Reaction scheme of 2-azidomethylbenzoyl reduction by methyldiphenylphosphine ($MePPh_2$) 67

Fig. 4.4. Structures of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, a novel class of cleavable fluorescent nucleotide reversible terminators 68

Fig. 4.5. Stucture of 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 70

Fig. 4.6. Single base extension reaction scheme (top) and resulting MALDI-TOF MS spectrum (bottom) of 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 71

Fig. 4.7. Structures of A) completely cleaved model nucleotide analogue and B) incompletely cleaved model nucleotide analogue with residual linker still attached 72

Fig. 4.8. TCEP cleavage scheme of 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 (top) and resulting MALDI-TOF MS spectrum (bottom) at various pHs: A) pH=5.5, B) pH=9.0, C) pH=10.0 73

Fig. 4.9. A polymerase reaction scheme (Top) to yield DNA extension products by incorporating each of the four 3'-*O*-N_3-dNTP-azidomethylbenzoyl-fluorophores and the subsequent cleavage reaction to remove the fluorophores and 3'-azidomethyl group from the DNA extension products. MALDI-TOF MS spectra (Bottom) showing efficient base-specific incorporation of the 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and the subsequent cleavage of the fluorophores from the DNA extension products: (A) Primer extended with 3'-O-N_3-dATP-azidomethylbenzoyl-ROX (1) (peak at 9,216 m/z), (B) its cleavage product (2) (8,437 m/z); (C) primer extended with 3'-O-N_3-dGTP-azidomethylbenzoyl-Cy5 (3) (peak at 9,358 m/z), (D) its cleavage product (4) (8,438 m/z); (E) primer extended with 3'-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-FL-510 (5) (peak at 8,855 m/z), (F) its cleavage product (6) (8,307 m/z); (G) primer extended with 3'-O-N_3-dUTP-azidomethylbenzoyl-R6G (7) (peak at 9,022 m/z), and (H) its cleavage product (8) (8,308 m/z). 75

Fig. 4.10. Comparison of structural difference of DNA incorporated A) CF-NRTs and B) NRTs after cleavage reactions 76

Fig. 4.11. Four-color DNA SBS with 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores. (A) A SBS with CF-NRTs scheme for four-color sequencing on a chip by using four 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3'-O-N_3-dNTPs. (B) Four-color fluorescence images for each step of the SBS: (1) incorporation of 3'-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-Fl-510 and

3’-O-N_3-dCTP; (2) cleavage of azidomethylbenzoyl-Bodipy-Fl-510 and 3’-azidomethyl group; (3) incorporation of 3’-O-N_3-dATP-azidomethylbenzoyl-Rox and 3’-O-N_3-dATP; (4) cleavage of azidomethylbenzoyl-Rox and 3’-azidomethyl group; images 5-18 were produced similarly. (C) A plot (four-color sequencing data) of raw fluorescence emission intensity obtained by using 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3’-O-N_3-dNTPs. The small groups of peaks between the identified bases are fluorescent background from the DNA chip. 80

Fig. 5.1. Scheme of “lagging” sequence during SBS due to incomplete incorporation of CF-NRTs 90

Fig. 5.2. Scheme of “read ahead” sequence during SBS due to incorporation of 3’ unblocked fluorescent nucleotides 91

Fig. 5.3. Scheme for sequencing by synthesis using CF-NRTs, NRTs, and ddNTPs 93

Fig. 5.4. Structures of the nucleotide reversible terminators (NRTs), 3’-O-N_3-dATP, 3’-O-N_3-dCTP, 3’-O-N_3-dGTP, 3’-O-N_3-dTTP 95

Fig. 5.5. Structures of cleavable fluorescent nucleotide reversible terminators (CF-NRTs: 3’-O-N_3-dNTP-N_3-fluorophores, with the 4 fluorophores having distinct fluorescent emissions: 3’-O-N_3-dCTP-N_3-Bodipy-FL-510 ($\lambda_{abs\ (max)}$ = 502 nm; $\lambda_{em\ (max)}$ = 510 nm), 3’-O-N_3-dUTP-N_3-R6G ($\lambda_{abs\ (max)}$ = 525 nm; $\lambda_{em\ (max)}$ = 550 nm), 3’-O-N_3-dATP-N_3-ROX ($\lambda_{abs\ (max)}$ = 585 nm; $\lambda_{em\ (max)}$ = 602 nm), and 3’-O-N_3-dGTP-N_3-Cy5 ($\lambda_{abs\ (max)}$ = 649 nm; $\lambda_{em\ (max)}$ = 670 nm). 97

Fig. 5.6. Scheme of TCEP cleavage mechanism of a DNA strand extended with a CF-NRT (3’-O-N_3-dCTP-N_3-Bodipy-Fl-510). TCEP is able to cleave both the azido linker to remove the fluorophore and the 3’-azidomethyl capping moiety to restore the 3’-OH group. 97

Fig. 5.7. A polymerase reaction scheme (top) to yield DNA extension products by incorporating each of the four 3’-O-N_3-dNTP-N_3-fluorophores and the subsequent cleavage reaction to remove the fluorophore from the DNA extension product. MALDI-TOF MS spectra (bottom) showing efficient base specific incorporation of the 3’-O-N_3-dNTP-N_3-fluorophores and the subsequent cleavage of the fluorophore from the DNA extension products: (A) Primer extended with 3’-O-N_3-dATP-N_3-ROX (1) (peak at 7,369 *m/z*), (B) its cleavage product 2 (6,554 *m/z*); (C) Primer extended with 3’-O-N_3-dGTP-N_3-Bodipy-FL-510 (3) (peak at 7,552 *m/z*), (D) its cleavage product (4) (6,609 *m/z*); (E) Primer extended with 3’-O-N_3-dCTP-N_3-Cy5 (5) (peak at 6,131 *m/z*), (F) its cleavage product (6) (5,603 *m/z*); (G) Primer extended with 3’-O-N_3-dUTP-N_3-R6G (7) (peak at 6,353 *m/z*) and (H) its cleavage product (8) (5,612 *m/z*). 99

Fig. 5.8. Four-color DNA SBS with 3’-O-N_3-dNTP-N_3-fluorophores (90mer3). (A) A SBS with CFNRTs scheme for four-color sequencing on a chip by using four 3’-O-N_3-dNTP-N_3-fluorophores and 3’-O-N_3-dNTPs with ddNTPs capping. (B) Four-color fluorescence images for each step of the SBS: (1) incorporation of 3'-O-N_3-dCTP-N_3-Bodipy-Fl-510 and 3'-O-N_3-dCTP; (2) cleavage of N_3-Bodipy-Fl-510 and 3'-CH_2N_3 group; (3) incorporation of 3'-O-N_3-dATP-N_3-Rox and 3'-O-N_3-dATP; (4) cleavage of N_3-Rox and 3'-CH_2N_3 group; images 5-60 were produced similarly. (C) A plot (four-color sequencing data) of raw fluorescence emission intensity obtained by using 3'-O-N_3-dNTP-N_3-fluorophores and 3'-O-N_3-dNTPs. The small groups of peaks between the identified bases are fluorescent background from the DNA chip. 104

Fig. 5.9. Scheme for parallel sequencing of multiple templates by four-color SBS 111

Fig. 5.10. A plot (4-color sequencing data) of raw fluorescence emission intensity of two more templates (90mer2 and 90mer4) sequenced simultaneously with previously presented template (90mer3) using four-color SBS with 3'-O-N_3-dNTPs and 3'-O-N_3-dNTP-N_3-fluorophores. 113

Fig. 6.1. Scheme for primer resetting strategy 1: template walking 120

Fig. 6.2. Template "walking" method 1 for SBS with CF-NRTs 123

Fig. 6.3. Template "walking" method 2 for SBS with CF-NRTs 124

Fig. 6.4. Template "walking" method 3 for SBS with CF-NRTs 125

Fig. 6.5. Structure of inosine in deoxynucleotide form 126

Fig. 6.6. Scheme for template resetting strategy 2: universal base walking......127

Fig. 6.7. Scheme for template resetting strategy 3: multiple primer hybridization 128

Fig. 6.8. MALDI-TOF MS spectrum of single base extension reaction using self-priming SP26T template and dTTP with DNA polymerase: A) Therminator II, B) Pfu DNA polymerase, C) Tgo DNA polymerase, and D) Thermosequenase 133

Fig. 6.9. MALDI-TOF MS spectrum of self-priming SP26T template extended with dATP, dGTP, and dTTP using Thermosequenase. Both the expected product (m/z = 8908) and product with extra base mis-incorporation (m/z = 9214) were present. 134

Fig. 6.10. MALDI-TOF MS spectrum of self-priming SP26T template extended with dATP, dGTP, and dTTP using Thermosequenase and PyroE solution. Only the expected product (m/z = 8908) was present, demonstrating the successful removal of excessive dNTPs in the incorporation mixture. 135

Fig. 6.11. Scheme for template walking using three natural nucleotides (dATP, dCTP, and dTTP) and one NRT (3'-O-N_3-dGTP) 136

Fig. 6.12. General scheme of SBS integrated with template walking 138

Fig. 6.13. Four-color sequencing data plot of raw fluorescence emission intensity obtained by 1st round SBS using CF-NRTs of a 90mer linear template. 31 consecutive bases were correctly identified. The small groups of peaks between the identified bases represent fluorescent background from the DNA chip after cleavage. The last small group of peaks, following base 31, revealed the removal of the sequencing primer extended with CF-NRTs after the denaturing process....140

Fig. 6.14. Four-color sequencing data plot of raw fluorescence emission intensity obtained by 2nd round SBS using CF-NRTs after four cycles of template walking. 23 consecutive bases were correctly identified. The small groups of peaks between the identified bases represent fluorescent background from the DNA chip after cleavage. 141

Fig. 6.15. Four-color sequencing data plot of fluorescence emission intensity obtained by combining 1st and 2nd round SBS using CF-NRTs. The first 30 bases were sequenced during the 1st round of SBS. After stripping away the sequencing primer extended with CF-NRTs, the original primer was reannealed to the template for four cycles of template walking using a mixture of three natural dNTPs (dATP, dCTP, and dTTP) and one NRT (3'-O-N_3-dGTP). Upon reaching the base at the end of the previous sequencing round, SBS was re-initiated to correctly identify the next 23 consecutive bases. By using SBS integrated with template walking, the target template of 53 bases was sequenced in its entirety.....142

Fig. 7.1. Photo of *Aplysia californica*, courtesy of Moroz lab.....152

Fig. 7.2. Scheme of adaptor ligation for genomic library DNA in Illumina Genome Analyzer.....154

Fig. 7.3. Scheme of cluster generation in flow cell prior to sequencing in Illumina Genome Analyzer.....155

Fig. 7.4. Illumina Genome Analyzer focus quality by fluorescent emission and cycle.....157

Fig. 7.5. Illumina Genome Analyzer fluorescent intensity of each base for selected sequencing cycles. 158

Fig. 7.6. (A) Scatter-plot reflecting differences in gene expression levels in 5HT-treated and intact (control) ganglia. Shown on the plot are the frequencies of gene tags which were log-transformed at base 2 and normalized using *Lowess* normalization. In yellow are tags corresponding to genes with lower than 2-fold changes in expression level. Shown in red are 247 gene tags with maximum expression level changes (LogRatio ranging from +4 to -6) that were chosen for further study. (B) Examples of differentially expressed genes: gene tag count in treated (blue) *vs* control cells (red), courtesy of Drs. Sergey Kalachikov and Irina Morozova. 159

Fig. 8.1. Theoretical SBS read length based on sequencing cycle efficiency........167

Acknowledgements

My journey as a graduate student at Columbia University started in 2005. At the time, I was just a kid fresh out of college with a big dream of becoming a scientist. For the past several years, I have had the fortune to learn and work with some of the greatest minds that I have the honor of calling my mentor and colleagues. They truly have helped me mature both as a student and a person. First and foremost, I would like to thank my mentor, Professor Jingyue Ju, for taking me under his wings as his student. His passion and dedication to science has been a great inspiration to me. He has guided me with his grand vision and taught me the ways of scientific thinking that I will truly treasure. It is and will always be my privilege to learn from Prof. Ju, a motivated scholar, scientist, and teacher. I would also like to thank Dr. Zengmin Li for his support and knowledge in chemistry as well as in life. The same appreciation is extended to Dr. James Russo, for his insightful advice and patient help. My special thanks goes out to my wonderful colleagues as well: Drs. Huanyan Cao, Xiaoxu Li, Mong Sang Marma, Shenglong Zhang, and Qingling Meng for their relentless work on the synthetic chemistry front; Drs. Shundi Shi, Sergey Kalachikov, Irina Morozova, and Ming Chen for teaching an engineer so much about molecular biology and biological bench work; Dr. Dae H. Kim for showing me all the nuances of different sequencing experimental procedures; Dr. Jia Guo for countless hours of collaboration; Chunmei Qiu and Wenjing Guo for their friendship; and Dr. John R. Edwards for being my big brother in science and life. I would also like to thank my defense committee members, Professor Faye McNeill (Department of Chemical

Engineering), Professor Chandima Abeywickrama (Department of Chemistry, Stevens Institute of Technology), and Professor Emil Kozarov (Columbia College of Dental Medicine) for their invaluable advice. Last but not least, I would like to thank two very special people that have always shown me love and care regardless of my ups and downs: my mother Yufeng Gu and my wife Vanessa Yu. I cannot imagine myself accomplishing anything without their unwavering support. Thank you very much!

Abbreviations and Symbols

ATP	adenosine 5’-triphosphate
Bodipy	4,4-difluoro-5,7-dimethyl-4-bora-3α,4α,-diaza-s-indacene
bp	base pair
CAE	capillary array electrophoresis
CF-NRT	cleavable fluorescent nucleotide reversible terminator
Cy5	cyanine-5
dA	deoxyadenosine
dC	deoxycytidine
ddNTP	dideoxynucleoside triphosphate
dG	deoxyguanosine
DMF	*N,N*-dimethylformamide
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
dT	deoxythymidine
EDTA	ethylenediaminetetraacetic acid
FMA	5-carboxyfluorescein
MALDI-TOF	matrix-assisted laser desorption ionization time-of-flight
MS	mass spectrometry
NHS	*N*-hydroxy succinimidyl

NRT	nucleotide reversible terminator
nt	nucleotide
OPC	oligonucleotide purification cartridge
PC	photocleavable
PCR	polymerase chain reaction
PEG	polyethylene glycol
PPi	pyrophosphate
ROX	6-carboxy-X-rhodamine
R6G	6-carboxyrhodamine 6G, hydrochloride
SBE	single base extension
SBS	sequencing by synthesis
SDS	sodium dodecyl sulfate
SNP	single nucleotide polymorphism
SPSC	sodium phosphate/sodium chloride
ss	single-stranded
Taq	*Thermus aquaticus*
TCEP	tris(2-carboxyethyl)phosphine
λ_{abs}	maximum absorption wavelength (nm)
λ_{em}	maximum emission wavelength (nm)

Chapter 1: Introduction to DNA Sequencing Technologies

1.1 Introduction

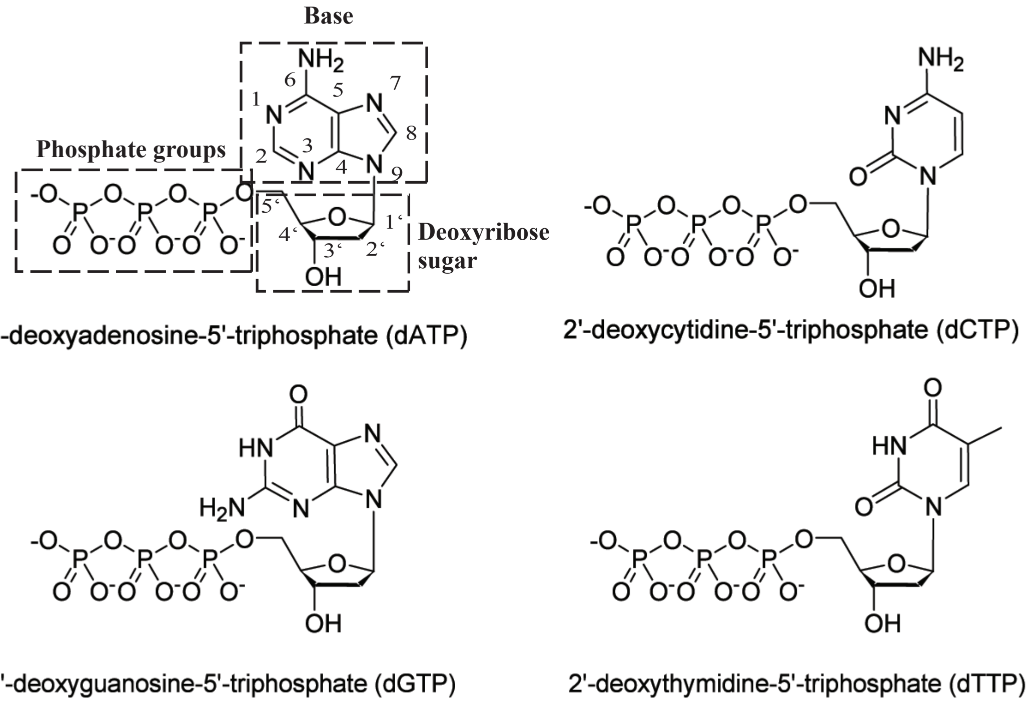
The history of science is filled with myriad marvels and wonders. For more than five thousand years, scientists have conquered countless challenges and obstacles through sheer ingenuity and diligence for the advancement of civilization. From the discovery of fire and invention of wheels in ancient times to the development of penicillin and landing on the moon in relatively recent memory, every single period of history can be defined by a specific monumental scientific triumph. At the dawn of the 21st century, one such achievement, the completion of the Human Genome Project,[1] revolutionized the way scientists approach biological and genetic studies, and stimulated an exponential growth in the development of high throughput genomic analysis technologies.[2] Through collaborative efforts in the fields of biology, chemistry, engineering, and computer science, such development has enabled the transition from studying individual genes separately to analyzing and comparing large genomes entirely. The driving force behind the Human Genome Project and this new revolution in biology and medicine is the DNA sequencing technology.

DNA sequencing is the process of determining the sequential order of nucleotides of a DNA strand. The sequence of DNA encodes genetic information that is transcribed and then translated into proteins to dictate the functioning of living organisms. To sequence a DNA segment is to decipher that perplexing code, hence, to lay the foundation for unraveling its specific role in genetic conservation and biological development. Recent data has shown that fundamental differences between species are accounted for at the subtle regulatory rather than by their overall number of genes.[3] Hence it is crucial to sequence genomes of closely related species as well as more human genomes to further understand the impact of these differences on biology. Other high profile projects, such as personalized medicine and large-scale genetic analysis, also require high-throughput and low-cost sequencing platforms that are beyond the current technologies.

The urgent need to overcome the limitations of the current sequencing technology, which is based on eletrophoresis and laser-induced fluorescence detection,[4-6] calls for innovative sequencing paradigms that can meet the challenge of ever-growing demands. The developers of several sequencing platforms that take advantage of sequencing by ligation,[7] nanopore sequencing,[8] mass-spectrometry based sequencing,[9-11] pyrosequencing,[12,13] and sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators,[14] have made great strides to usher in the next-generation sequencing era. In this chapter, the current generation of traditional and newer sequencing technologies will be reviewed.

1.2 Background and Significance

Deoxyribonucleic acid, or DNA, is a threadlike macro-polymer composed of deoxyribonucleotide units. Each nucleotide unit consists of a nitrogenous base, a sugar, and a phosphate group.[15] Genetic information is stored in the bases of DNA molecules while the sugar and phosphate groups provide the structural backbone. The sugar in a deoxyribonucleotide is 2’-deoxyribose, a derivative of the five-membered ribose ring. There are four different nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The first two, A and G, are purine derivatives while the latter two, C and T, are pyrimidines. Collectively, the four nucleotides corresponding to the four bases are referred to as deoxyribonucleoside 5’-triphosphates (dNTPs, N = A, G, C, or T) (Fig. 1.1). They serve as the structural and functional foundation of all DNA molecules.



2'-deoxyadenosine-5'-triphosphate (dATP)

2'-deoxycytidine-5'-triphosphate (dCTP)

2'-deoxyguanosine-5'-triphosphate (dGTP)

2'-deoxythymidine-5'-triphosphate (dTTP)

Fig. 1.1. Chemical structures of 2’-deoxyribonucleotide triphosphates. Each nucleotide is composed of a base (adenine, guanine, cytosine, or thymine), a sugar, and a phosphate group.

The 3-dimensional structure of DNA was first characterized by James Watson and Francis Crick in their 1953 *Nature* article.[16] DNA consists of two helical polynucleotide chains coiled around a common axis. Each chain has a backbone composed of deoxyriboses linked by phosphate groups. More specifically, the 3’-hydroxyl of the sugar ring of one deoxyribonucleotide is bridged to the 5’-hydroxyl of the subsequent sugar via a phosphodiester bond. Thus every DNA strand has a precise orientation characterized by the phosphate group on the 5’-terminal sugar unit

(5’-end) and a hydroxyl group at the 3’-terminal sugar unit (3’-end). Running in opposite directions, the two DNA strands are held together by hydrogen bonds between pairs of bases. Base pairing in DNA is very explicit: adenine is always paired with thymine and guanine is always paired with cytosine (Fig. 1.2). There is no restriction on the sequential order of bases alone the polynucleotide chain, but the precise sequence of those bases carries the genetic information.

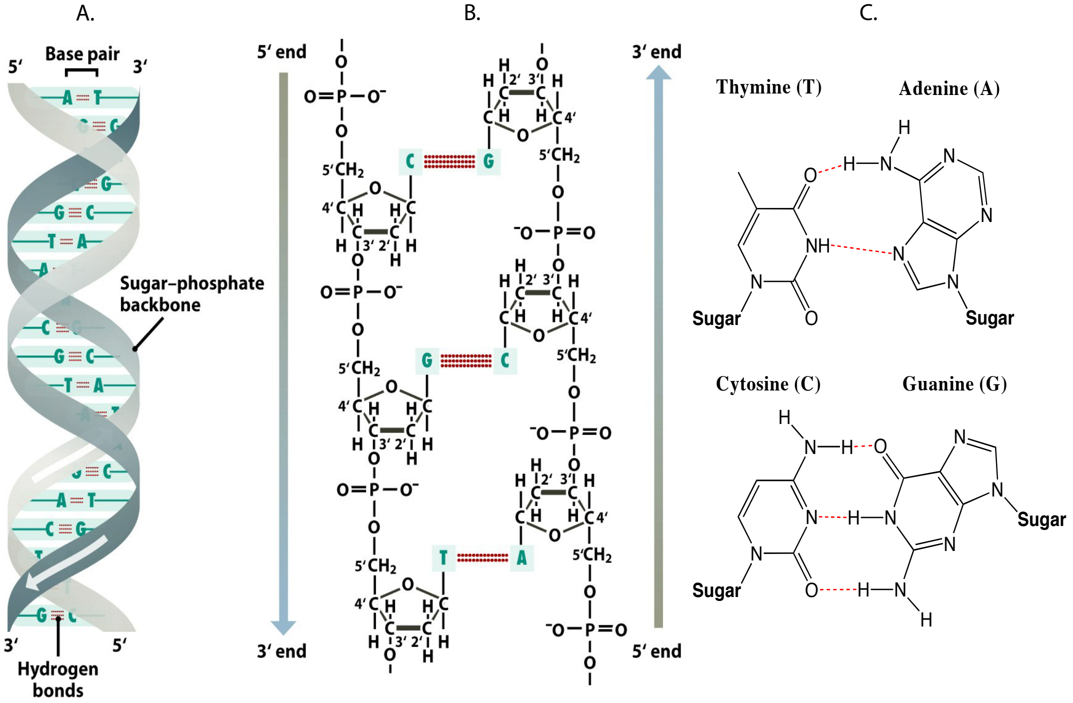
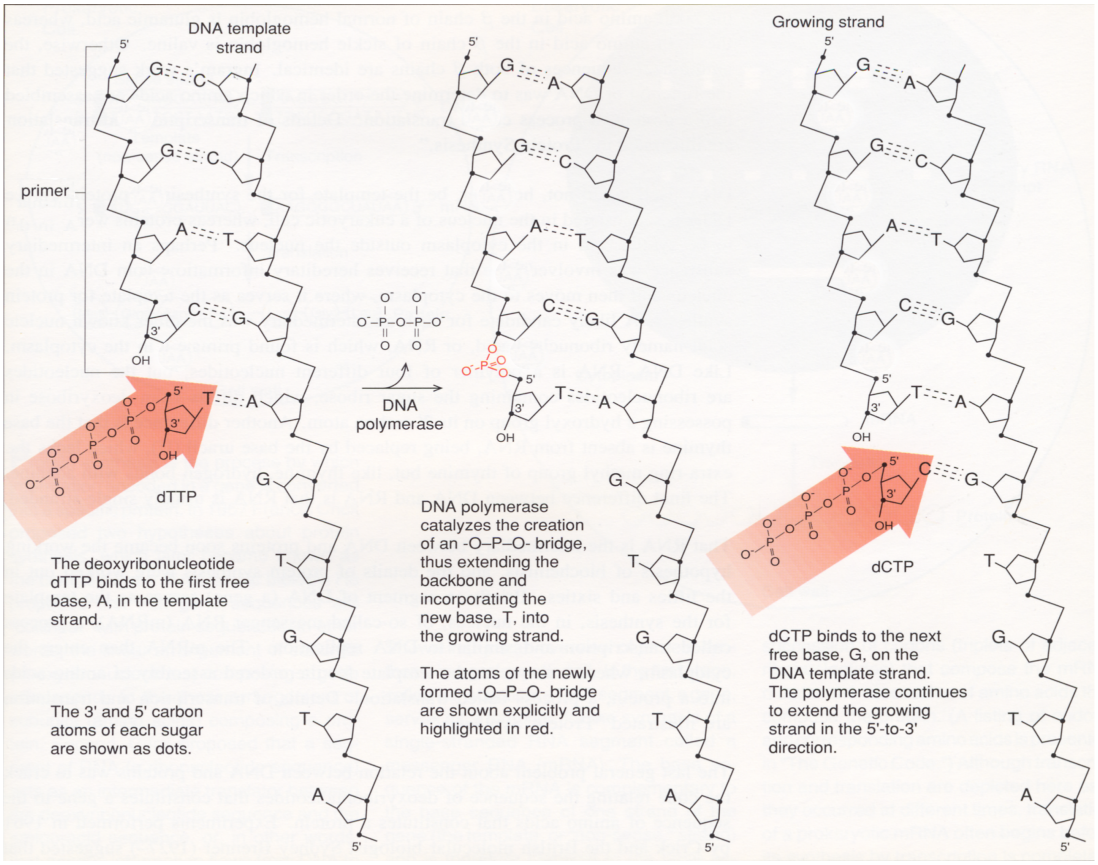


Fig. 1.2. (A) A cartoon illustrating the double helical structure of DNA. Each strand is supported by the sugar-phosphate backbone. They are held together via hydrogen bonds in anti-parallel fashion (the 5’ end of one strand aligns with the 3’ end of the other one). (B) A figure depicting two DNA strands held together by hydrogen bonds between paired bases. (C) More detailed chemical structure showing the hydrogen bonding between bases.

DNA undergoes two major biological reactions: transcription and replication. Transcription is the synthesis of RNA under the direction of DNA. The transcribed messenger RNA is then translated into an amino acid sequence to complete the flow of genetic information. DNA replication takes place during cell division to transmit the genetic material to the new daughter cells. A precise replication process is needed to guarantee the conservation of genetic information in each new generation of cells. The structural property of DNA enables such a highly accurate copying mechanism. During replication, the two strands of the double helix first unwind and separate. Each strand serves as a template for the synthesis of its new complementary DNA, resulting in the production of two identical daughter strands. This semiconservative nature of replication assures the preservation of our hereditary genomes. Many proteins are involved in DNA replication. Among them, DNA polymerase plays a vital role in synthesizing the new daughter strand. This enzyme catalyzes the addition of complementary nucleotides to the 3'-hydroxyl terminus of the growing DNA stand. It facilitates the formation of a phosphodiester bond through a nucleophilic attack by the terminal 3'-OH group of the primer strand on the α phosphorus atom of the incoming dNTP, releasing a pyrophosphate (PPi) as the by-product (Fig. 1.3).

Inside cells, DNA polymerase is highly accurate, incorporating the incoming nucleotide only if it is complementary to the base on the template, and extremely efficient, with the ability to process millions of bases at an error rate less than 10^{-8} per base in hours.[17] Due to the speed and accuracy of the DNA replication process, it has become the foundation for most of the current and emerging sequencing technologies. The following sections will review these sequencing methods in detail.



The deoxyribonucleotide dTTP binds to the first free base, A, in the template strand.

The 3' and 5' carbon atoms of each sugar are shown as dots.

DNA polymerase catalyzes the creation of an -O–P–O- bridge, thus extending the backbone and incorporating the new base, T, into the growing strand.

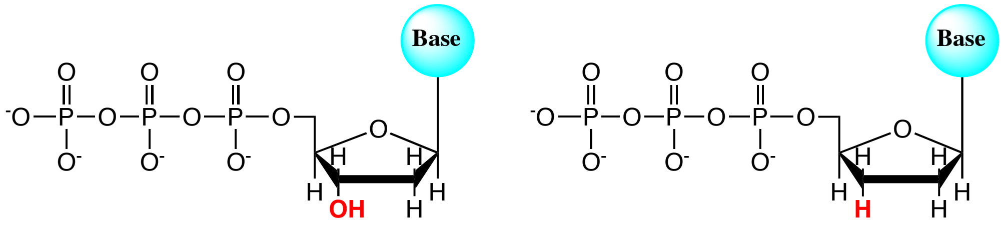
The atoms of the newly formed -O–P–O- bridge are shown explicitly and highlighted in red.

dCTP binds to the next free base, G, on the DNA template strand. The polymerase continues to extend the growing strand in the 5'-to-3' direction.

Fig. 1.3. Scheme of DNA polymerase reaction. DNA synthesis takes place via the addition of a nucleotide to the 3'-OH end of a DNA primer strand. The base-pairing between the incoming nucleotide and the DNA template strand dictates which nucleotide is added. DNA polymerase facilitates the addition of the incoming nucleotide by catalyzing the formation of a phosphodiester bond between the terminal 3'-OH group of the primer strand and the alpha phosphorus atom of the nucleotide. A pyrophosphate (PPi) group is released as a by-product.

1.2.1. Sanger dideoxynucleotide sequencing

One of the true pioneers in DNA sequencing technology is Frederick Sanger. Sanger and his colleagues developed a DNA sequencing method based on the generation of DNA ladder fragments using 2',3'-dideoxyribonucleotide (ddNTP) chain-termination in 1977.[18] The nucleotide analogs, ddNTPs, are terminators of the polymerase reaction due to the lack of a hydroxyl group on the 3' position on the sugar moiety compared to the natural dNTPs (Fig. 1.4).



2'-deoxyribonucleotide (dNTP)

2', 3'-dideoxyribonucleotide (ddNTP)

Fig. 1.4. Chemical structures of 3'-deoxyribonucleotide (dNTP) and 2', 3'-dideoxyribonucleotide (ddNTP). Since ddNTPs do not have the 3'-OH group, which is necessary for DNA synthesis, they terminate further extension of the DNA strand once incorporated.

The absence of this 3'-OH group in the ddNTPs significantly changes the dynamics of the DNA polymerase reaction. The phosphodiester bond simply cannot form with the incoming nucleotide once a ddNTP is incorporated, and DNA synthesis is terminated. Therefore, when a mixture of dNTPs and a miniscule amount of ddNTPs is used for DNA replication reactions, the resulting DNA fragments will be of various length, all terminated by 2,'3'-dideoxy nucleotide analogs at the 3' ends (Fig. 1.5). Gel electrophoresis is used to separate these fragments based on length with single base resolution. To facilitate the process of identifying the sequence of target DNA, each of the four ddNTPs analogs are labeled with four different fluorescent dyes. The sequence of the DNA is then obtained by reading the color of the electrophoretic bands in length order.

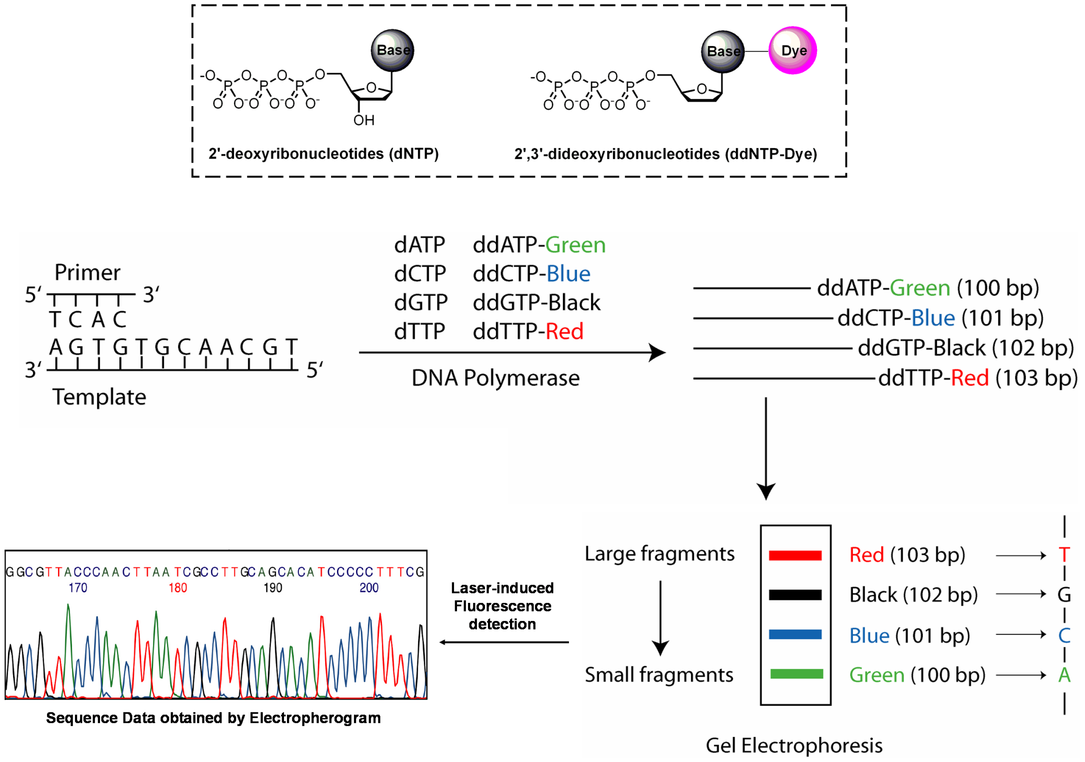


Fig. 1.5. Sanger dideoxy chain-termination sequencing method. DNA fragments are generated by extending the primer with a mixture of dNTPs and ddNTPs. Upon incorporation of a ddNTP, the DNA strand ceases to participate in polymerase reaction due to the lack of the 3'-OH group. Thus, a mixture of DNA strands with different length complementary to the template DNA is produced. To determine the sequence of the template, these DNA fragments are separated based on size by electrophoresis, and the resulting bands of DNA are detected by their fluorescent signals.

For over three decades, the Sanger dideoxy chain-termination method has been the technique of choice for large-scale DNA sequencing projects. Many aspects of the original Sanger sequencing process have evolved to a high-level of automation by introducing engineered DNA polymerases,[19] capillary electrophoresis,[4] and laser induced fluorescent excitation of energy transfer dyes.[5] Among the numerous improvements, application of the laser induced fluorescent energy transfer dyes was a major advancement that made large-scale genome sequencing initiatives possible. An "ideal" set of fluorophores for 4-color DNA sequencing consists of four different fluorophores. These fluorophores should have similar high molar absorbance at a common excitation wavelength, high fluorescence quantum yields, strong and well-separated fluorescence emissions, and constant relative mobility shifts for the DNA sequencing fragments. These criteria cannot be met optimally by the spectroscopic properties of single fluorescent dye molecules, and indeed were poorly satisfied by the initially used sets of fluorescent tags. Ju *et al.* overcame these obstacles imposed by the use of single dyes and developed fluorescence energy transfer dyes for DNA sequencing that fulfilled the performance criteria set out above.[5] The higher sensitivity offered by these new sets of fluorescent dyes also allowed the direct sequencing of large DNA templates (> 30 kilo bases) with read lengths of over 700 bases per sequencing reaction, leading to significant progress in the large scale genome sequencing and mapping projects.[20-22]

However, Sanger sequencing has its limitations due to the decreasing resolving power of the gel as fragment length increases[23] and the overlapping emissions of even the best fluorophore combinations. Currently, the automated Sanger sequencing method can allow up to 384 samples to be analyzed simultaneously, which largely limits its output. High cost associated with this technology poses another major obstacle for analyzing and comparing entire genomes.

1.2.2. MALDI-TOF MS based DNA sequencing

DNA sequencing by mass spectrometry (MS) is a sequencing method that involves the detection of DNA ladder fragments based on their masses. The concept was originally proposed in the early 1990s[24, 25] and developed by several research groups in the following decade[9, 10, 26-28]. The core rationale behind sequencing by MS is to use a mass spectrometer as the mean of separating and detecting DNA fragments instead of gel electrophoresis in Sanger sequencing. Implementing of the system does not require fluorescent labeling of terminators, hence significantly reducing the cost of sequencing reagents. As in Sanger sequencing, the same DNA ladder generation method, which involves the incorporation of ddNTPs, is carried out. The resulting sequencing fragments are then isolated and purified in preparation for MS analysis. DNA sequences are determined based on the generated mass spectrum. Since all four nucleotide analogs have unique masses, peak (and sequence) identification is achieved by comparing observed mass and expected nucleotide mass values.

Two principal criteria for DNA sequencing by MS are the resolution of mass spectrometric instrumentation and the purity of DNA fragments. Since conventional mass spectrometers lack the sensitivity and range to meet the first criterion, the powerful matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has been used for rapid and accurate analysis of DNA fragments due to its high mass range and resolution (Fig. 1.6). In order to detect the mass of DNA, a matrix solution composed of UV or IR absorbing organic molecules is mixed with the target fragments. The mixture is left to crystallize on a sample plate. Upon bombardment by a laser, the matrix is first ionized, causing charge transfer to the DNA fragments through collision. The ionized DNA fragments are released from the surface of the plate and introduced into the evacuated flight tube along with the matrix. An electrical field is set up inside the long tube so that the fragments fly toward the ion detector at different rates according to their mass to charge ratio. Lower mass molecules arrive at the detector faster than their heavier counterparts. Hence the flight time is used to determine the masses of target DNA fragments and a mass spectrum of intensity versus mass to charge ratio (m/z) is reported. Current generation MALDI-TOF MS is capable of detecting masses ranging from 500 to 20,000 Daltons, making it suitable to handle the masses of shorter DNA fragments.

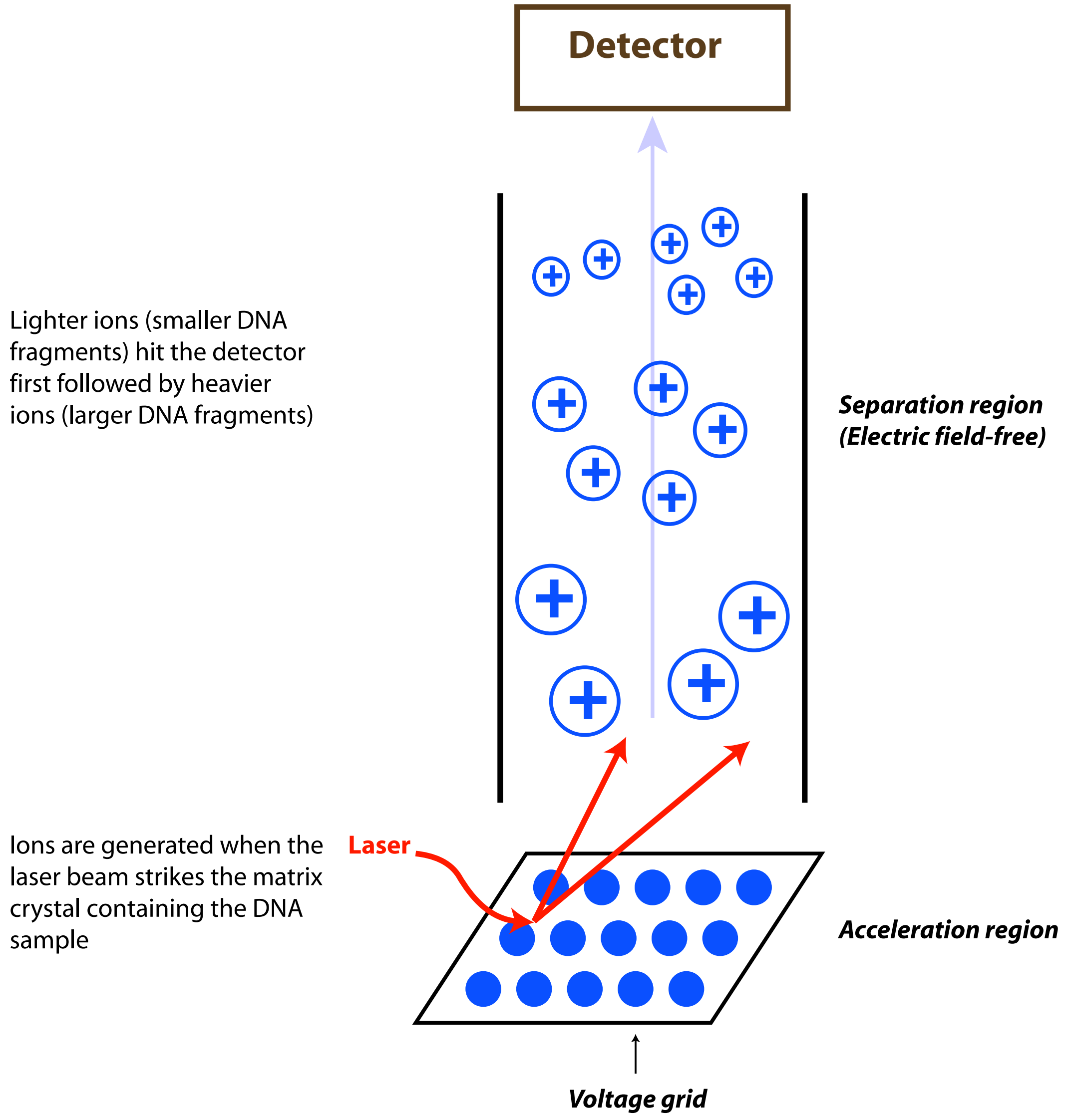

Detector
Lighter ions (smaller DNA fragments) hit the detector first followed by heavier ions (larger DNA fragments)
Separation region (Electric field-free)
Ions are generated when the laser beam strikes the matrix crystal containing the DNA sample
Laser
Acceleration region
Voltage grid

Fig. 1.6. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). A mixture of analyte (e.g. DNA sequencing fragments) and matrix molecules (blue)

are spotted on the sample plate and allowed to co-crystallize prior to loading into the vacuum chamber. After UV laser irradiation, the desorbed and ionized analyte and matrix molecules are accelerated under a constant electric voltage, causing them to fly towards the detector. The charged molecules arrive at the detector at different times based on their masses. Therefore, the masses of the charged particles can be determined from their time-of-flight.

The second important requirement for DNA analysis by MS is the purity of samples introduced into the mass spectrometer. As the products of Sanger sequencing reactions, the DNA fragments are often contaminated by alkaline earth salts (Na+, K+, etc.) that are included in the polymerase reaction buffer and falsely terminated fragments (those terminated with dNTP instead of ddNTP). This problem is tackled by subjecting the samples to desalting procedures[27] and molecular affinity purification of ladders that have incorporated biotin-labeled ddNTPs on streptavidin columns (Fig. 1.7B).[11, 28, 29] With drastically improved purity of DNA fragments, there have been attempts to push the read length of MS-based sequencing up to 100 bases.[28] Several limitations still exist to prevent sequencing by MS from becoming a high-throughput sequencing method. Among the chief obstacles are the inability to resolve larger DNA fragments and high cost associated with sample preparation and purification. With a reliable read length around 30 bases, MALDI-TOF MS sequencing is more suitable for directed mutation screening, especially in insertion/deletion regions.

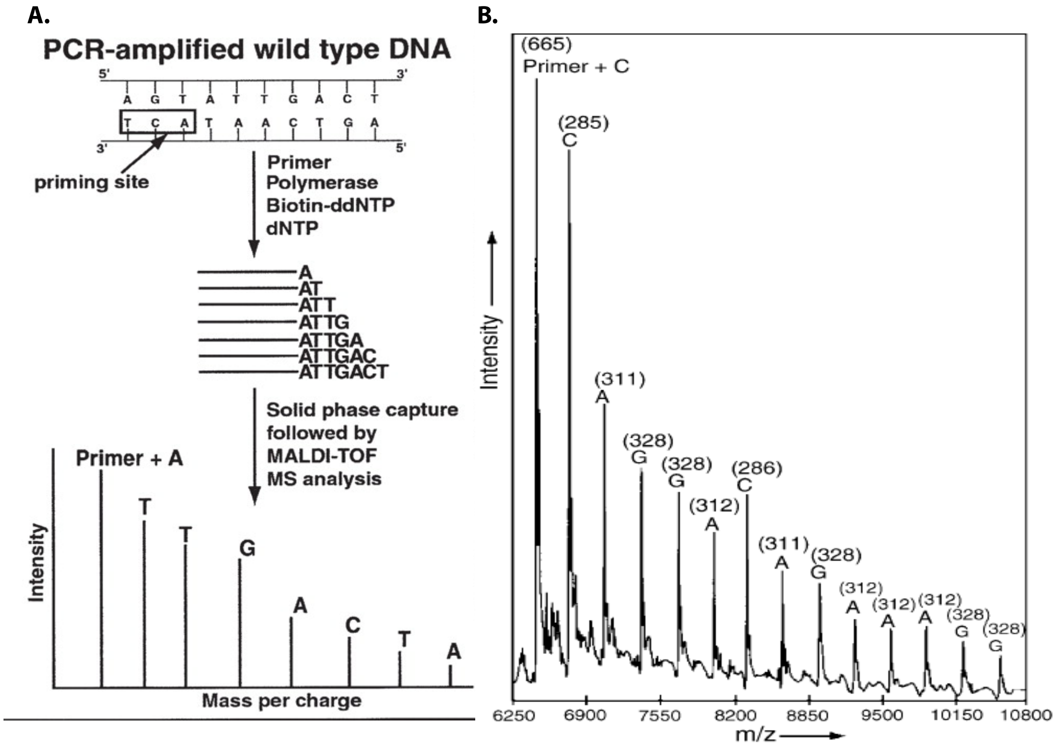


Fig. 1.7. DNA sequencing using MALDI-TOF MS. (A) Sanger sequencing fragments generated using biotin-labeled ddNTPs. (B) Example of mass sequencing spectrum using biotin-labeled ddNTPs.

1.2.3. Pyrosequencing

While Sanger sequencing and sequencing by MALDI-TOF MS became widely popular in the genomic field, a trend is emerging to move sequencing platforms away from the costly generationand resolution of DNA ladder fragments. Pyrosequencing, a method belonging to the broader group of the sequencing paradigm known as sequencing by synthesis (SBS), was first established in 1988 as an attempt to sequence DNA as the strand is synthesized. It reveals the sequence of DNA by detecting the pyrophosphate group that is released when a nucleotide is incorporated during the DNA polymerase reaction.[30] In this approach, each of the four dNTPs is added sequentially with a cocktail of enzymes, substrates, and the usual polymerase reaction components (Fig. 1.8). If the added nucleotide is complementary to the first available base on the template, the nucleotide will be incorporated and a pyrophosphate will be released. Through an enzyme cascade, the released pyrophosphate is converted to ATP, and then turned into a visible light signal by firefly luciferase. On the other hand, if the added nucleotide is not incorporated, no light will be produced and the nucleotide will simply be degraded by the enzyme apyrase.

Pyrosequencing has been applied to single nucleotide polymorphism (SNP) detection[12] and DNA sequencing.[13] In 2005, 454 Life Science Corp. developed a commercial sequencing platform combining pyrosequencing and DNA template amplification on individual microbeads for high-throughput DNA sequencing.[13] However, there are inherent difficulties in pyrosequencing for determining the number of incorporated nucleotides in homopolymeric regions (e.g. a string of several T's in a row) of the template. In theory, homopolymeric regions can be deciphered since the signal intensity is directly proportional to the number of bases added. However, this relationship breaks down for regions with more than 4 or 5 bases.[12] Wu *et al.* have solved this problem by using nucleotide reversible terminators to decipher the homopolymeric regions in pyrosequencing.[31] But other aspects of pyrosequencing still need improvement. For example, each of the four nucleotides has to be added and detected separately. The accumulation of undegraded nucleotides and other components could also lower the accuracy of the method when sequencing a long DNA template. Ideally, as one examines the fundamental limitations towards miniaturization, one would prefer a simple method to directly detect a reporter group attached to the nucleotide that is incorporated into a growing DNA strand during the polymerase reaction rather than relying on a complex enzymatic cascade.

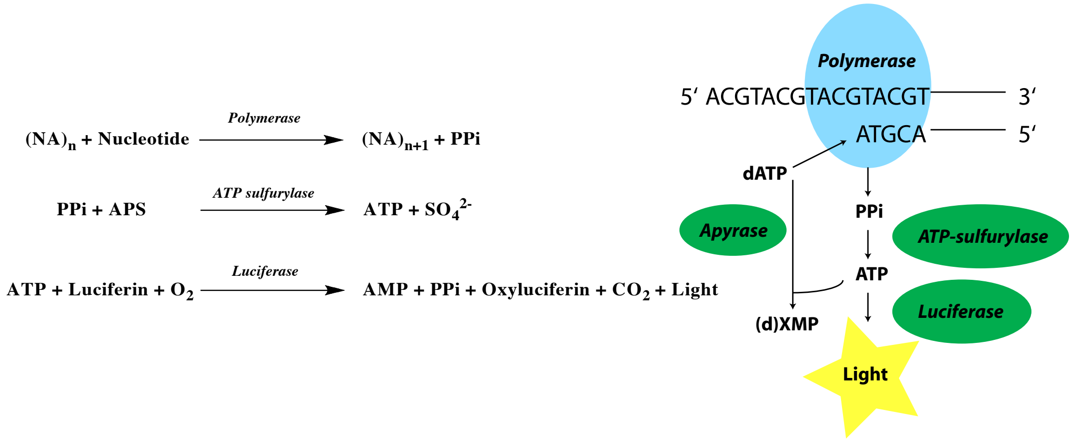

(NA)n + Nucleotide —Polymerase→ (NA)n+1 + PPi
PPi + APS —ATP sulfurylase→ ATP + SO4 2-
ATP + Luciferin + O2 —Luciferase→ AMP + PPi + Oxyluciferin + CO2 + Light
Polymerase
5′ ACGTACGTACGTACGT 3′
ATGCA 5′
dATP
PPi
ATP
(d)XMP
Apyrase
ATP-sulfurylase
Luciferase
Light

Fig. 1.8. General scheme of pyrosequencing. As the polymerase catalyzes the incorporation of nucleotide(s) into a growing strand of DNA, PPi molecules are released and then converted to ATPs by ATP-sulfurylase. The ATPs participate in the luciferase reaction in which a luciferin molecule is oxidized to produce oxyluciferin and light. The resulting luminescence can be registered using a photon detector.

1.2.4. DNA sequencing by ligation

In 2005, George Church and his colleagues at Harvard Medical School presented their large-scale DNA sequencing method termed sequencing by ligation.[7] They used emulsion PCR to amplify sequencing templates on 1-micron beads. After the beads were allowed to self-assemble into a monolayer on a glass surface, a polyacrylamide gel was polymerized around the beads to fix their positions. Reagents were flowed through the gel to carry out the sequencing by ligation reactions. Sequencing by ligation consists of four general steps in each sequencing cycle (Fig. 1.9). In order to initiate the cycle, an anchor primer is first hybridized to the template strands immobilized on the beads. A set of four fluorophore-labeled degenerate nonamers competes for ligation to the anchor primer. After the ligation, the fluorescent signal is detected on an epifluorescent microscope to identify the base. Following the detection step, the fluorophore is cleaved to continue the next cycle of ligation, detection, and cleavage. Finally, after a series of ligation cycles, the extended primer is removed and a new primer is hybridized to the n-1 position for a second round of ligation cycles.

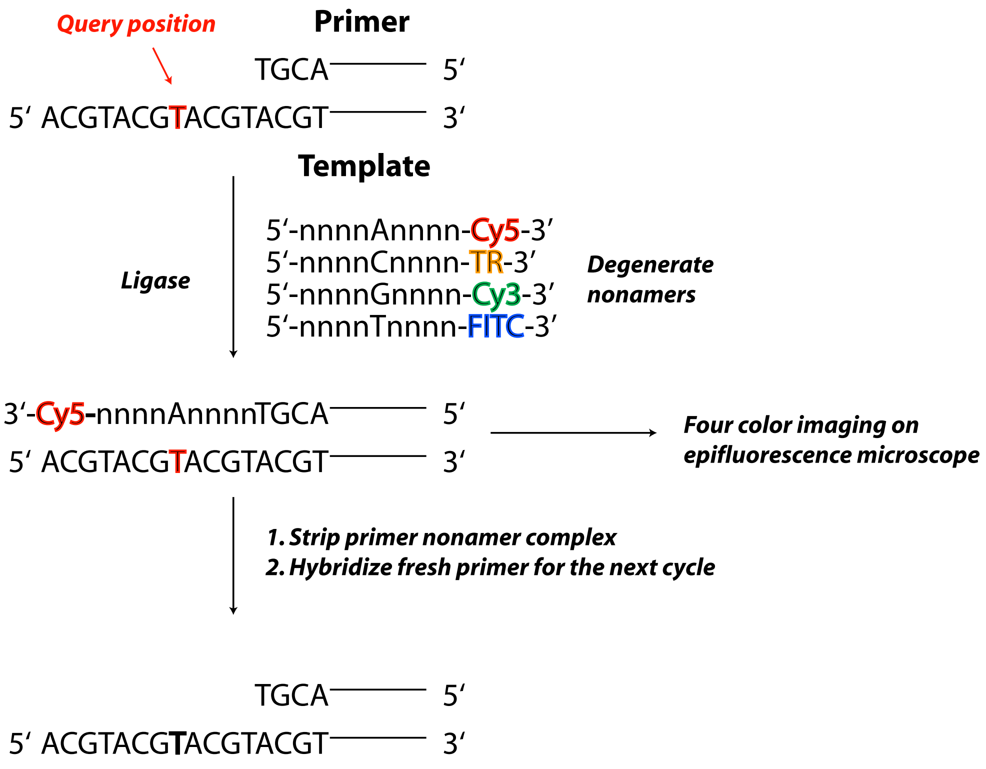


Fig. 1.9. Scheme of DNA sequencing by ligation using degenerate nonamers.

One of the features of sequencing by ligation, as developed by Church and commercialized by ABI Life Sciences, is the large number of beads that can be placed on the surface. This is directly responsible for more than a 600-fold increase in the number of templates that can be simultaneously sequenced. However, since the method relies on single base discrimination from a primer sequence during the ligation step, the read length has been limited to 6-7 bases per cycle. In addition, with increasing distance of the nucleotide to be sequenced from the ligation point, the ligase’s ability to discriminate falls drastically, along with accuracy.

1.2.5. DNA sequencing by engineered nanopores

Recently, a unique method that utilizes ion channels (nanopores) for detection of individual DNA or RNA strands has been proposed for sequencing purposes. The concept of nanopore sequencing was first explored by Kasianowicz and co-workers when they demonstrated the possibility of using an α-hemolysin channel to detect nucleic acids at the single molecule level.[8] The α-hemolysin, an exotoxin secreted by a bacterium, self-assembles into a lipid bilayer membrane to form a heptameric pore with a 2.6 nm-diameter vestibule and 1.5 nm-diameter limiting aperture (Fig. 1.10).[32-34]

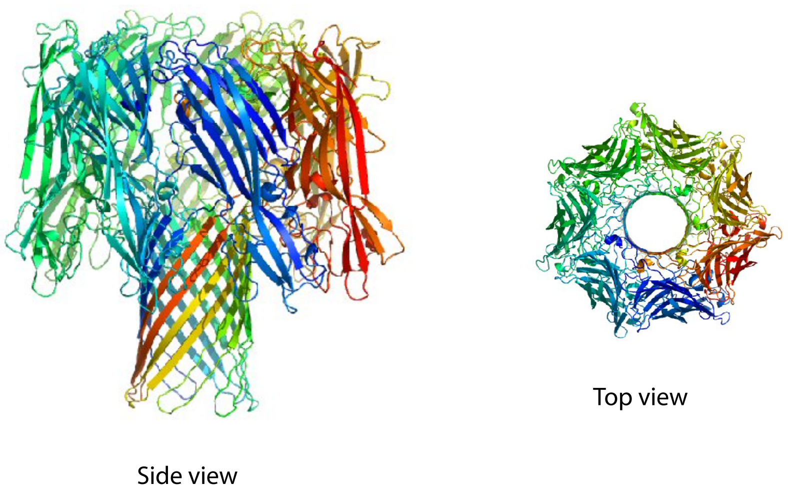


Fig. 1.10. 3-D rendering of α-hemolysin structure.

The pore was able to conduct a strong and steady ionic current when an appropriate voltage was applied across the membrane in the presence of an aqueous salt solution such as KCl. Due to the size of the limiting aperture, only single-stranded polyanionic nucleic acid molecules could pass through the pore when an electric field was applied. When such an event took place, the otherwise steady ionic current was either blocked or reduced, generating an electronic signature.[34, 35] It was expected that different nucleic acids would create characteristic signatures based on their steric and electronic properties to distinguish them from each other and uncover their identities. A plot of translocation time versus blockade current (Fig. 1.11) would be produced to reveal the composition of the polynucleotides based on translocation current, translocation duration, and corresponding dispersion.[32]

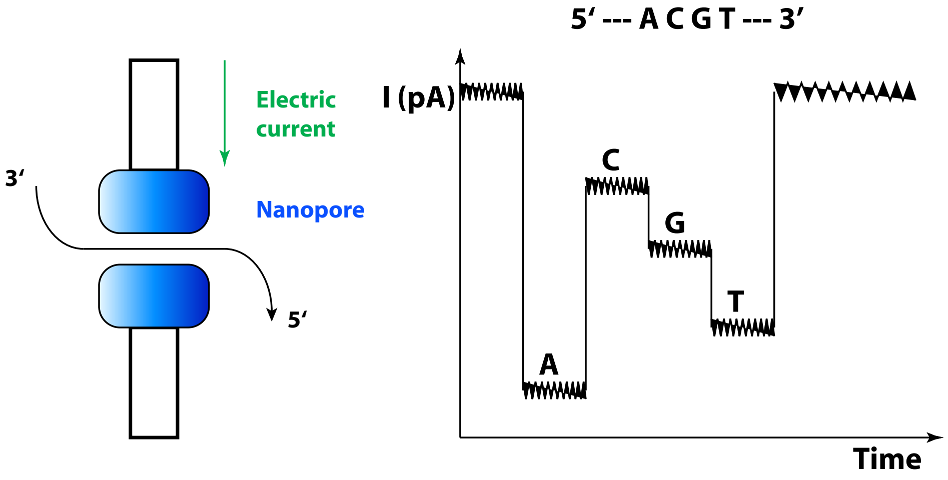


Fig. 1.11. Scheme for DNA sequencing using nanopore and theoretical plot of translocation time versus blockade current elicited by DNA strand.

The concept of single DNA molecule detection using a nanopore is very well established. However, the effort to implement the idea into a robust base-to-base sequencing platform has met several obstacles. At this point, it is still very difficult to achieve accurate and obvious discrimination between the four bases (A, C, G, and T) due to the strength and resolution of signals. In addition, the translocation rate of nucleic acids through the pore is sporadic, which also contributes to the discrepancy in base-to-base discrimination. The nanopores themselves have yet to be improved through synthesis and engineering for better differentiation of the four different nucleotides. Therefore, DNA sequencing by nanopore is still at the early proof of principle stage. While gaining much attention and publicity due to its potential, the nanopore sequencing approach has yet to make the leap to become a viable commercial system.[36, 37]

1.3 Conclusion

Since Sanger's introduction of his dideoxynucleotide chain termination method, the field of DNA sequencing has rapidly evolved from a lab-bench based experiment to a potential multi-billion dollar industry. In addition to the sequencing technolgies described in the previous sections, two platforms for single molecule DNA sequencing by synthesis have been reported recently. In the Helicos Heliscope sequencer,[38] single-molecule DNA templates are directly immobilized on a solid surface to allow SBS to take place. Pacific Biosciences SMRT single molecule sequencing system[39] captures the polymerase rather than the DNA library on the surface to carry out real-time sequence determination. The National Human Genome Research Institute (NHGRI) has mandated new sequencing technologies to reduce the cost of the current Sanger based method by 100 fold for the near term ($100K Genome) and by a further 10,000 fold in the near future ($1,000 Genome). Among the new sequencing technologies, sequencing by synthesis (SBS) approaches such as pyrosequencing and SBS with cleavable fluorescent nucleotide reversible terminators (CF-NRTs) have shown much promise for the so-called next generation sequencing. All the methods are capable of highly parallel readouts and identifying the DNA sequence with high accuracy. Based on some of the ideas formulated in the author's laboratory, this thesis will focus on the development of SBS using CF-NRTs. It is desirable to push this particular promising platform towards the $1000 Genome goal through innovative molecular engineering.

1.4 References

1. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W. *et al.* **Initial sequencing and analysis of the human genome.** *Nature* **2001**, *409***,** 960-921.
2. Collins, F. S.; Green, E. D.; Guttmacher, A. E.; Guyer, M. S. A vision for the future of genomics research. *Nature* **2003**, *422*, 835–847.
3. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431,* 931-945.
4. Smith, L. M.; Sanders, J. Z.; Kaiser, R. J.; Hughes, P.; Dodd, C.; Connell, C. R.; Heiner, C.; Kent, S. B.; Hood, L. E. Fluorescence detection in automated DNA sequence analysis. *Nature* **1986**, *321*, 674-679.
5. Ju, J.; Ruan, C.; Fuller, C. W.; Glazer, A. N.; Mathies, R. A. Fluorescence energy transfer dy-labeled primers for DNA sequencing and analysis. *Proc Natl Acad Sci USA* **1995**, *92*, 4347-4351.
6. Kan, C. W.; Doherty, E. A.; Barron, A. E. Thermoresponsive N,N-dialkylacrylamide copolymer blends as DNA sieving matrices with a thermally tunable mesh size. *Electrophoresis* **2003**, *24*, 4161-4169.
7. Shendure, J.; Porreca, G. J.; Reppas, N. B.; Lin, X.; McCutcheon, J. P.; Rosenbaum, A. M.; Wang, M. D.; Zhang, K.; Mitra, R. D.; Church, G. M. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* **2005**, *309,* 1728-1732.
8. Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **1996**, *93*, 13770-13773.
9. Fu, D. J.; Tang, K.; Braun, A.; Reuter, D.; Iverson, B. L.; Darnhofer-Demar, B.; Little, D. P.; O'Donnell, M. J.; Cantor, C. R.; Koster, H. Sequencing exons 5 to

8 of the *p53* gene by MALDI-TOF mass spectrometry. *Nat Biotechnol* **1998**, *16*, 381-384.

10. Roskey, M. T.; Juhasz, P; Smirnov, I. P.; Takach, E. J.; Martin, S. A.; Haff, L. A. DNA sequencing by delayed extraction-matrix-assisted laser desorption/ionizaton time of flight mass spectrometry. *Proc Natl Acad Sci U S A* **1996**, *93*, 4724–4729.

11. Edwards, J. R.; Itagaki, Y.; Ju, J. DNA sequencing using biotinylated dideoxynucleotides and mass spectrometry. *Nucleic Acids Res* **2001**, *29*, E104.

12. Ronaghi, M.; Uhlen, M.; Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **1998**, *281*, 363-365.

13. Margulies, M.; Egholm, M.; Altman, W. E.; Atiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376-380.

14. Seo, T. S.; Bai, X.; Kim, D. H.; Meng, Q.; Shi, S.; Ruparel, H.; Li, Z.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci U S A* **2005**, *102,* 5926-5931.

15. Stryer, L. *Biochemistry, 4th Edition* **1995**, *W.H. Freeman and Co., New York.*

16. Watson, J. D.; Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737-738.

17. Alberts, B.; Bray, D.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walkter, P. *Essential Cell Biology* **1998**, *Garland Publishings, Inc., New York & London.*

18. Sanger, F.; Nicklen, S.; Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **1977**, *74,* 5463-5467.

19. Tabor, S.; Richardson, C. C. A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc Natl Acad Sci U S A* **1995**, *92*, 6339-6343.

20. Marra, M.; Weinstock, L. A.; Mardis, E. R. End sequence determination from large insert clones using energy transfer fluorescent primers. *Genome Res* **1996**, *6,* 1118-1122.

21. Lee, L. G., *et al*. New energy transfer dyes for DNA sequencing. *Nucleic Acids Res* **1997**, *25,* 2816-2822.

22. Heiner, C. R., Hunkapiller K. L., Chen, S., Glass, J.I. & Chen, E.Y. Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res* **1998**, 8*,* 557-561.

23. Bowling, J. M.; Bruner, K. L.; Cmarik, J. L.; Tibbetts, C. Neighboring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Res.* **1991**, *19*, 3098-3097.

24. Smith, L. M. The future of DNA sequencing. *Science* **1993**, *262*, 530-532.

25. Romano, L. J.; Levis, R. J. Nondestructive laser vaporization of high molecular weight, single-stranded DNA. *J Am Chem Soc* **1991**, *113*, 9665-9667.

26. Shaler, T. A.; Tan, Y.; Wickham, J. N.; Wu, K. J.; Becker, C. H. Analysis of enzymatic DNA sequencing reactions by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **1995**, *9*, 942-947.

27. Mouradian, S.; Rank, D. R.; Smith, L. M. Analyzing sequencing reactions from Bacteriophage M13 by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **1996**, *10*, 1475-1478.

28. Monforte, J. A.; Becker, C. H. High-throughput DNA analysis by time-of-flight mass spectrometry. *Nature Med* **1997**, *3*, 360-362.

29. Ruparel, H.; Ulz, M. E.; Kim, S.; Ju, J Digital detection of genetic mutations using SPC-sequencing. *Genome Res.* **2004**, *14,* 296-300.

30. Hyman, E. D. A new method of sequencing DNA. *Anal Biochem* **1988**, *174,* 423-436.

31. Wu, J., *et al.* 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* **2007**, *104,* 16462-16467.

32. Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D. Rapid nanopore discrimination between single oligonucleotide molecutes. *Proc Natl Acad Sci USA* **2000**, *97*, 1079-1084.

33. Akeson, M.; Branton, D.; Kasianowicz, J. J.; Brandin, E.; Deamer, D. W. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys J* **1999**, *77*, 3227-3233.

34. Deamer, D. W.; Branton, D. Characterization of nucleic acids by nanopore analysis. *Acc Chem Res* **2002**, *35*, 817-825.

35. Vercoutere, W.; Winters-Hilt, S.; Olsen, H.; Deamer, D.; Haussler, D.; Akeson, M. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nat Biotech* **2001**, *19*, 248-252.

36. Shendure, J.; Mitra, R. D.; Varma, C.; Church, G. M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **2004**, *5*, 335-344.

37. Rhee, M.; Burns, M.A. Nanopore sequencing technology: research trends and applications. *TRENDS in Biotechnology* **2006**, *24*, 580-586.

38. Harris, T. D.; Buzby, P. R.; Babcok, H.; Beer, E.; Bowers, J.; Braslavsky, I.; Causey, M.; Colonell, J.; DiMeo, J.; Efcavitch, J. W.; Giladi, E.; Gill, J.; Healy,

J.; Jarosz, M.; Lapen, D.; Moulton, K.; Quake, S. R.; Steinmann, K.; Thayer, E.; Tyurina, A.; Ward, R.; Weiss, H.; Xie, Z. Single-molecule DNA sequencing of a viral genome. *Science* **2008**, *320*, 106-109.

39. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; *et al*. Real-time DNA sequencing from single polymerase molecule. *Science* **2008**, *323*, 133-138.

Chapter 2: Overview of DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators

2.1 Introduction

Among the novel approaches for DNA sequencing, the sequencing by synthesis (SBS) approach has emerged as a viable candidate for a massively parallel high throughput sequencing platform. SBS takes advantage of the polymerase reaction, a key process for DNA replication inside cells. The basic concept of SBS is to use DNA polymerase to extend a primer that is hybridized to a template by a single nucleotide, determine its identity, and then proceed to the next nucleotide, eventually reading out the entire DNA sequence serially. In contrast to Sanger sequencing, in which fluorescently labeled DNA fragments of different sizes are all generated in a single reaction and then separated and detected, SBS approaches have an advantage in that individual bases are detected consecutively and multiple templates simultaneously without the need for separation. Thus, SBS can easily scale-up over Sanger's dideoxy-sequencing techniques. Currently array scanners already exist that can easily detect over 100,000 sample spots arrayed on a glass surface.[1] Advanced array scanners enable fast screening of large areas with high resolution, allowing automated detection of hundreds of thousands and even millions of samples simultaneously.

2.2 General Methodology for DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators

The concept of DNA sequencing by synthesis (SBS) was first established in 1988 with an attempt to sequence DNA by detecting the pyrophosphate group that is released when a nucleotide is incorporated during the DNA polymerase reaction.[2] Pyrosequencing, which was developed based on this concept, has been applied to single nucleotide polymorphism (SNP) detection[3] and DNA sequencing[4] as described in the previous chapter.[5] However, there are inherent difficulties in this method for determining the number of incorporated nucleotides in homopolymeric regions (e.g. a string of several T's in a row) of the template. Wu *et al.* have solved this problem by using nucleotide reversible terminators to decipher the homopolymeric regions for pyrosequencing.[6] Yet, other aspects of pyrosequencing still need improvement. Ideally, it is desirable to develop a simple method to directly detect a reporter group attached to the nucleotide that is incorporated into a growing DNA strand during the polymerase reaction rather than relying on a complex enzymatic cascade.

Ju *et al.* have developed an integrated SBS approach for a high throughput sequencing platform as shown in Fig. 2.1.[7] This method relies on using the polymerase reaction to read out the DNA sequence through the incorporation of novel reporter nucleotides (Fig. 2.2). After each nucleotide is added, the attached reporter group is detected to determine the identity of the added nucleotide. In order to temporarily pause the sequencing reaction and to accurately sequence through homopolymeric regions, the 3' hydroxyl group of the nucleotide must be blocked by a moiety to stop the polymerase reaction during the identification of the added nucleotide. This blocking group then needs to be easily removed to regenerate a free hydroxyl group for subsequent extension. In order to design an ideal system for SBS, new nucleotide analogues, termed cleavable fluorescent nucleotide reversible terminators (CF-NRTs), with the above properties must be developed.

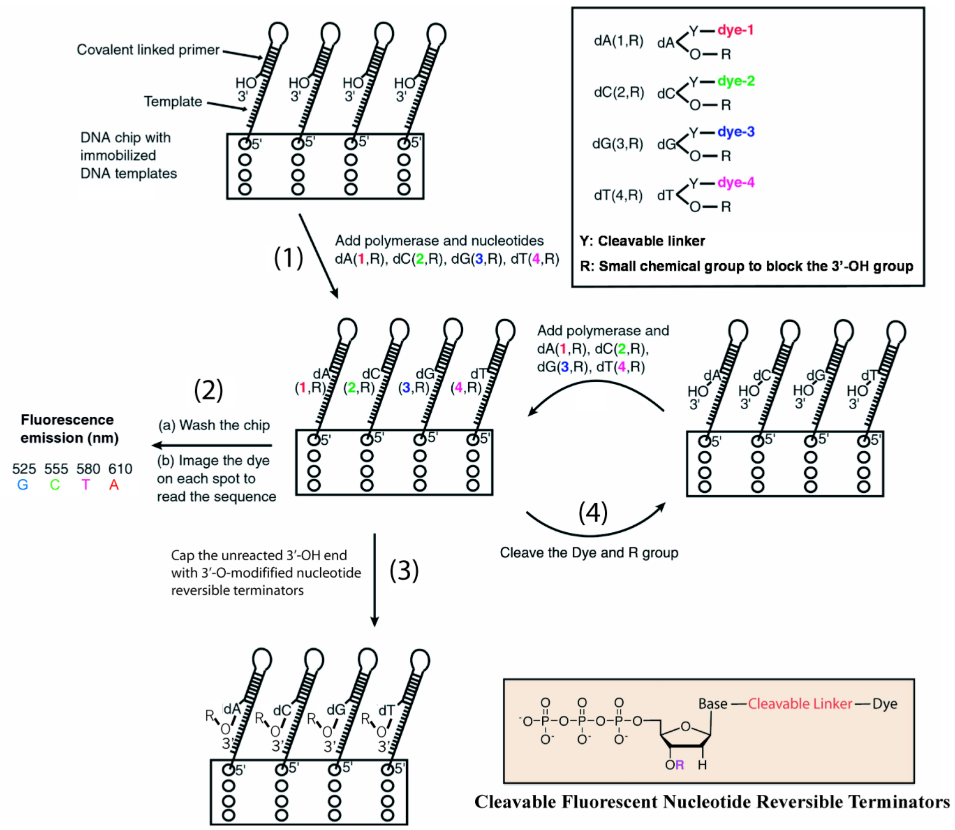

Covalent linked primer
Template
HO 3'
DNA chip with immobilized DNA templates
5'
dA(1,R) dA Y—dye-1 O—R
dC(2,R) dC Y—dye-2 O—R
dG(3,R) dG Y—dye-3 O—R
dT(4,R) dT Y—dye-4 O—R
Y: Cleavable linker
R: Small chemical group to block the 3'-OH group
(1)
Add polymerase and nucleotides dA(1,R), dC(2,R), dG(3,R), dT(4,R)
dA (1,R) dC (2,R) dG (3,R) dT (4,R)
(2)
(a) Wash the chip
(b) Image the dye on each spot to read the sequence
Fluorescence emission (nm)
525 555 580 610
G C T A
Add polymerase and dA(1,R), dC(2,R), dG(3,R), dT(4,R)
(4)
Cleave the Dye and R group
Cap the unreacted 3'-OH end with 3'-O-modified nucleotide reversible terminators
(3)
R O 3'
Base — Cleavable Linker — Dye
OR H
Cleavable Fluorescent Nucleotide Reversible Terminators

Fig. 2.1. In the SBS approach, a chip is constructed with immobilized DNA templates that are able to self-prime for initiating the polymerase reaction. Four nucleotide analogues are designed such that each is labeled with a unique fluorescent dye on the specific location of the base, and a small chemical group (R) to cap the 3'-OH group. Upon adding the four nucleotide analogues and DNA polymerase, only the nucleotide analogue complementary to the next nucleotide on the template is incorporated by polymerase on each spot of the chip (step 1). After removing the excess reagents and washing away any unincorporated nucleotide analogues, a 4-color fluorescence scanner is used to image the surface of the chip, and the unique fluorescence emission from the specific dye on the nucleotide analogues on each spot of the chip will yield the identity of the nucleotide (step 2). After imaging, the small amount of unreacted 3'-OH group on the self-primed template moiety will be capped by excess 3'-O-modified nucleotide reversible terminators and DNA polymerase to avoid interference with the next round of synthesis for synchronization (step 3). The dye moiety and the R protecting group will be removed to generate

a free 3'-OH group with high yield (step 4). The self-primed DNA moiety on the chip at this stage is ready for the next cycle of the reaction by repetition of steps 1, 2, 3, and 4 to identify the next nucleotide in the sequence of the template DNA.

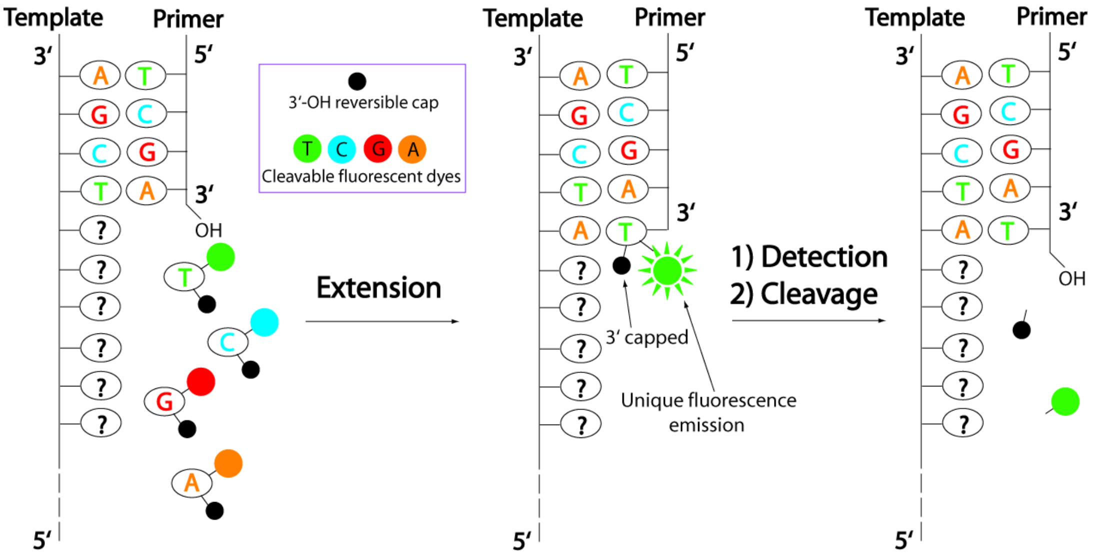


Fig. 2.2. DNA sequencing by synthesis based on modified nucleotide analogues. First, the primer is extended by the addition of four nucleotide analogues, each of which has a unique fluorescent label and can act as a reversible terminator in the DNA polymerase reaction. Upon detection of the incorporated nucleotide analogue by its unique fluorescent label, both the label and the 3' protecting group are removed to reinitiate the next DNA polymerase reaction cycle.

Taking this into account, the following requirements should be met to establish an entire SBS system: *(1)* standard cloning techniques to amplify DNA must be replaced by a high-throughput method for DNA template preparation; *(2)* after initial amplification, DNA templates must be physically arrayed in a format that allows each template to be probed multiple times; *(3)* nucleotides must be reversible terminators (3'-OH is blocked) so that only a single nucleotide is added at each step during SBS; *(4)* the 3'-OH blocking group and the fluorescent label used in SBS must be easily removed after detection for subsequent nucleotide addition; *(5)* the entire system must allow for simple washing and reagent additions between detection cycles. For SBS based on single fluorescent molecule detection, there is no need for the template amplification step. Emulsion PCR, which has been shown to have the potential to address DNA template preparation for various sequencing platforms,[4, 8] can be readily adapted to the approach shown in Fig. 2.1 for SBS. A recently developed SBS system based on a similar design of the CF-NRTs has already found wide applications in genome biology.[9-12]

2.3 Four color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators

2.3.1. Overview

In order to design the functional reporter nucleotides used in the SBS extension reaction, it is important to examine the structure of the polymerase enzyme complex with a DNA template, a primer and an incoming nucleotide during polymerase reaction. The 3-D structure of the ternary complexes of a rat DNA polymerase, a DNA template-primer, and a dideoxycytidine triphosphate (ddCTP) is shown in Fig. 2.3.[13]

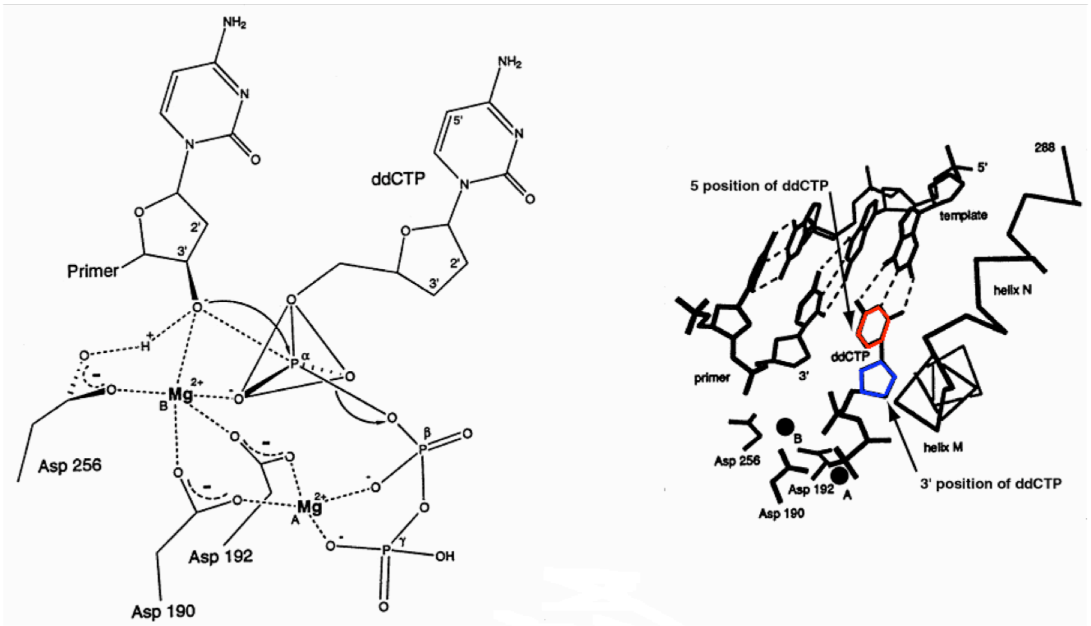


Fig. 2.3. The 3-D structure of the ternary complexes of a rat DNA polymerase, a DNA template-primer, and dideoxycytidine triphosphate (ddCTP). The left side of the illustration shows the mechanism for the addition of ddCTP and the right side shows the active site of the polymerase. It is important to notice that the 3' position of the dideoxyribose ring is very crowded, while ample space is available at the 5 position of the cytidine base.

It is apparent from this structure that the 5-position of the cytosine points away from the catalytic pocket of the enzyme, while the 3'-position of the ribose ring in ddCTP is in a very crowded space near the active amino acid residues of the polymerase. Any group that is attached at the 3' position of the sugar must be small so as to not interfere with the polymerase reaction. Large bulky dye molecules have been attached at the 5 position of pyrimidines and the 7 position of purines and used in enzymatic incorporation reactions, especially in Sanger dideoxy-sequencing.[14-16] Thus if a unique fluorescent dye is attached to the 5-position of the pyrimidines (T and C) and 7-position of purines (G and A) through a cleavable linker, and a small chemical moiety is used to cap the 3'-OH group, the resulting nucleotide analogues should be able to incorporate into the growing DNA strand. Based on this rationale, Ju *et al.*

proposed a SBS methodology (Fig. 2.1) using cleavable fluorescent nucleotide analogues as reversible terminators to sequence surface-immobilized DNA.[7, 17] In this approach, the nucleotides are modified at two specific locations so that they are still recognized by DNA polymerase as substrates (Fig. 2.4): (i) a different fluorophore with a distinct fluorescent emission is linked to each of the four bases through a cleavable linker and (ii) the 3'-OH group is capped by a small chemically reversible moiety. DNA polymerase incorporates only a single nucleotide analogue complementary to the base on a DNA template covalently linked to a surface. After incorporation, the unique fluorescence emission is detected to identify the incorporated nucleotide. The fluorophore is subsequently removed and the 3'-OH group is regenerated, allowing the next cycle of the polymerase reaction to proceed. Because the large surface on a DNA chip can have a high density of different DNA templates spotted, each cycle can identify many bases in parallel, allowing the simultaneous sequencing of a large number of DNA molecules.

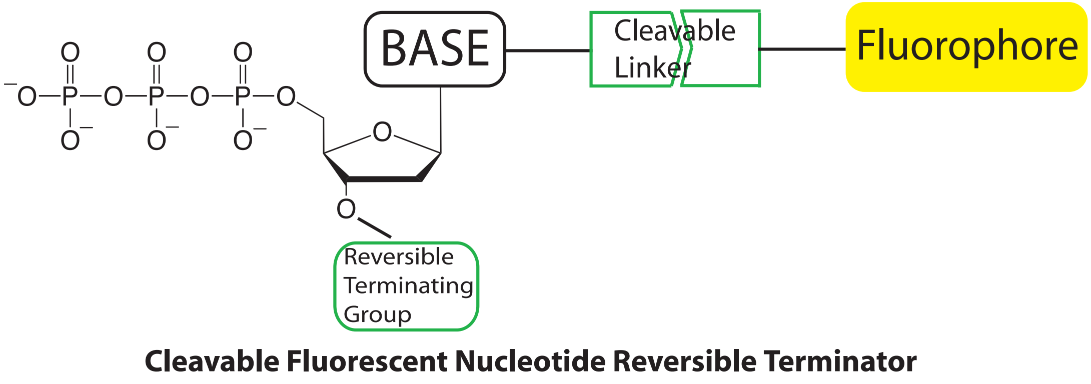
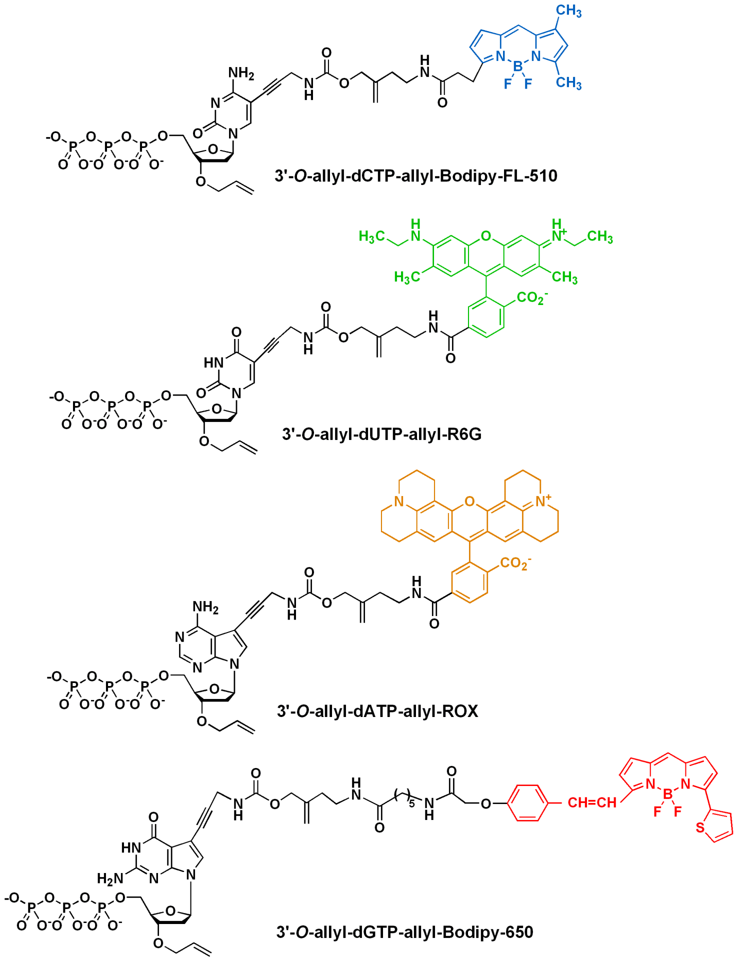

BASE
Cleavable Linker
Fluorophore
Reversible Terminating Group
Cleavable Fluorescent Nucleotide Reversible Terminator

Fig. 2.4. General structure of CF-NRT: the dual modified nucleotide analogue has a small reversible terminating moiety capping the 3' position and a cleavable linker tethering a unique fluorophore to the base of the nucleotide analgue.

2.3.2. Design, synthesis, and characterization of cleavable fluorescent nucleotide reversible terminators

Through previous research that established the feasibility of performing SBS on a chip using four photocleavable fluorescent nucleotide analogues[18] and successful use of an allyl group as a cleavable linker to bridge a fluorophore to a nucleotide,[19-21] Ju *et al* reported the design and synthesis of nucleotide analogues containing a 3'-O-allyl group and a unique fluorophore tethered by a cleavable allyl linker for SBS.[17] The four chemically cleavable fluorescence nucleotide reversible terminators (3'-O-allyl-dCTP-allyl-Bodipy-FL-510, 3'-O-allyl-dUTP-allyl-R6G, 3'-O-allyl-dATP-allyl-ROX and 3'-O-allyl-dGTP-allyl-Bodipy-650/Cy5) (Fig. 2.5) were designed according to the general rationale for nucleotide modification described in the previous section. Since modified DNA polymerases have been shown to be highly tolerant to nucleotide modifications with bulky groups at the 5-position of pyrimidines (C and U) and the 7-position of purines (A and G), each unique fluorophore was attached to the 5-position of C/U and the 7-position of A/G through an allyl carbamate linker. However, due to the close proximity of the 3' position on the sugar ring of a nucleotide to the amino acid residues of the active site of the DNA polymerase, a relatively small allyl moiety was chosen as the 3'-OH reversible capping group. After the incorporation of these nucleotide analogues and the detection of the fluorescent signal, the fluorophore and the 3'-O-allyl group on the DNA extension product are removed simultaneously in 30 seconds by Pd-catalyzed deallylation in aqueous solution. Such an efficient one-step dual-deallylation reaction allows the re-initiation of the polymerase reaction to incorporate the next base.



3'-*O*-allyl-dCTP-allyl-Bodipy-FL-510

3'-*O*-allyl-dUTP-allyl-R6G

3'-*O*-allyl-dATP-allyl-ROX

3'-*O*-allyl-dGTP-allyl-Bodipy-650

Fig. 2.5. Structures of allyl modified CF-NRTs (3'-*O*-ally-dNTP-allyl-fluorophores), with the 4 fluorophores having distinct fluorescent emissions: 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510 [$\lambda_{abs(max)}$ = 502 nm; $\lambda_{em(max)}$ = 510 nm], 3'-*O*-allyl-dUTP-allyl-R6G [$\lambda_{abs(max)}$ = 525 nm; $\lambda_{em(max)}$ = 550 nm], 3'-*O*-allyl-dATP-allyl-ROX [$\lambda_{abs(max)}$ = 585 nm; $\lambda_{em(max)}$ = 602 nm], and 3'-*O*-allyl-dGTP-allyl-bodipy-650 [$\lambda_{abs(max)}$ = 649 nm; $\lambda_{em(max)}$ = 670 nm].

To verify that these cleavable fluorescent nucleotide reversible terminators are incorporated accurately in a base-specific manner in a polymerase reaction, four continuous steps of DNA extension and deallylation were carried out in solution. This allows the isolation of the DNA product at each step for detailed molecular structure characterization by MALDI-TOF mass spectrometry (MS) as shown in Fig. 2.6. These results demonstrate that the four dual-allyl modified chemically cleavable fluorescent nucleotide analogues are successfully incorporated with high fidelity into the growing DNA strand in a polymerase reaction, and furthermore, both the fluorophore and the 3’-O-allyl group are efficiently removed by using a Pd-catalyzed deallylation reaction, which makes it feasible to use them for SBS on a chip.

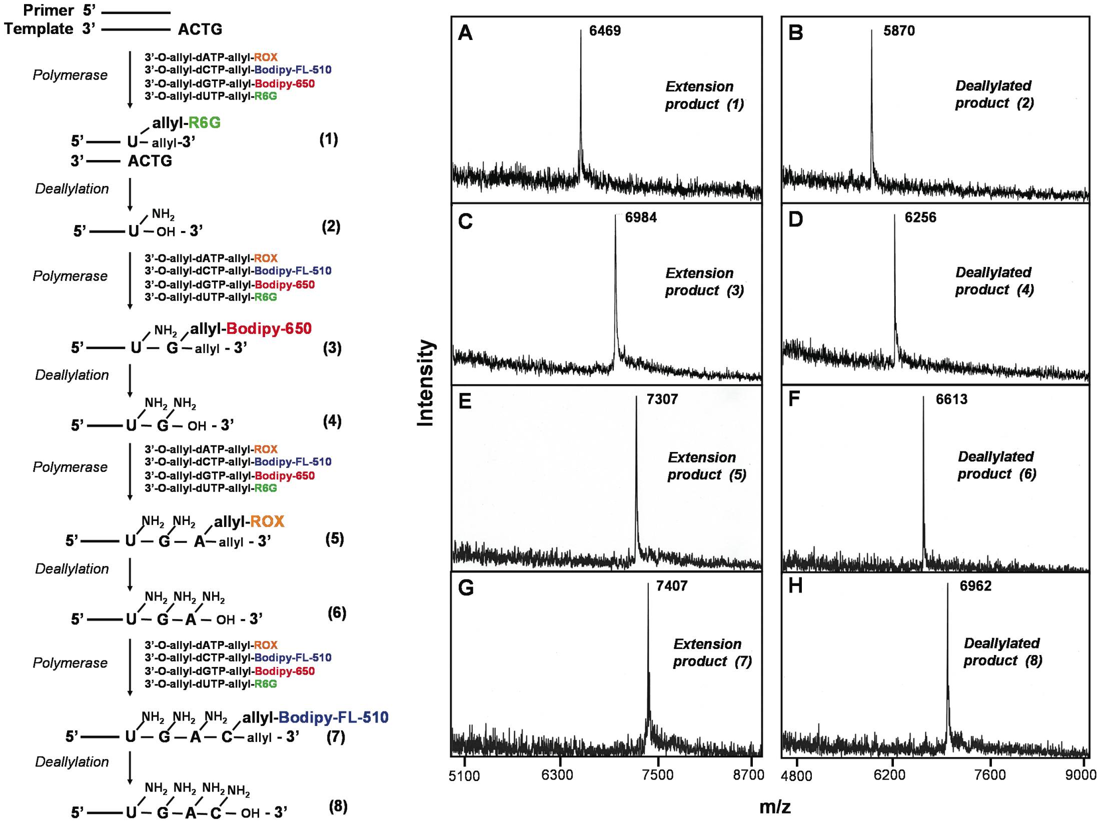


Fig. 2.6. The polymerase extension scheme (*Left*) and MALDI-TOF MS spectra of the four consecutive extension products and their deallylated products (*Right*). Primer extended with 3'-*O*-allyl-dUTP-allyl-R6G (1), and its deallylated product (2); Product (2) extended with 3'-*O*-allyl-dGTP-allyl-bodipy-650 (3), and its deallylated product (4); Product (4) extended with 3'-*O*-allyl-dATP-allyl-ROX (5), and its deallylated product (6); Product (6) extended with 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510 (7), and its deallylated product (8). After 30s of incubation with the palladium/TPPTS mixture at 70°C, deallylation is complete with both the fluorophores and the 3'-*O*-allyl groups cleaved from the extended DNA products.

2.3.3. DNA chip construction

In order to construct a DNA chip for SBS, a site-specific 1,3-dipolar cycloaddition coupling chemistry was used to covalently immobilize the alkyne-labeled self-priming DNA template on the azido-functionalized surface in the presence of a Cu(I) catalyst, as shown in Fig. 2.7.[17] The principal advantage offered by the use of a self-priming moiety as compared with using separate primers and templates is that the covalent linkage of the primer to the template in the self-priming moiety prevents any possible dissociation of the primer from the template during the process of SBS. To prevent nonspecific absorption of the unincorporated fluorescent nucleotides on the surface of the chip, a PEG linker is introduced between the DNA templates and the chip surface.[17] This approach was shown to produce very low background fluorescence after cleavage to remove the fluorophore as demonstrated by the DNA sequencing data described below.

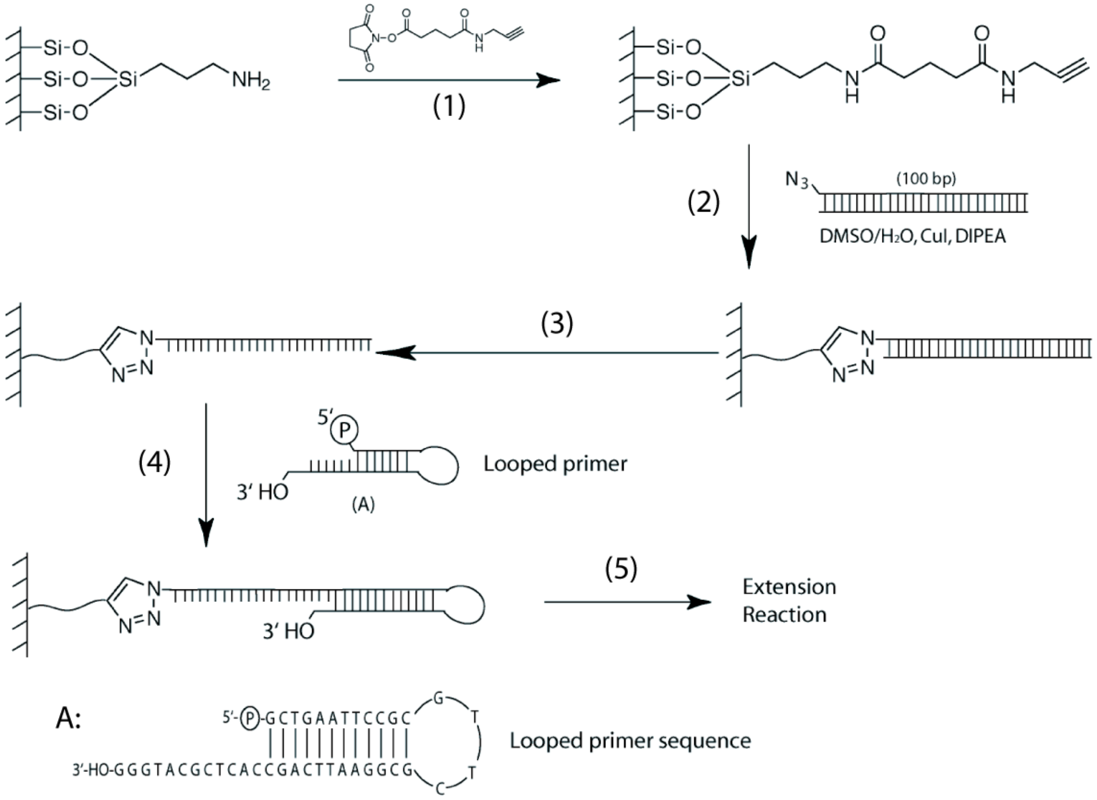


Fig 2.7. DNA immobilization on a surface with click chemistry. An amino modified glass surface is reacted with a linker containing an NHS-ester on one end and an alkyne on the other to functionalize the surface (step 1). Azide labeled PCR product is then attached to the surface using click chemistry (step 2). The unattached DNA strand is removed under alkaline conditions (step 3). A loop DNA primer is ligated to the single-stranded DNA template (step 4). DNA extension reactions are carried out to identify the sequence of the template (step 5). The sequence of the loop primer is shown in (A).

2.3.4. Four color sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators

SBS on a chip-immobilized DNA template that had no homopolymer sequences was carried out using the four chemically cleavable fluorescent nucleotide reversible terminators (3’-O-allyl-dCTP-allyl-bodipy-FL-510, 3’-O-allyl-dUTP-allyl-R6G, 3’-O-allyl-dATP-allyl-ROX and 3’-O-allyl-dGTP-allyl-Cy5), and the results are shown in Fig. 2.8. The *de novo* sequencing reaction on the chip was initiated by extending the self-priming DNA using a solution containing all four 3’-O-allyl-dNTP-allyl-fluorophores, and the 9°N mutant DNA polymerase. That primer was extended by only the complementary fluorescent nucleotide was confirmed by observing a red signal (the emission from Cy5) in a four-color fluorescent scanner (Fig. 2.8B1). After detection of the fluorescent signal, the chip surface was immersed in a deallylation mixture [1X Thermopol I reaction buffer/Na_2PdCl_4/$P(PhSO_3Na)_3$] to cleave both the fluorophore and 3’-O-allyl group simultaneously. The chip was then immediately immersed in a 3M Tris HCl buffer (pH 8.5) to remove the Pd complex. A negligible residual fluorescent signal was detected to confirm cleavage of the fluorophore. The entire process of incorporation, synchronization, detection, and cleavage was performed multiple times to identify 13 successive bases in the DNA template. The same method was applied to sequence a DNA template with two separate homopolymeric regions as shown in Fig. 2.9. All 20 bases, including the individual base (A, G, C, T), the 10 repeated A’s, and the 5 repeated A’s were clearly identified. In contrast, the pyrosequencing data for the same DNA template (Fig. 2.9) displayed two large peaks, from which it was very difficult to obtain the exact sequence.

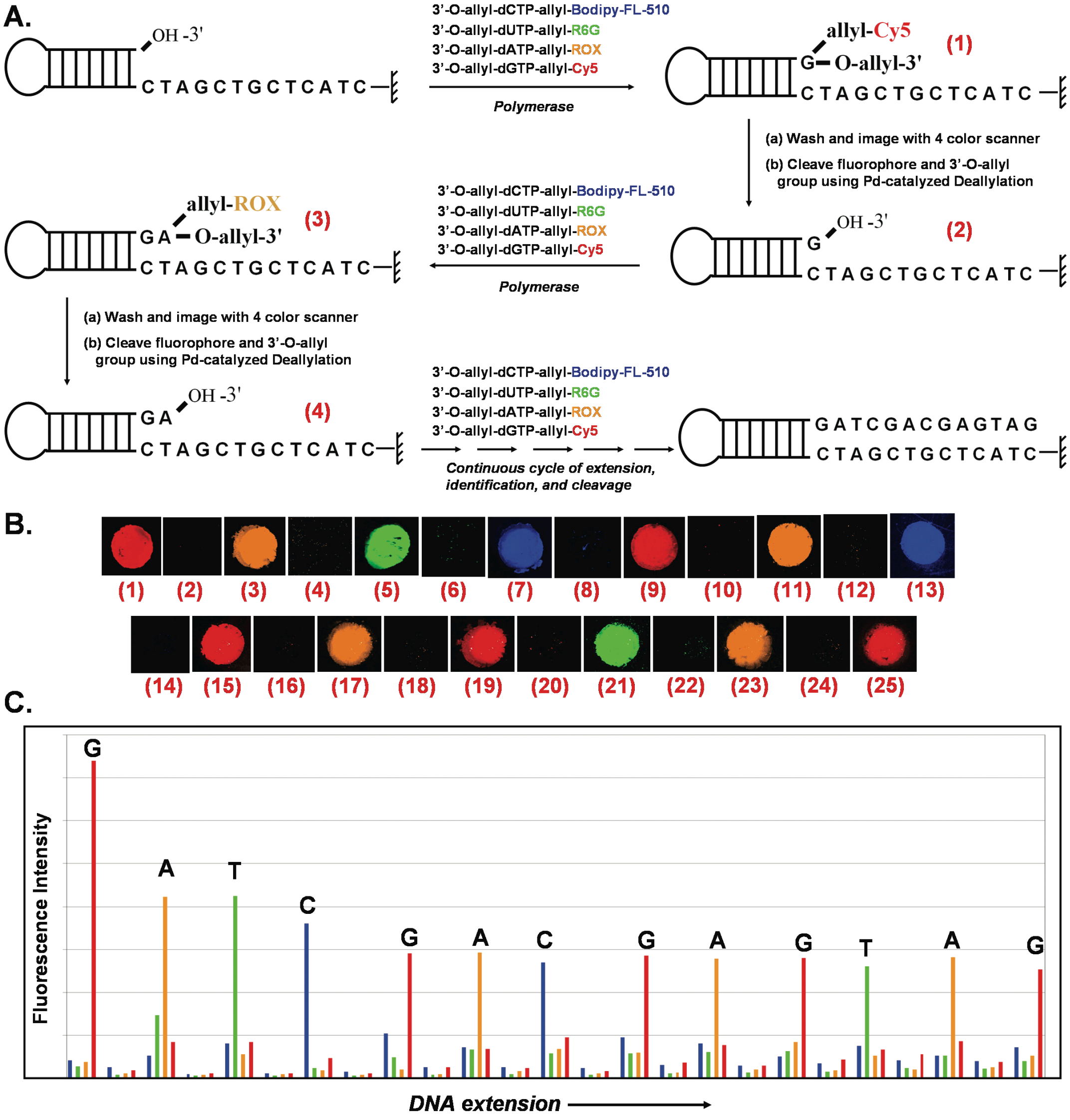

A.
OH -3'
CTAGCTGCTCATC
3'-O-allyl-dCTP-allyl-Bodipy-FL-510
3'-O-allyl-dUTP-allyl-R6G
3'-O-allyl-dATP-allyl-ROX
3'-O-allyl-dGTP-allyl-Cy5
Polymerase
allyl-Cy5
G
O-allyl-3'
(1)
CTAGCTGCTCATC
(a) Wash and image with 4 color scanner
(b) Cleave fluorophore and 3'-O-allyl group using Pd-catalyzed Deallylation
allyl-ROX
GA
O-allyl-3'
(3)
CTAGCTGCTCATC
3'-O-allyl-dCTP-allyl-Bodipy-FL-510
3'-O-allyl-dUTP-allyl-R6G
3'-O-allyl-dATP-allyl-ROX
3'-O-allyl-dGTP-allyl-Cy5
Polymerase
G
OH -3'
(2)
CTAGCTGCTCATC
(a) Wash and image with 4 color scanner
(b) Cleave fluorophore and 3'-O-allyl group using Pd-catalyzed Deallylation
GA
OH -3'
(4)
CTAGCTGCTCATC
3'-O-allyl-dCTP-allyl-Bodipy-FL-510
3'-O-allyl-dUTP-allyl-R6G
3'-O-allyl-dATP-allyl-ROX
3'-O-allyl-dGTP-allyl-Cy5
Continuous cycle of extension, identification, and cleavage
GATCGACGAGTAG
CTAGCTGCTCATC
B.
(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13)
(14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25)
C.
Fluorescence Intensity
G A T C G A C G A G T A G
DNA extension

Fig. 2.8. Four-color sequencing by synthesis data on a DNA chip. (*A*) Reaction scheme of SBS on a chip using four chemically cleavable fluorescent nucleotides. (*B*) The scanned four-color fluorescence images for each step of SBS on a chip: (1) incorporation of 3'-*O*-allyl-dGTP-allyl-Cy5; (2) cleavage of allyl-Cy5 and 3'-allyl group; (3) incorporation of 3'-*O*-allyl-dATP-allyl-ROX; (4) cleavage of allyl-ROX and 3'-allyl group; (5) incorporation of 3'-*O*-allyl-dUTP-allyl-R6G; (6) cleavage of allyl-R6G and 3'-allyl group; (7) incorporation of 3'-*O*-allyl-dCTP-allyl-bodipy-FL-510; (8) cleavage of allyl-bodipy-FL-510 and 3'-allyl group; images 9–25 are similarly produced. (*C*) A plot (four-color sequencing data) of raw fluorescence emission intensity at the four designated emission wavelengths of the four chemically cleavable fluorescent nucleotides vs. the progress of sequencing extension.

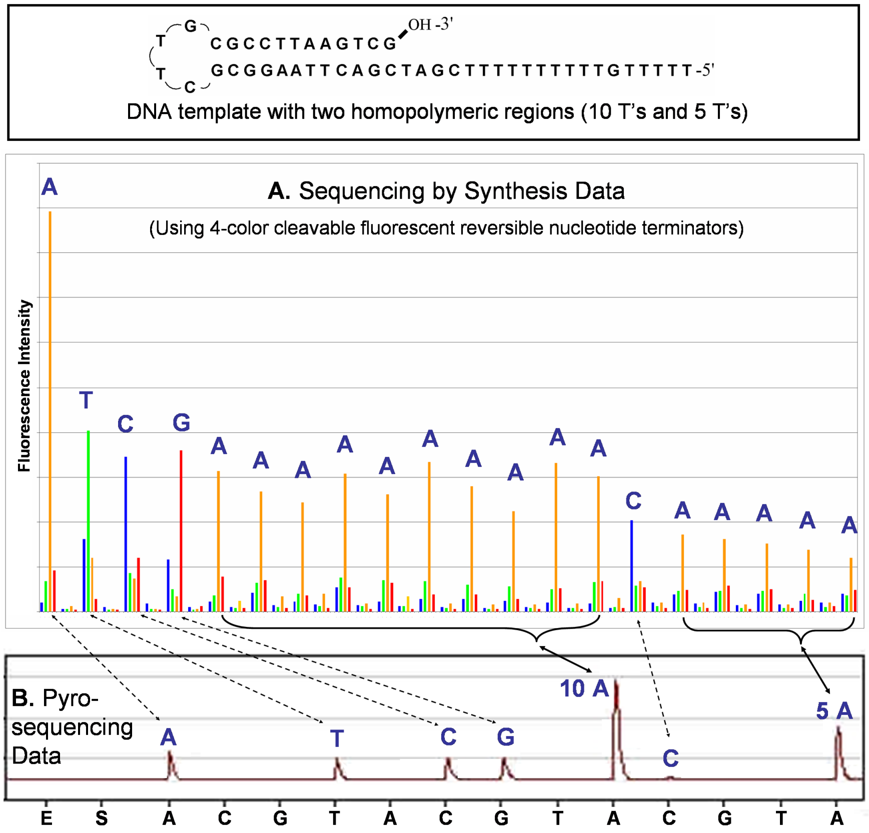


Fig. 2.9. Comparison of four-color sequencing by synthesis and pyrosequencing data. (*A*) Four-color DNA sequencing raw data with our sequencing by synthesis chemistry using a template containing two homopolymeric regions. The individual bases (A, T, C, G), the 10 repeated A's, and the five repeated A's are clearly identified. The small groups of peaks between the identified bases are fluorescent background from the DNA chip, which does not build up as the cycle continues. (*B*) The pyrosequencing data of the same DNA template containing the homopolymeric regions (10 T's and five T's). The first four individual bases are clearly identified. The two homopolymeric regions (10 A's) and (five A's) produce two large peaks, from which it is very difficult to identify the exact sequence.

2.4 References

1. Schena, M.; Shalon D.; Davis, R.W.; Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270,* 467-470.

2. Hyman, E. D. A new method of sequencing DNA. *Anal Biochem* **1988**, *174,* 423-436.

3. Ronaghi, M.; Karamohamed S.; Pettersson, B.; Uhlen, M.; Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **1996**, *242,* 84-89.

4. Margulies, M.; Egholm, M.; Altman, W. E.; Atiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376-380.

5. Ronaghi, M.; Uhlen M.; Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **1998**, *281,* 363, 365.

6. Wu, J., *et al.* 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* **2007**, *104,* 16462-16467.

7. Ju, J.; Li Z.; Edwards, J.; Itagaki, Y. Massive parallel method for decoding DNA and RNA. United States Patent 6,664,079 USA, **2003**.

8. Shendure, J.; Porreca, G. J.; Reppas, N. B.; Lin, X.; McCutcheon, J. P.; Rosenbaum, A. M.; Wang, M. D.; Zhang, K.; Mitra, R. D.; Church, G. M. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* **2005**, *309,* 1728-1732.

9. Mikkelsen, T. S., *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **2007**, *448*, 553–560.

10. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502.

11. Barski, A., *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **2007**, *129*, 823–837.

12. Bentley, D. R.; *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53-59.

13. Pelletier, H.; Sawaya M. R.; Kumar, A.; Wilson, S. H.; Kraut, J. Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **1994**, *264,* 1891-1903.

14. Zhu, Z.; Chao J.; Waggoner, A. S. Directly labeled DNA probes using fluorescent nucleotides with different length linkers. *Nucleic Acids Res* **1994**, *22,* 3418-3422.

15. Rosenblum, B. B.; *et al.* New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res* **1997**, *25,* 4500-4504.

16. Duthie, R. S.; *et al.* Novel cyanine dye-labeled dideoxynucleoside triphosphates for DNA sequencing. *Bioconjug Chem* **2002**, *13,* 699-706.

17. Ju, J., *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **2006**, *103,* 19635-19640.

18. Seo, T. S.; Bai, X.; Kim, D. H.; Meng, Q.; Shi, S.; Ruparel, H.; Li, Z.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci U S A* **2005**, *102,* 5926-5931.

19. Bi, L.; Kim, D. H.; Ju, J. Design and synthesis of a chemically cleavable fluorescent nucleotide, 3'-O-allyl-dGTP-allyl-Bodipy-FL-510, as a reversible terminator for DNA sequencing by synthesis. *J. Am Chem Soc*. **2006**, *128,* 2542-2543.

20. Ruparel, H., *et al.* Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc Natl Acad Sci U S A* **2005**, *102,* 5932-5937.

21. Meng, Q., *et al.* Design and synthesis of a photocleavable fluorescent nucleotide 3'-O-allyl-dGTP-PC-bodipy-FL-510 as a reversible terminator for DNA sequencing by synthesis. *J Org Chem,* **2006**, *71,* 3248-3252.

Chapter 3: Exploration of a New Chemical Moiety for Nucleotide Reversible Terminator Modification in DNA Sequencing by Synthesis

3.1 Introduction

The successful implementation of sequencing by synthesis is essentially dependent on the modified nucleotides used during the sequencing reactions. As discussed in the previous chapter, the design of cleavable fluorescent nucleotide reversible terminators (CF-NRTs) must encompass a suitable chemical moiety for capping the 3’-OH of the nucleotide such that it temporarily terminates the polymerase reaction, enabling the unambiguous identification of the incorporated base. The 3’-OH capping group of the nucleotides with a reversible moiety allows for the simultaneous addition of all four nucleotides, eliminating the necessity of adding each nucleotide sequentially as is done in some next-gen sequencing technologies.[1] This results in increased accuracy and reduction of the number of cycles needed during SBS. Our previous research efforts have firmly established the molecular level strategy to rationally modify the nucleotides by capping the 3’-OH with a small chemically reversible moiety for SBS.[2-6] Although DNA templates were accurately sequenced using CF-NRTs modified with the allyl capping group and an allyl-based cleavable linker, the use of palladium in cleavage solution could damage the templates, thus limiting sequencing read length. We decided to explore other chemical groups that could potentially be used as linkers and capping groups for nucleotide modification so that linker cleavage and 3’ capping group removal conditions could be milder and more DNA-compatible. Such improved conditions would ultimately lead to a more efficient sequencing by synthesis process, drastically increasing the read-length of the target templates. Building on our successful nucleotide modification strategy, we have explored alternative chemically reversible groups for capping the 3'-OH and linking the fluorophore to the nucleotides.

3.2 Experimental Rationale and Overview

Among the criteria for choosing the chemical moiety to modify CF-NRTs, size and reversibility are of foremost importance. According to the 3-D structure study of the polymerase complex with a DNA template, a primer and an incoming nucleotide during the polymerase reaction,[7] the 3'-position of the ribose ring in a nucleotide occupies a very crowded space near the active amino acid residues of the polymerase. Hence any group that is attached at the 3' position of the sugar must be small enough so that it does not pose steric hindrance to the polymerase activity. The chemical moiety capping the 3'-position of the nucleotide also needs to be reversible, meaning it can be cleaved to restore the 3'-OH group on the nucleotide so the next base can be incorporated. Taking these two considerations into account, we focused on the azido functional group (Fig. 3.1) as a potential chemical moiety for nucleotide modification due to its small size and ability to be reduced under mild conditions.

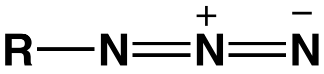


Fig. 3.1. Structure of the azido functional group

As an exemplar of the Staudinger reaction, an azido group can be effectively converted into an amine with phosphine in DNA-friendly aqueous solution.[8] This efficient reduction is further enhanced through the utilization of Tris(2-carboxyethyl) phosphine (TCEP), an odorless and stable agent often used to digest peptide disulfide bonds (Fig. 3.2).

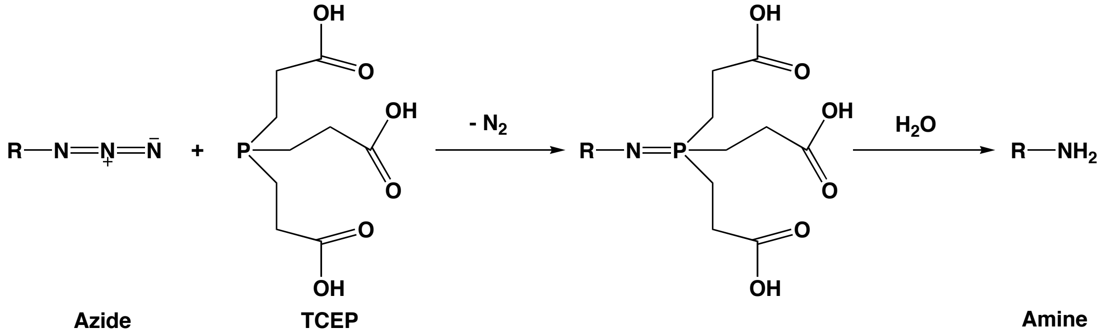


Fig. 3.2. Staudinger reduction of azido functional group using TCEP

In 1991, Zavgorodny et al. reported the capping of the 3'-OH group of the nucleoside with an azidomethyl moiety, which can be chemically cleaved under mild conditions with triphenylphosphine in aqueous solutions.[9] According to Zavogorodny, TCEP can reduce the azido-methyl capping group to methylamine at the 3' sugar base of the nucleotide. Since the carbon of the methylamine is highly unstable due to its position between two electron-withdrawing elements (oxygen and nitrogen), the methylamine is hydrolyzed in the presence of water to recover the hydroxyl group at the 3' position (Fig. 3.3).

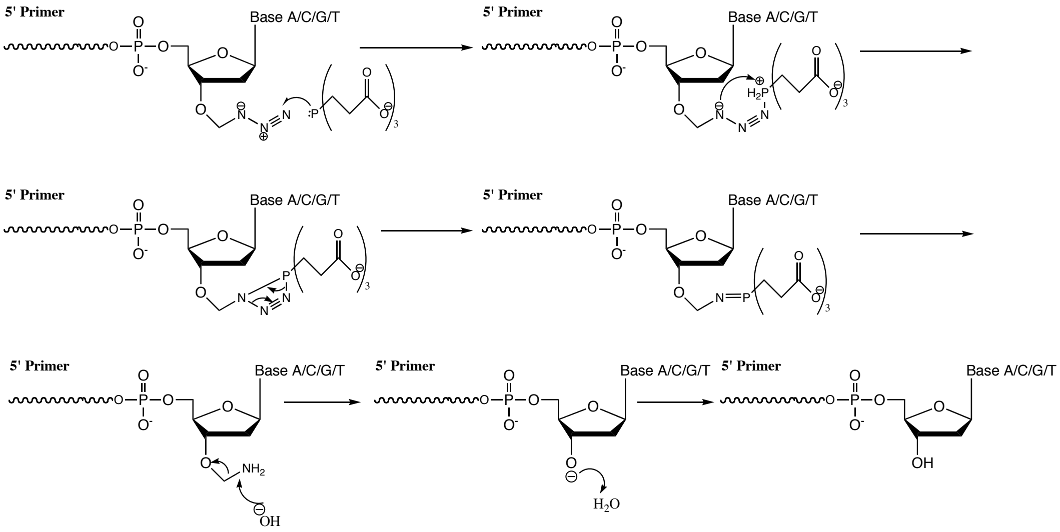


Fig. 3.3. Staudinger reduction of 3’-O-N_3 group of an incorporated NRT to recover its 3’-OH group

In order to verify the feasibility of using an azido group to modify nucleotide reversible terminators (NRTs) and evaluate the potential implementation of these NRTs in sequencing by synthesis, we synthesized a model NRT, 3’-O-azidomethyl-dUTP (Fig. 3.4). The incorporation and reduction conditions of the NRT were tested and optimized in preparation for its utilization in SBS.

3.3 Results and Discussion

The key requirements for evaluating nucleotide reversible terminators (NRTs) to be used in SBS are their ability to terminate the DNA polymerase reaction immediately after their incorporation and the efficient removal of their terminating group to allow the subsequent base incorporation.

3.3.1. Design and synthesis of 3’-O-azidomethyl-dUTP-NH_2, a model NRT compound

The model NRT, 3’-O-azidomethyl-dUTP-NH_2 (Fig. 3.4), capped at the 3’ position with an azidomethyl moiety, was synthesized and used to evaluate the feasibility of its use in SBS. The 3'-azidomethyl capping group served as a reversible terminating moiety that was expected to temporarily stop the polymerase reaction. After removal of the azidomethyl group with TCEP, the 3'-OH would be restored to continue the incorporation of the next base. An amino-propargyl group was attached to the 5-position of the base in preparation for future cleavable linker attachment.

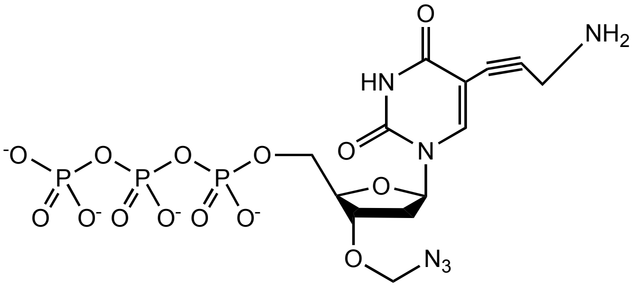


Fig. 3.4. Structure of 3'-O-azidomethyl-dUTP

3.3.2. Polymerase reaction using 3'-O-azidomethyl-dUTP-NH_2 and characterization by MALDI-TOF MS

In order to establish if the 3'-azidomethyl modified NRT could be used in SBS, it was crucial to first verify that it would be recognized by DNA polymerase as a substrate during the polymerase reaction. Hence, we performed a single base extension reaction using 3'-O-azidomethyl-dUTP-NH_2 along with a self-priming loop template (m.w.=7,966), manganese (Mn^{2+}, 20 mM), and 9°N Thermopolymerase at 65°C for 25 minutes (Fig. 3.5). The extension products were analyzed by MALDI-TOF mass spectrometry. If the modified nucleotide reversible terminator was successfully incorporated, the template would have been extended by one base, increasing its mass from 7,966 to 8,363. As shown in the resulting MALDI-TOF MS spectrum (Fig. 3.5), the initial peak of the primer at m/z=7,966 completely disappeared, replaced by a single peak at m/z=8,363 that corresponded to the correct extension product. Thus, we verified that the model NRT was a suitable substrate for DNA polymerase.

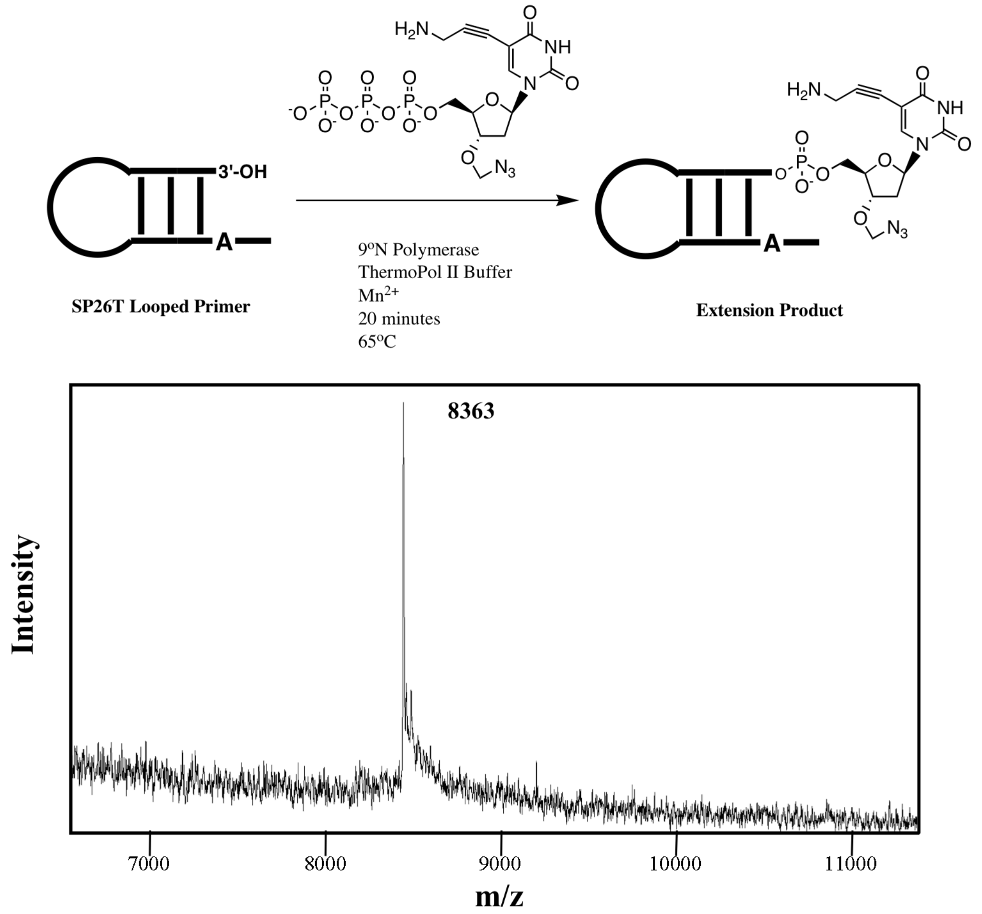


Fig. 3.5. Single base extension reaction scheme (top) and resulting MALDI-TOF MS spectrum (bottom) of 3’-O-azidomethyl-dUTP-NH$_2$.

We then investigated the efficiency of the polymerase reaction with the model NRT. A set of single base extension reactions was carried out with varying reaction times to determine the minimum time for complete incorporation. MALDI-TOF MS was again used for analysis of the extended primer products. As shown in Fig. 3.6, four single base extension reactions, with reaction times of 20min (A), 10min (B), 5min (C), and 2min (D), all yielded the same extension product at m/z=8,363, confirming the efficiency of incorporation of our model NRT, 3’-O-azidomethyl-dUTP-NH_2.

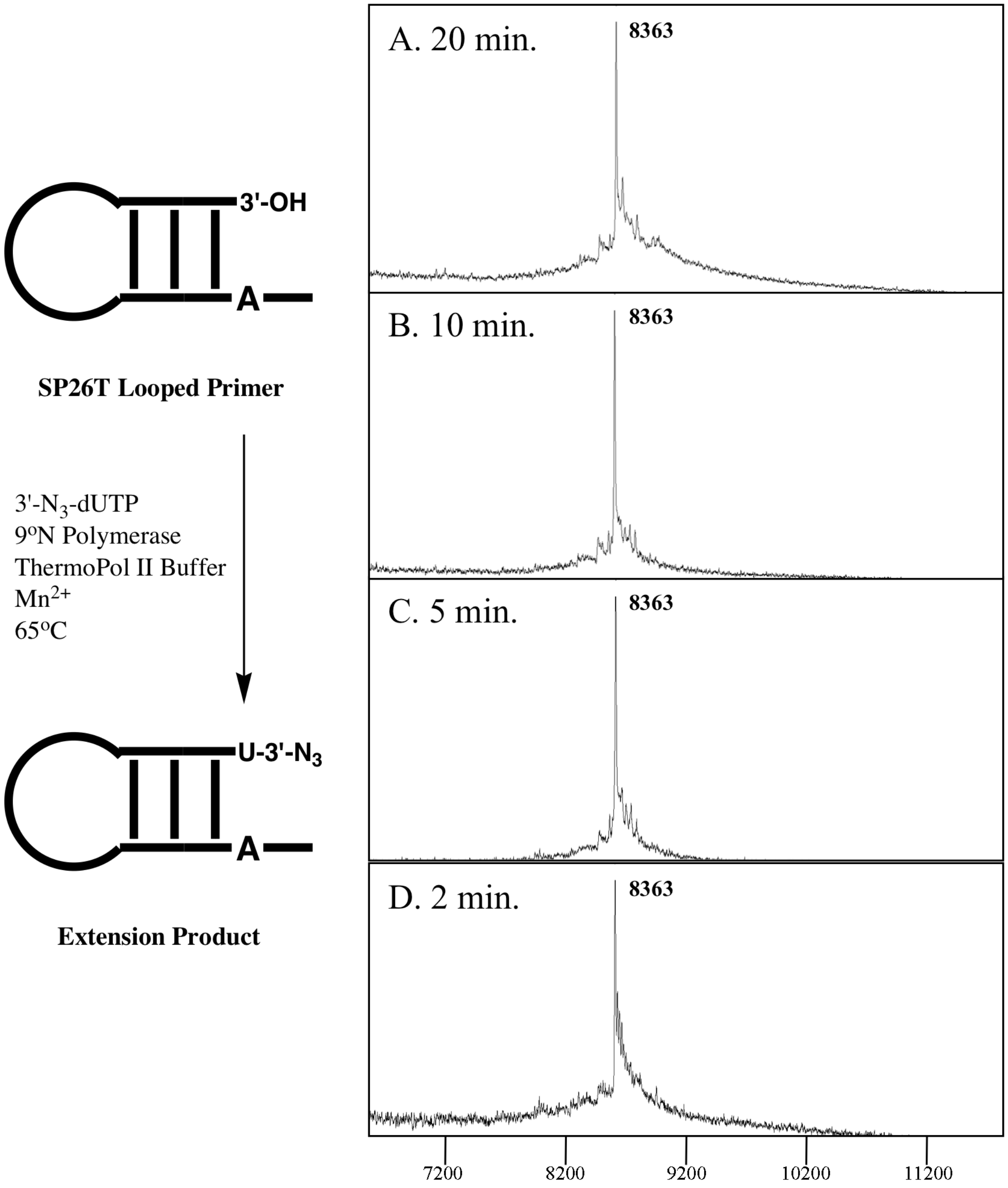


Fig. 3.6. Single base extension reaction scheme (left) and MALDI-TOF MS spectra of 3’-O-azidomethyl-dUTP with varying reaction times of A) 20min, B) 10min, C) 5min, and D) 2min.

3.3.3. Cleavage reaction to restore 3’-OH of DNA extension product and its optimization

Once we confirmed that the NRT, 3’-O-azidomethyl-dUTP-NH_2, was a good substrate for DNA polymerase and could be incorporated very efficiently to extend the primer by a single base, we went on to carry out the cleavage reaction on the extended primer to remove the 3’-azidomethyl group so that the 3’-OH could be restored for the next base incorporation. TCEP was used for the deprotection/cleavage reaction. Two sets of cleavage reactions were carried out, one varying reaction time and the other with different TCEP concentrations and the products were analyzed using MALDI-TOF MS (Fig. 3.7). The removal of the azidomethyl group at the 3’ position of the NRT would shift the mass of the extension product from 8,636 to 8,308 (replacing azidomethyl with hydrogen = 55 Dalton mass difference). Fig. 3.7 demonstrates the results of the cleavage reactions. The MALDI-TOF MS spectra on the left reveal the successful cleavage of the azidomethyl group using TCEP (25mM, pH=7.5) at 65°C for 30min (A), 15min (B), and 5min (C). In each case, the extension product peak at m/z=8,363 was replaced by the cleaved extension product peak at m/z=8,308. Similar MALDI-TOF MS spectra are shown on the right side, for cleavage reactions performed using TCEP (pH=7.5, 30min) at 65°C with concentrations of 25mM (D), 10mM (E), and 5mM (F). Thus it was concluded that even under the least stringent cleavage conditions, TCEP was an effective deprotection agent to reversibly convert the 3’-azidomethyl group back to 3’-OH on the NRT.

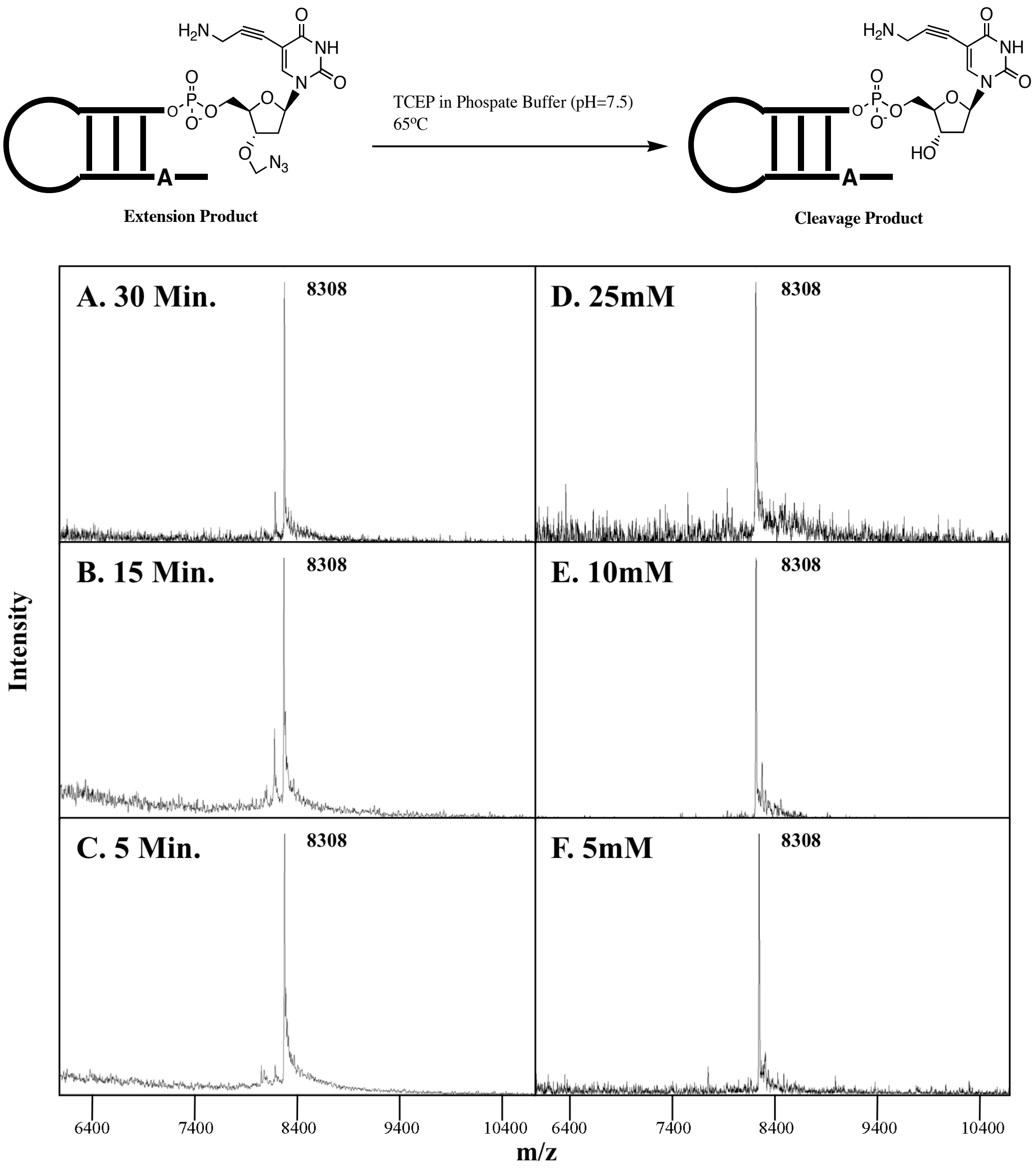
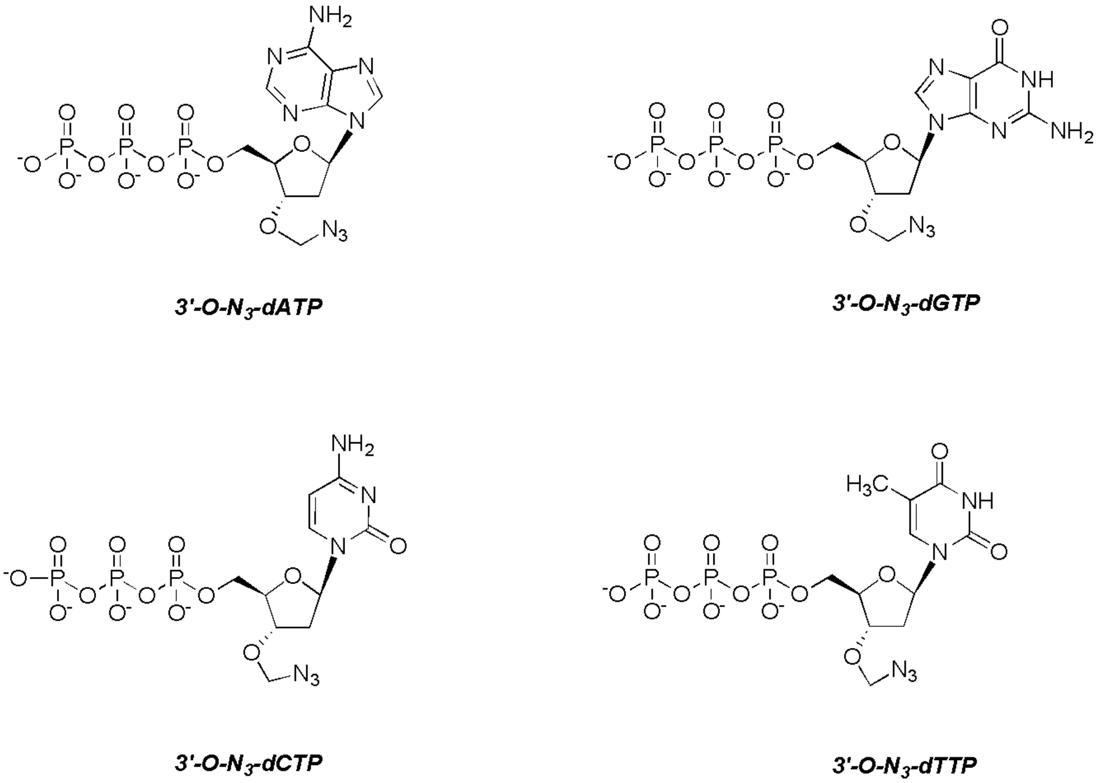


Fig. 3.7. MALDI-TOF MS spectra of cleavage products after treating the single base extension product using TCEP for varying times: A) 30min, B) 15min, C) 5min, and varying TCEP concentration: D) 25mM, E) 10mM, F) 5mM. In each case, the 3'-OH is recovered completely, showing a m/z of 8,308.

3.3.4. Design, synthesis, and evaluation of a complete nucleotide reversible terminator set: 3’-O-N_3-dNTPs

Through our previous experiments, we firmly established that the azidomethyl group (N_3) was a good reversible terminating moiety to modify NRTs. The model 3’-azidomethyl-dUTP-NH_2 was efficiently incorporated into a primer by the polymerase reaction, and easily cleaved to restore the 3’-OH group of the extension product for the next base incorporation. Hence, we designed and synthesized a complete set of nucleotide reversible terminators with 3’-azidomethyl modification (3’-O-N_3-dNTPs, Fig. 3.8) based on a method similar to that reported by Zavgorodny *et al.*[9, 10]



3'-O-N_3-dATP

3'-O-N_3-dGTP

3'-O-N_3-dCTP

3'-O-N_3-dTTP

Fig. 3.8. Structures of the nucleotide reversible terminators (NRTs): 3’-*O*-N_3-dATP, 3’-*O*-N_3-dCTP, 3’-*O*-N_3-dGTP, 3’-*O*-N_3-dTTP.

The 3’-O-azidomethyl group on the DNA extension product generated by incorporating each of the NRTs was able to terminate further elongation of the template. Upon efficient removal of the reversible terminating moiety by the Staudinger reaction using aqueous Tris(2-carboxyethyl) phosphine (TCEP) solution[11, 12] followed by hydrolysis to yield a free 3'-OH group, the DNA strand was able to carry out the next cycle of the SBS.

Each NRT (3'-O-N_3-dATP, mw=541; 3'-O-N_3-dGTP, mw=558; 3'-O-N_3-dCTP, mw=518; 3'-O-N_3-dTTP, mw=533) was incorporated into its corresponding looped primer in solution with manganese (Mn^{2+}, 20mM) and the mutant 9°N Thermopolymerase. Each extension reaction was carried out at 65°C for 5 minutes. Complete incorporation was confirmed with MALDI-TOF mass spectroscopy (MS) by observing the total disappearance of the primer peak (m/z=7966) and the emergence of extended product peaks (m/z=8332, 8348, 8323, and 8308 for 3'-O-N_3-dA/G/T/CTP, respectively, Fig. 3.9). After obtaining the extended product for each NRT, deprotection was carried out at 50°C with 5mM TCEP. The 3'-azidomethyl capping group could be removed completely in under 5 minutes. MALDI-TOF mass spectroscopy was again used to ascertain the results (Fig. 3.9).

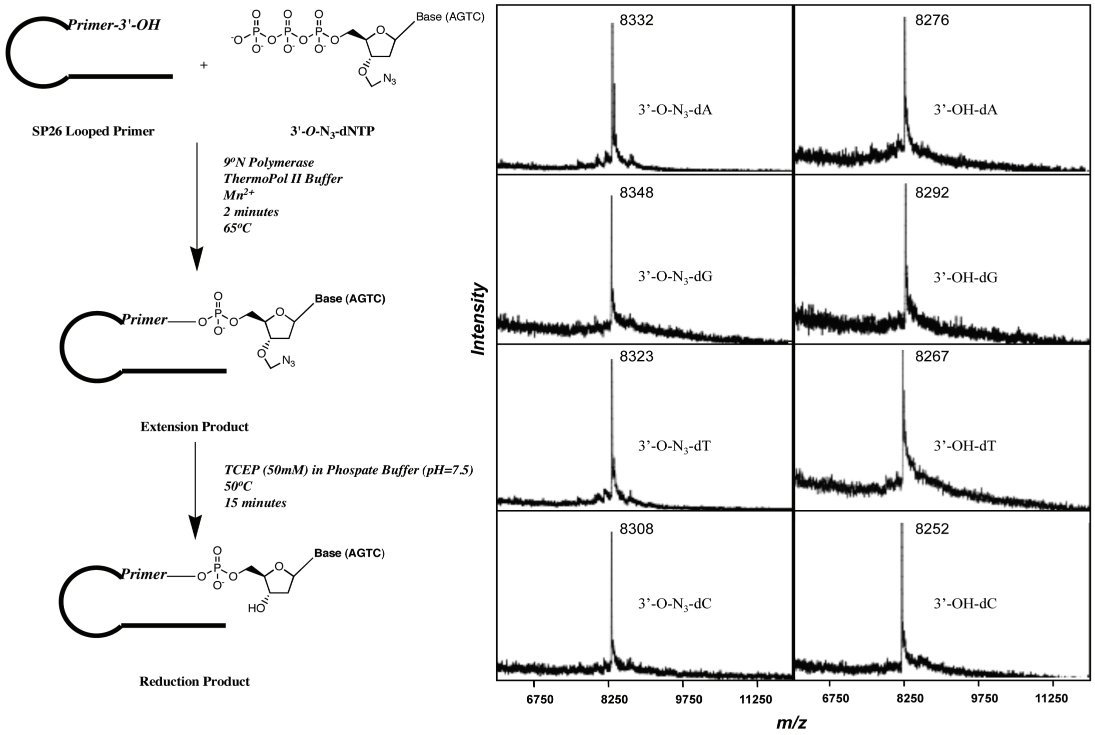


Fig. 3.9. Scheme (left) and MALDI-TOF MS spectra (right) of single base extension reaction and 3'-O-azidomethyl cleavage reaction for each modified nucleotide reversible terminator (3'-O-N_3-dATP, 3'-O-N_3-dGTP, 3'-O-N_3-dTTP, and 3'-O-N_3-dCTP)

3.3.5. Continuous polymerase extension using 3'-O-modified NRTs and characterization by MALDI-TOF mass spectrometry

In order to verify that 3'-O-modified NRTs (3'-O-N_3-dNTPs) are incorporated accurately in a base specific manner in the polymerase reaction, four continuous DNA extension and cleavage reactions were carried out in solution using 3'-O-N_3-dNTPs as substrates.[13] This allowed the isolation of the DNA product at each step for detailed molecular structure characterization as shown in Fig. 3.10. The first extension product 5'-primer-C-N_3-3' (**1**) was desalted and analyzed using MALDI-TOF MS (Fig. 3.10A). This product was then incubated in aqueous TCEP solution to remove the azidomethyl moiety to yield the cleavage product (**2**) with a free 3'-OH group, which was also analyzed using MALDI-TOF MS (Fig. 3.10B). As can be seen from Fig. 3.10A, the MALDI-TOF MS spectrum consists of a distinct peak corresponding to the DNA extension product 5'-primer-C-N_3-3' (**1**) (*m/z* 8,310), which confirms that the NRT is incorporated base specifically by DNA polymerase into a growing DNA strand. Fig. 3.10B shows the cleavage result on the DNA extension product. The extended DNA mass peak at *m/z* 8,310 completely disappeared while the peak corresponding to the cleavage product 5'-primer-C-3' (**2**) appears as the sole dominant peak at *m/z* 8,255, which establishes that TCEP incubation completely cleaves the 3'-O-azidomethyl group with high efficiency. The next extension reaction was carried out using this cleaved product, which now has a free 3'-OH group, as a primer to yield a second extension product, 5'-primer-CG-N_3-3' (**3**) (*m/z* 8,639, Fig. 3.10C). As described above, the extension product (**3**) was cleaved to generate product (**4**) for further MS analysis yielding a single peak at *m/z* 8,584 (Fig. 3.10D). The third extension reaction to yield 5'-primer-CGA-N_3-3' (**5**) (*m/z* 8,952, Fig. 3.10E), the fourth extension to yield 5'-primer-CGAT-N_3-3' (**7**) (*m/z* 9,256, Fig. 3.10G) and their cleavage to yield products (**6**) (*m/z* 8,897, Fig. 3.10F) and (**8**) (*m/z* 9,201, Fig. 3.10H) were similarly carried out and analyzed by MALDI-TOF MS. These results demonstrate that all four 3'-O-azidomethyl modified NRTs are successfully synthesized and efficiently incorporated base-specifically into the growing DNA strand in a continuous polymerase reaction as reversible terminators and the 3'-OH capping group on the DNA extension products is quantitatively cleaved by TCEP.

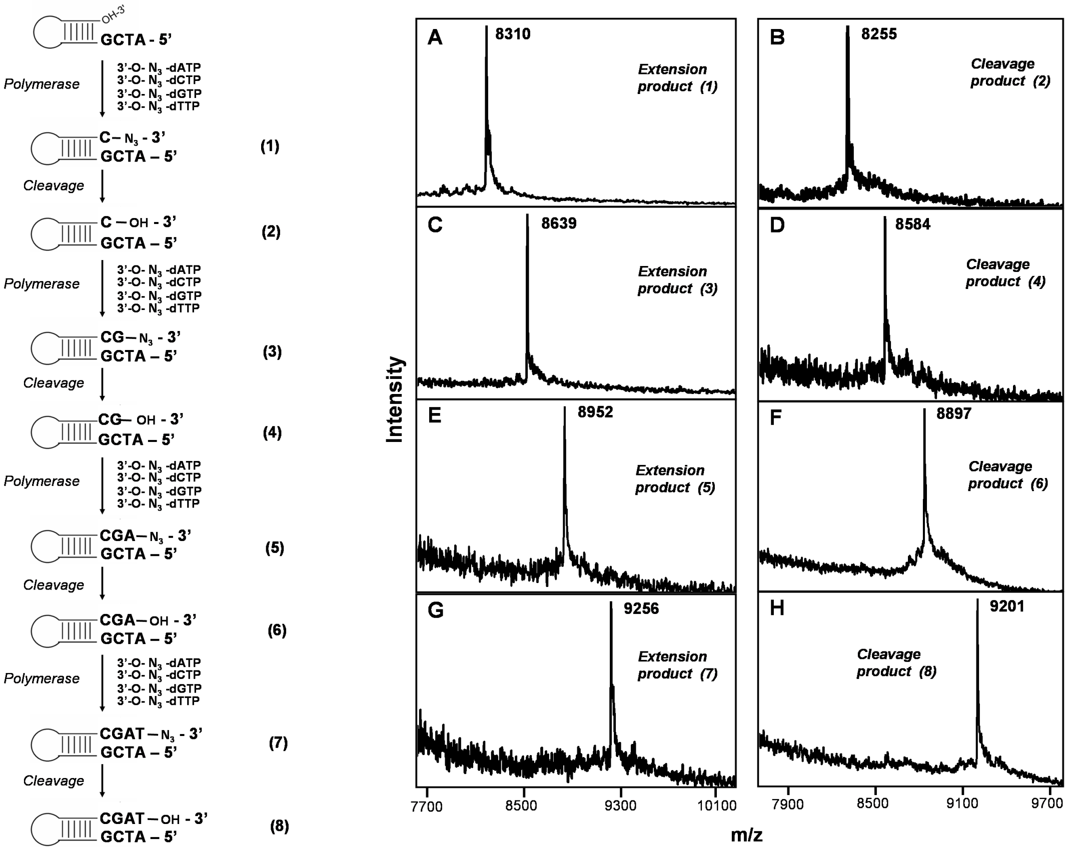


Fig. 3.10. The polymerase extension scheme (left) and MALDI-TOF MS spectra of the four consecutive extension products and their cleaveage products (right) using the nucleotide reversible terminators, 3'-O-N_3-dNTPs. Primer extended with 3'-O-N_3-dCTP (1) (right, A), and its cleavage product (2) (right, B); Product (2) extended with 3'-O-N_3-dGTP (3) (right, C), and its cleavage product (4) (right, D); Product (4) extended with 3'-O-N_3-dATP (5) (right, E), and its cleavage product (6) (right, F); Product (6) extended with 3'-O-N_3-dTTP (7) (right, G), and its cleavage product (8) (right, H). After brief incubation in a TCEP aqueous solution the azidomethyl moiety capping the 3'-OH group of the DNA extension products is completely removed to continue the polymerase reaction.

3.4 Materials and Methods

General Information. All solvents and reagents were reagent grade, purchased commercially and used without further purification. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Oligonucleotides used as primers or templates were synthesized on an Expedite nucleic acid synthesizer (Applied BioSystems) or purchased from Midland. Mass measurement of DNA was performed on a Voyager DE MALDI-TOF mass spectrometer (Applied Biosystems). 9^{o}N polymerase (exo-) A485L/Y409V was obtained from New England Biolabs. Phosphoramidite reagents and columns for oligonucleotide synthesis were purchased from Glen Research (Sterling, VA). The 3’-O-modified nucleotides were purified with reverse-phase HPLC on a 150×4.6 mm C18 column (Supelco), mobile phase: A, 8.6 mM Et_3N / 100 mM 1,1,1,3,3,3-hexafluoro-2-propanol in water (pH 8.1); B, methanol. Elution was performed from 100% A isocratic over 10 minutes followed by a linear gradient of 0-50% B for 20 minutes and then 50% B isocratic over another 30 minutes.

3.4.1. Synthesis of 3’-O-azidomethyl-dUTP-NH_2, a model NRT compound

The synthesis of the model azido modified nucleotide reversible terminator, 3’-O-azidomethyl-dUTP-NH_2, was accomplished by the organic chemistry synthesis team of our group. Drs. Mong Sang Marma and Zengmin Li spearheaded the synthesis efforts leading to the completion of the novel nucleotide analogue. The detailed synthesis procedures for 3’-O-azidomethyl-dUTP-NH_2 are described in our recent publications.[13]

3.4.2. Polymerase reaction using 3’-O-azidomethyl-dUTP and characterization by MALDI-TOF MS

We characterized the azido modified nucleotide reversible terminator, 3’-O-azidomethyl-dUTP-NH_2, by first performing a single base extension reaction using a self-priming DNA template (SP26T: 26-mer hairpin DNA with a 4-base 5’-overhang, 5’-GTCAGCGCCGCGCCTTGGCGCGGCGC-3’). The extension reaction (total volume of 20 μL) consisted of 3’-O-azidomethyl-dUTP (120 pmol) along with 60 pmol of the SP26T self-priming DNA template, 1X Thermopol II reaction buffer, 40 nmol of $MnCl_2$ and 1 unit of 9^oN DNA polymerase (exo-) A485L/Y409V. The reaction was performed at 65^oC for 25 minutes. Subsequently, the extension product was purified by ethanol precipitation and Zip-Tip desalting procedure, and then analyzed using MALDI-TOF MS. In order to investigate the efficiency of polymerase incorporation, the same extension reaction was repeated for various reaction times (20min, 10min, 5min, and 2min). The extended products were again analyzed using MALDI-TOF MS.

3.4.3. Cleavage reaction to restore 3'-OH of DNA extension product and its optimization

The cleavage of the DNA extension product incorporated with 3'-O-azidomethyl-dUTP-NH_2 was accomplished by resuspending the product in 40 μL of 25mM TCEP solution (pH 7.5) at 65^oC for 30 minutes. The cleaved products were then analyzed by MALDI-TOF MS. The same cleavage reaction was carried out at different times (15min and 5min) and TCEP concentrations (25mM, 10mM, and 5mM). Again, the cleaved products were analyzed using MALDI-TOF MS without further purification.

3.4.4. Design, synthesis, and evaluation of a complete nucleotide reversible terminator set: 3'-O-N_3-dNTPs

The synthesis of the complete set of azido modified nucleotide reversible terminators, 3'-O-N_3-dNTPs, was accomplished by the organic chemistry synthesis team in our group. Drs. Zengmin Li, Huanyan Cao, Shenglong Zhang, and Mong Sang Marma led the synthesis efforts for the successful completion of the novel nucleotide analogues. The detailed synthesis procedures for this whole set of NRTs are described in Dr. Shenglong Zhang's thesis as well as our recent publication.[13]

We characterized the four nucleotide reversible terminators, 3'-O-N_3-dNTPs (3'-O-N_3-dATP, 3'-O-N_3-dGTP, 3'-O-N_3-dCTP, and 3'-O-N_3-dUTP) by performing four separate single base DNA extension reactions, each with a different self-priming DNA template allowing the four NRTs to be incorporated. The resulting DNA extension products were analyzed by MALDI-TOF MS. The following four self-priming DNA templates (26-mer hairpin DNA with a 4-base 5'-overhang) were used for the extension: 5'-*GACT*GCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dATP; 5'-*GATC*GCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dGTP; 5'-*ATCG*GCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dCTP; and 5'-*GTCA*GCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dUTP. Each extension reaction (total volume of 20μL) consisted of one of the four 3'-O-N_3-dNTPs (120pmol) along with 60pmol of its corresponding self-priming DNA template, 1X Thermopol II reaction buffer, 40nmol $MnCl_2$ and 1 unit 9^oN DNA polymerase (exo-) A485L/Y409V. The reaction was performed at 65^oC for 5 minutes. Subsequently, the extension product was purified by reverse-phase HPLC using established procedures.[6] The fraction containing the desired product was collected and freeze-dried for analysis by MALDI-TOF MS. The cleavage of the DNA extension products incorporated with 3'-O-N_3-dNTP was accomplished by resuspending each product in 40 μL of 5mM TCEP solution (pH 7.5) at 50^oC for 5 minutes. The cleaved products were then analyzed with MALDI-TOF MS.

3.4.5. Continuous polymerase extension using 3'-O-modified NRTs and characterization by MALDI-TOF mass spectrometry

In order to further characterize the four NRTs (3'-O-N_3-dCTP, 3'-O-N_3-dTTP, 3'-O-N_3-dATP and 3'-O-N_3-dGTP) in a base specific manner during the polymerase reaction, we performed four continuous DNA extension reactions using a self-priming DNA template (5'-*ATCG*GCGCCGCGCCTTGGCGCGGCGC-3'). The four nucleotides in the template immediately adjacent to the annealing site of the primer are 3'-GCTA-5', which allows the evaluation of the incorporation and cleavage efficiency of the 4 NRTs. First, a polymerase extension reaction using a pool of all four NRTs along with the self-priming DNA template was performed producing a single base extension product. The reaction mixture for this, and all subsequent extension reactions, consisted of 80pmol of self-priming DNA template, 160pmol of each 3'-O-N_3-dNTP, 1X Thermopol II reaction buffer, 40nmol $MnCl_2$ and 1 unit 9^{o}N DNA polymerase (exo-) A485L/Y409V in a total reaction volume of 20 μL. The reaction was performed at 94^{o}C for 5 minutes, 4^{o}C for 5 minutes, and 65^{o}C for 20 minutes. Subsequently, the extension product was desalted by using a ZipTip and analyzed by MALDI-TOF mass spectrometry. For cleavage, the desalted DNA extension product bearing the 3'-O-azidomethyl group was first resuspended with 5 μL of 50 mM EDTA solution to quench the polymerase activity. This DNA solution was then mixed with 10μL of 225mM TCEP solution (pH 9.0) and incubated at 65^{o}C for 15 minutes to yield a cleaved DNA product which was characterized by MALDI-TOF MS. The DNA product with the 3'-O-azidomethyl group removed to generate a free 3'-OH group was purified using an Oligonucleotide Purification Cartridge (Applied Biosystems) and used as a primer for a second extension reaction using 3'-

O-N_3-dNTPs. The second extended DNA product was then purified by ZipTip and cleaved as described above. The third and the fourth extensions were carried out in a similar manner using the previously extended and cleaved products as primers.

3.5 Conclusion

We explored the potential of using an azido group as a chemical moiety for nucleotide modification. Based on our established rationale for nucleotide reversible terminator (NRT) design, we synthesized a complete set of NRTs capped at the 3' position with an azidomethyl group (3'-O-N_3-dATP, 3'-O-N_3-dCTP, 3'-O-N_3-dGTP, 3'-O-N_3-dTTP). Each nucleotide analog was successfully incorporated into its corresponding looped primer during a polymerase reaction at 65°C for various time spans ranging from 20 minutes to 2 minutes. Thus, it was apparent that these NRTs were good substrates for a DNA polymerase. After obtaining the extended product for each nucleotide, cleavage reactions were carried out under varying conditions such as time (1~20 minutes), temperature (20~60°C), and concentration of TCEP (1~100mM). Through testing and optimization, we worked out an optimal chemical cleavage protocol to remove the azidomethyl group capping the 3'-OH of the nucleotide analogues under conditions that were compatible with DNA. We also performed four continuous cycles of DNA polymerase reactions incorporating each NRT in sequential order in order to validate the use of the NRTs in SBS.

3.6 References

1. Harris, T. D.; Buzby,P. R.; Babcock, H.; Beer, E.; Bowers, J.; Braslavsky, I.; Casey, M.; Colonell, J.; DiMeo, J.; Efcavitch, J. W.; *et al*. Single-molecule DNA sequencing of a viral genome. *Science,* **2008**, *320*, 106-109.

2. Ju, J.; Li Z.; Edwards, J.; Itagaki, Y. Massive parallel method for decoding DNA and RNA. United States Patent 6,664,079 USA, **2003**.

3. Li. Z.; Bai, X.; Ruparel, H.; Kim, S.; Turro, N. J.; Ju, J. A photocleavable fluorescent nucleotide for DNA sequencing and analysis. *Proc Natl Acad Sci USA,* **2003**, *100*, 414-419.

4. Ruparel, H., *et al.* Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc Natl Acad Sci U S A* **2005**, *102,* 5932-5937.

5. Seo, T. S.; Bai, X.; Kim, D. H.; Meng, Q.; Shi, S.; Ruparel, H.; Li, Z.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci U S A* **2005**, *102,* 5926-5931.

6. Ju, J., *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **2006**, *103,* 19635-19640.

7. Pelletier, H.; Sawaya M. R.; Kumar, A.; Wilson, S. H.; Kraut, J. Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **1994**, *264,* 1891-1903.

8. Gololobov, Y. G.; Zhmurova, I. N.; Kasukin, L. F., Sixty Years of Staudinger Reaction. *Tetrahedron* **1981,** 37, 437-472.

9. Zavgorodny, S.; Polianski, M.; Besidsky, E.; Kriukov, V.; Sanin, A.; Pokrovskaya, M.; Gurskaya, G.; Lonnberg, H.; Azhayev, A. 1-Alkylthioalkylation of nucleoside hydroxyl functions and its synthetic applications: a new versatile method in nucleoside chemistry. *Tetrahedron Letters,* **1991**, *32*, 7593-7596.

10. Zavgorodny, S.; Pechenov, A. E.; Shvets, V. I.; Miroshnikov, A. I. S,X-acetals in nucleoside chemistry. III. Synthesis of 2'-and 3'-*O*-azidomethyl derivatives of ribonucleosides. *Nucleosides, Nucleotides & Nucleic Acids,* **2000**, *19*, 1977-1991.

11. Saxon, E.; Bertozzi, C. R. Cell surface engineering by a modified Staudinger reaction. *Science,* **2000**, *287*, 2007-2010.

12. Milton, J.; Ruediger, S.; Liu, X. Nucleosides/nucleotides conjugated to labels via cleavable linkages and their use in nucleic acid sequencing. *United States Patent Application US20060160081A1*, **2006**.

13. Guo, J.; Xu, N.; Li, Z.; Zhang, S.; Wu, J.; Kim, D. H.; Marma, M. S.; Meng, Q.; Cao, H.; Li, X.; Shi, S.; Yu, L.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* **2008**, *105*, 9145-4150.

Chapter 4: Design, Synthesis, and Evaluation of a Novel Class of Cleavable Fluorescent Nucleotide Reversible Terminators Containing Substituted 2-Azidomethyl Benzoic Acid Linker for DNA Sequencing by Synthesis

4.1 Introduction

With the successful synthesis and evaluation of nucleotide reversible terminators (NRTs) modified at the 3’ position with an azidomethyl functional group (3’-O-N_3-dNTPs), we verified the feasibility of nucleotide modification using the azido moiety. In order to implement nucleotide analogues into our DNA sequencing by synthesis (SBS) platform, we need to design, synthesize, and evaluate a set of cleavable fluorescent nucleotide reversible terminators (CF-NRTs) that are tailored to our SBS approach. Ju *et al.* proposed the SBS methodology using CF-NRTs to sequence surface-immobilized DNA.[1, 2] In this approach, the CF-NRTs are modified at two specific locations (Fig. 4.1) so that they are still recognized by DNA polymerase as substrates: (i) a different fluorophore with a distinct fluorescent emission is linked to each of the four bases through a cleavable linker and (ii) the 3’-OH group is capped by a small chemically reversible moiety.

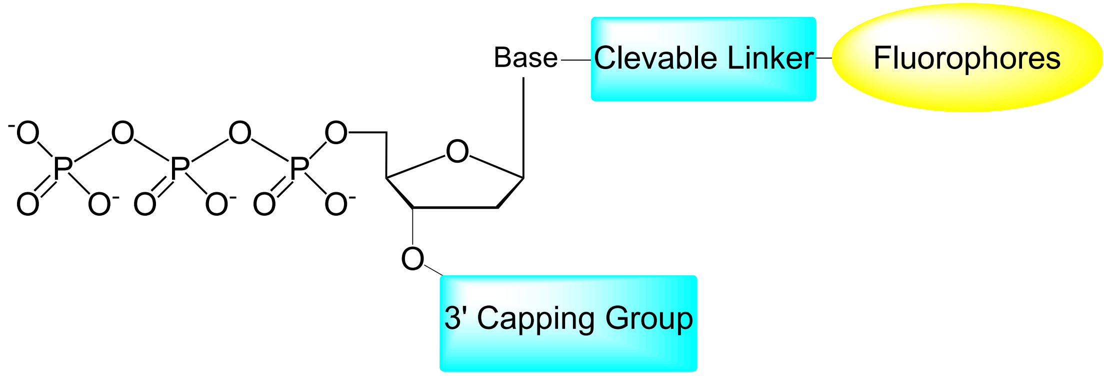


Fig. 4.1. General Structure of Cleavable Fluorescent Nucleotide Reversible Terminators

During the sequencing reaction, DNA polymerase incorporates only a single CF-NRT complementary to the base on a DNA template covalently linked to a surface. In order to temporarily pause the sequencing reaction and to accurately sequence through homopolymeric regions, the 3' capping group of the nucleotide stops the polymerase reaction during the identification of the added nucleotide. After incorporation, the unique fluorescence emission is detected to reveal the incorporated nucleotide. The fluorophore is then detached through the cleavage of the linker and the 3'-OH group is regenerated by removing the capping moiety, allowing the next cycle of the polymerase reaction to proceed. Because the large surface on a DNA chip can have a high density of different DNA templates spotted, each cycle can identify many bases in parallel, allowing the simultaneous sequencing of a large number of DNA molecules.

4.2 Experimental Rationale and Overview

According to our general strategy for SBS, four nucleotides (A, C, G, T) were modified as reversible terminators by capping the 3'-OH of those nucleotides with a small reversible moiety, so that those nucleotides could still be recognized by DNA polymerase.[2] Unique fluorophores were attached to the 5 position of pyrimidines (C and U) and the 7 position of purines (A and G) through cleavable linkers. Both a 2-nitrobenzyl linker[3-6] and an allyl linker[2, 7] were extensively studied. Near-UV irradiation was used to cleave the 2-nitrobenzyl linker, while removal of the allyl linker was achieved via Pd-catalyzed deallylation. Although DNA templates were accurately sequenced using both sets of nucleotides containing the linkers and allyl capping group described above, to discover new linker chemisties and cleavage conditions that could be milder and more DNA-compatible, we decided to explore other chemical groups that could potentially be used as linkers and capping groups in modified nucleotides for DNA sequencing by synthesis.

In 1991, Zavgorodny *et al*[8, 9] first reported capping of the 3'-OH of the nucleoside with an azidomethyl group, which can be cleaved under mild conditions with triphenylphosphine. Both Milton's group[10] and our group[11, 12] have proposed the azidomethyl moiety to cap the 3'-OH of modified nucleotides in DNA sequencing by synthesis based on the general principles that we reported for design of nucleotide reversible terminators.[1, 2, 4] Through research efforts established in the previous chapter, we confirmed that the azidomethyl moiety is a very efficient and effective 3' capping group. To design a class of CF-NRTs for sequencing by synthesis, azido based linkers, which can be cleaved under the same conditions as are used for the removal of the 3'-O-azidomethyl capping group, are preferred for attaching fluorophores to bases. This dual modification of nucleotides with a similar chemical moiety allows a single cleavage reaction to remove the linker and the 3' capping group. In 2001, Sekine *et al* reported that 2-azidomethylbenzoyl (Fig. 4.2) can be used as a protecting group in nucleosides and it can be cleaved efficiently by triphosphine.[13]

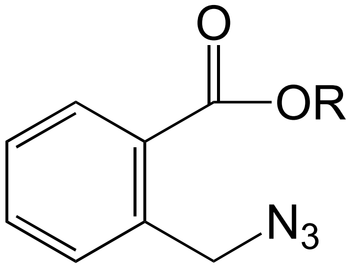


Fig. 4.2. Structure of 2-azidomethylbenzoyl group

It was expected that once the 2-azidomethylbenzoyl group is reduced to an amino group, rapid ring closure would occur, leading to cleavage from the base at this site leaving behind a small remnant (Fig. 4.3). Based on Sekine's study, we suggest that substituted 2-azidomethyl benzoyl linkers could be used to attach fluorophores to the bases of nucleotides.[14]

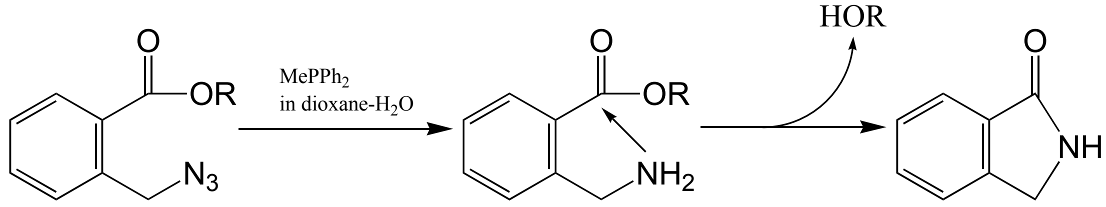


Fig. 4.3. Reaction scheme of 2-azidomethylbenzoyl reduction by methyldiphenylphosphine ($MePPh_2$)

By integrating our nucleotide modification rationale with our sequencing strategy, we report the design, synthesis, and evaluation of a novel class of cleavable fluorescent nucleotide reversible terminators (3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, Fig. 4.4) for DNA sequencing by synthesis.

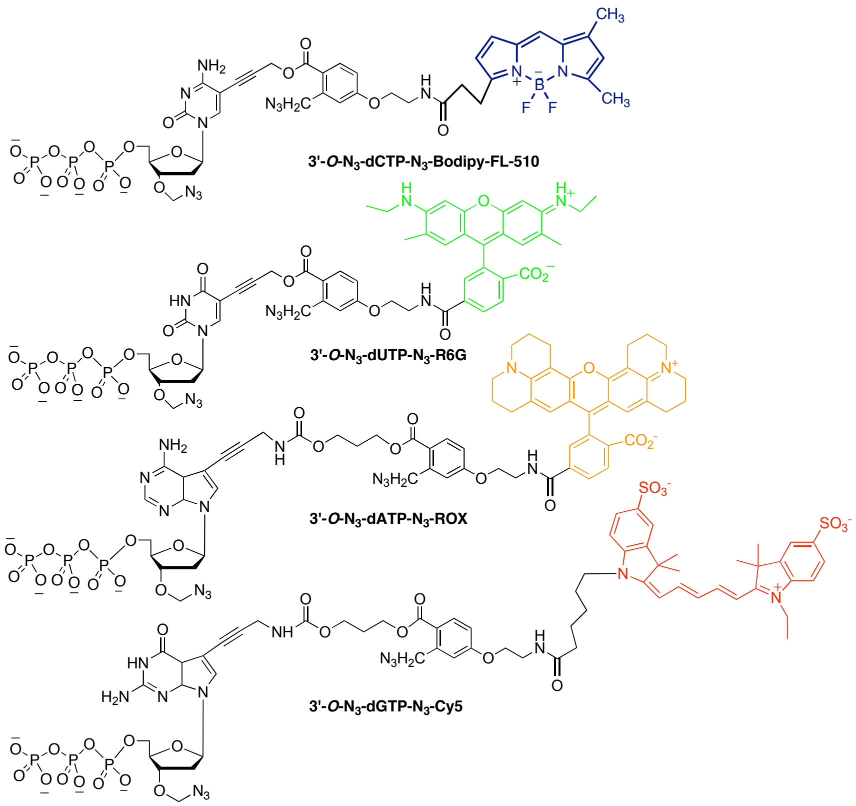


Fig. 4.4. Structures of 3'-O-N$_3$-dNTP-azidomethylbenzoyl-fluorophores, a novel class of cleavable fluorescent nucleotide reversible terminators

Those nucleotide analogs were designed to have fluorescent dyes attached to bases via novel substituted 2-azidomethylbenzoyl linkers. The 3'-OH of this class of nucleotides were also capped by an azidomethyl moiety. We demonstrate here that these nucleotides can be efficiently incorporated into a growing DNA strand by the mutant 9^{o}N DNA polymerase and temporarily terminate the polymerase reaction. After determination of the incorporated nucleotide by fluorescence detection, the 3'-OH capping group and the fluorophore are simultaneously removed under DNA-compatible Staudinger conditions. The regeneration of the free 3'-OH group allows reinitiation of the polymerase reaction. Accurately sequencing nine bases of a DNA template was demonstrated using this set of CF-NRTs, 3'-O-N$_3$-dNTP-azidomethylbenzoyl-fluorophores, in combination with a set of unlabeled nucleotides, 3’-O-N_3-dNTPs (NRTs) on a DNA chip using a four-color fluorescent scanner.

4.3 Results and Discussion

4.3.1. Synthesis of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores

Synthetic work for the whole set of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores was accomplished by Dr. Huanyan Cao in our organic chemistry laboratory. This was achieved with two different synthetic strategies. For the synthesis of modified pyrimidine nucleotides, 3’-O-N_3-dUTP-azidomethylbenzoyl-R6G and 3’-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-FL-510, linkers were attached to nucleosides before the triphosphorylation step. After triphosphorylation and deprotection, fluorophores were then attached to the nucleotides in one step. The strategy of synthesis of modified purine nucleotides, 3’-O-N_3-dATP-azidomethylbenzoyl-ROX and 3’-O-N_3-dGTP-azidomethylbenzoyl-Cy5, was slightly different due to the chemical properties of the purines. First the nucleotides without linkers were synthesized, and then the fluorophores were separately attached to linkers in 3 steps. Finally the nucleotides were coupled to the linker-fluorophore intermediates.

4.3.2. Polymerase single base extension and subsequent cleavage reactions of 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2 in solution and characterization by MALDI-TOF MS

In order to establish that the newly synthesized set of CF-NRTs, 3’-O-N_3-dUTP-azidomethylbenzoyl-fluorophores, could be used in SBS, it was crucial to verify that they would be recognized by DNA polymerase as a substrate during the polymerase reaction and both the 3'-azidomethyl group and the azidomethylbenzoyl linker could be cleaved after incorporation. Hence, we performed single base extension using a model compound, 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 (Fig. 4.5, m.w.=791: analogous to 3'-O-N_3-dUTP-azidomethylbenzoyl-R6G, but without fluorophore attachment for synthetic convenience) along with a self-priming loop template (SP26T, m.w.=7,966), 20 mM Mn^{2+}, and 9°N Thermopolymerase at 65°C for 25 minutes.

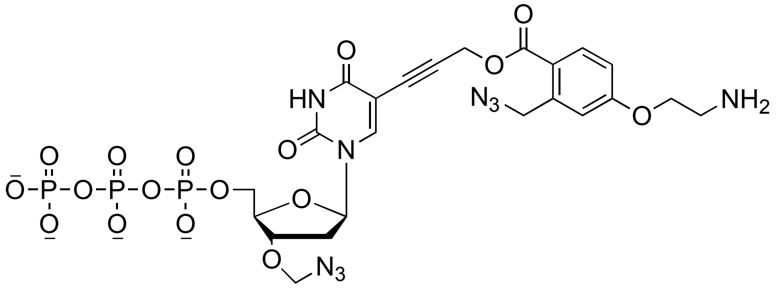


Fig. 4.5. Stucture of 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2

The extension products were analyzed with a MALDI-TOF mass spectrometer. If the modified nucleotide analogue was successfully incorporated, the template would have been extended by one base, increasing its mass from 7,966 to 8,582. As shown in the resulting MALDI-TOF MS spectrum (Fig. 4.6), the initial peak of the primer at m/z=7,966 completely disappeared, replaced by a single peak at m/z=8,582 that corresponded to the correct extension product. Thus, we verified that the dual-modified nucleotide analogue was a suitable substrate for DNA polymerase.

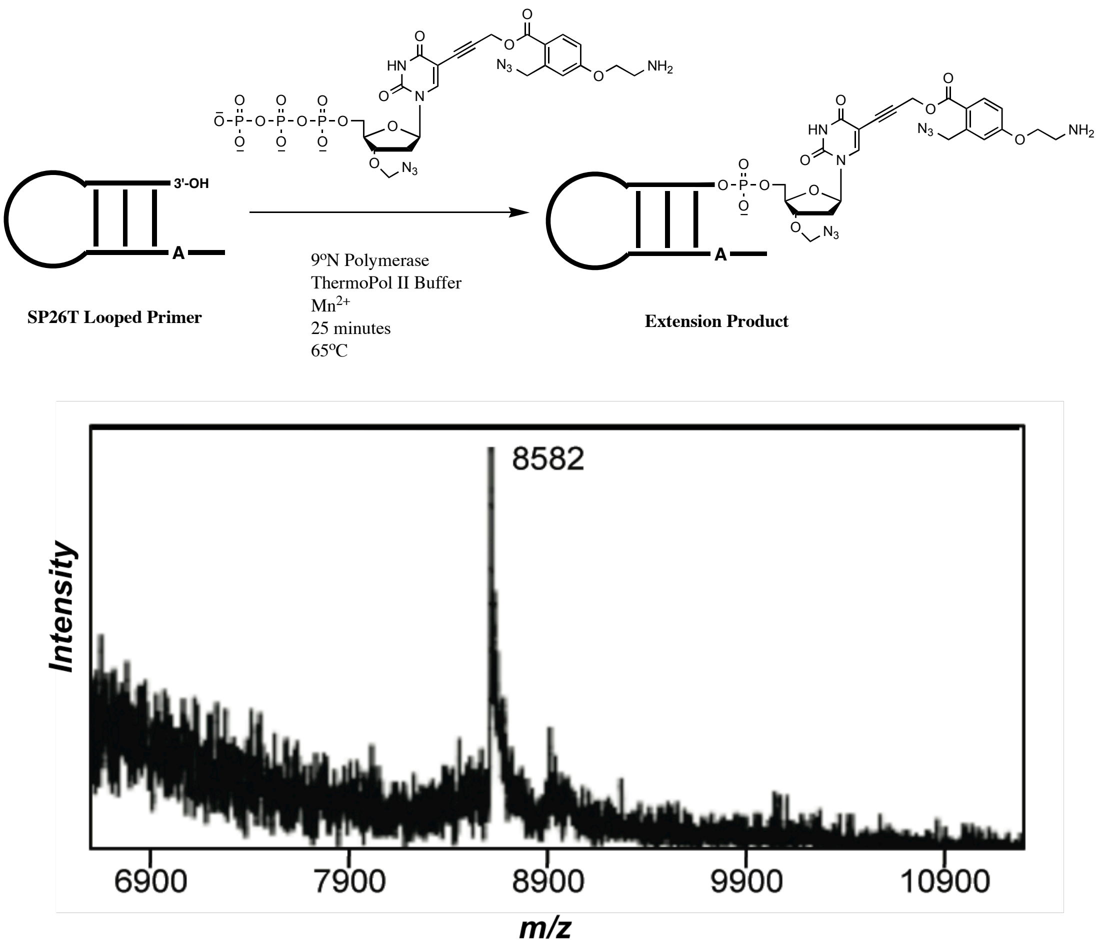


Fig. 4.6. Single base extension reaction scheme (top) and resulting MALDI-TOF MS spectrum (bottom) of 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2

The next evaluation of the model nucleotide analogue was to investigate whether the 3’ capping group and the azidomethylbenzoyl linker could be cleaved simultaneously. TCEP solutions of various pH were tested for the deprotection/cleavage reaction. Under acidic conditions (pH=5.5), the 3’ azidomethyl moiety was removed to restore the hydroxyl group at the 3’ position. However, the azidomethylbenzoyl linker was not completely cleaved due to failure to complete the ring-closure reaction (Fig. 4.7).

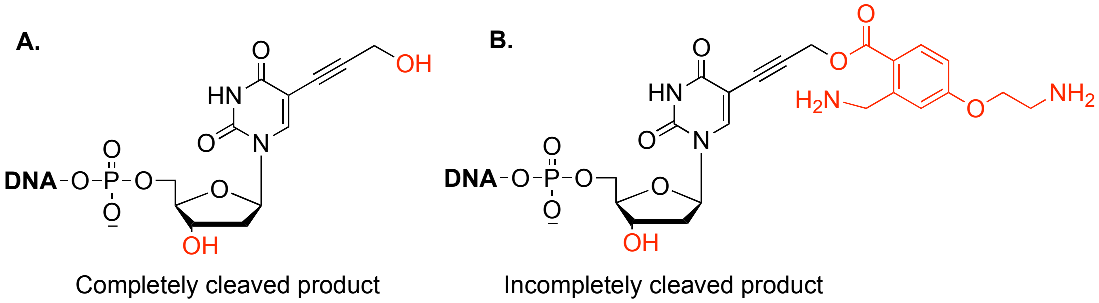
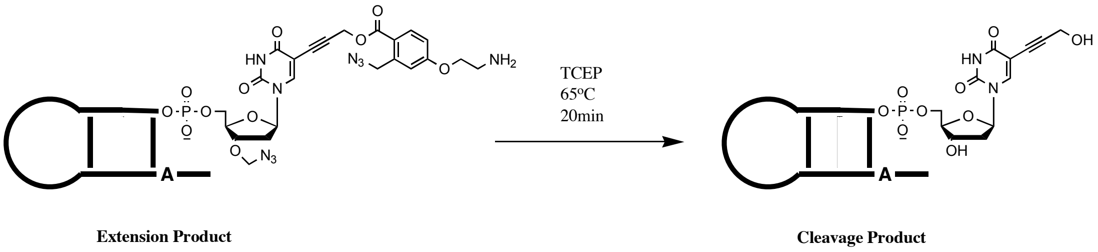


Fig. 4.7. Structures of A) completely cleaved model nucleotide analogue and B) incompletely cleaved model nucleotide analogue with residual linker still attached

As a result, two mass peaks, one corresponding to complete linker cleavage product (m/z=8309, relative abundance = ~45%) and one to incomplete linker cleavage product (m/z=8501, relative abundance = ~55%), appeared in MALDI-TOF MS spectra (Fig. 4.8A). When the pH of the TCEP solution was raised to a basic level (pH9.0), the relative abundance of completely cleaved products increased dramatically (Fig. 4.8B), confirming the correlation between pH and cleavage efficiency. Finally, at pH10.0, there was almost total removal of the azidomethylbenzoyl linker (Fig. 4.8C).



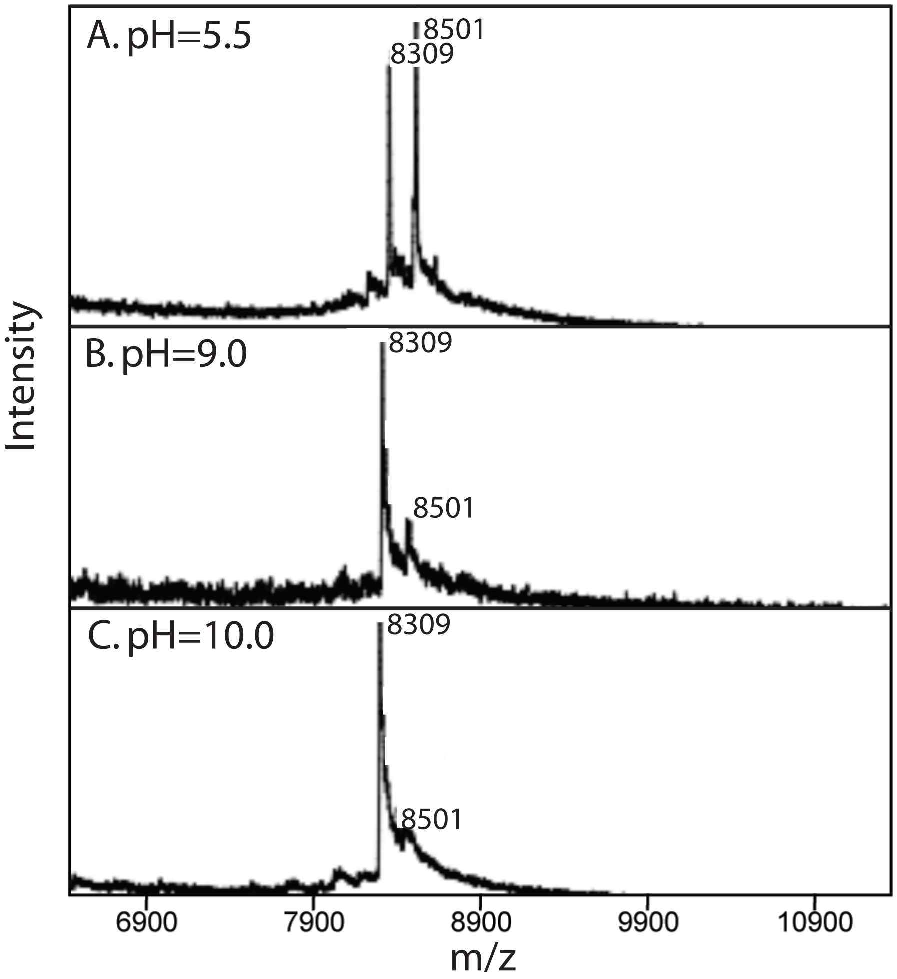


Fig. 4.8. TCEP cleavage scheme of 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2 (top) and resulting MALDI-TOF MS spectrum (bottom) at various pHs: A) pH=5.5, B) pH=9.0, C) pH=10.0

4.3.3. Polymerase extension of the complete set of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores as reversible fluorescent nucleotide terminators in solution and characterization by MALDI-TOF MS

Through previous experiments with the model nucleotide analogue, 3’-O-N_3-dUTP-azidomethylbenzoyl-NH_2, we established what we expected would be general incorporation and cleavage reaction conditions for the set of novel CF-NRTs, 3’-O-

N_3-dUTP-azidomethylbenzoyl-fluorophores. To verify that the four 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores we synthesized can work as reversible fluorescent nucleotide terminators and be incorporated accurately in a base-specific manner in a polymerase reaction, single base extension reactions with four different self-priming DNA templates whose subsequent complementary base was either A, C, G, or T were carried out in solution. After the reaction, the 4 different primer extension products were analyzed using MALDI-TOF MS. The spectra are shown in Fig. 4.9. Evidently each nucleotide was successfully incorporated into its corresponding primer, leaving none of the starting primers unextended. The single mass peaks at 9216, 9358, 8855 and 9022 (*m/z,* A, G, C, and U respectively) are good indications for the efficiency and completeness of the single base extension reactions (Fig. 4.9A, C, E, G). To demonstrate that both the fluorophores and the 3'-azidomethyl group could be removed simultaneously, cleavage reactions were carried out by incubating the single base extension products in an aqueous TCEP (pH=10.0) solution at 65^{o}C for 20 minutes. As MALDI MS spectra in Fig. 4.9B, D, F and H show, the mass peaks representing the DNA extension products have completely disappeared while single peaks corresponding to the cleavage products appear at 8437, 8438, 8307, and 8308 (*m/z*) respectively. These results demonstrated that the 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores could be successfully incorporated during the DNA polymerase reaction and efficiently cleaved at both the 3’ blocking and the fluorophore linking positions. Therefore, this set of 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores meet the key requirements necessary for carrying out SBS.

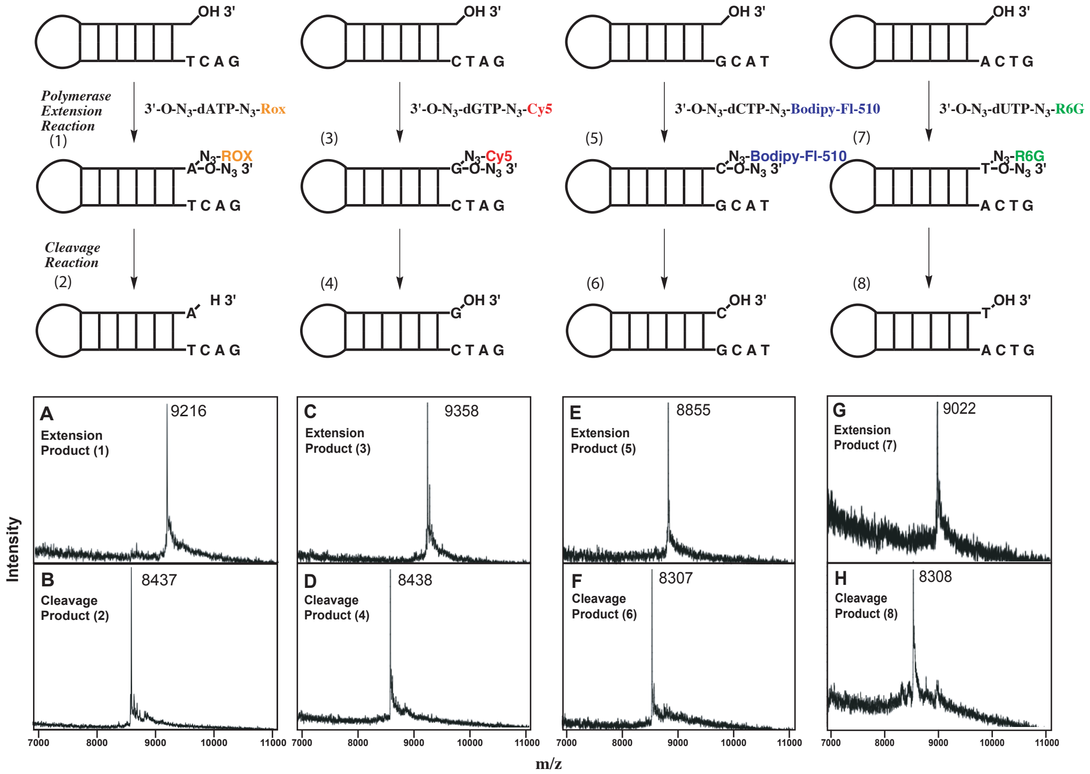


Fig. 4.9. A polymerase reaction scheme (Top) to yield DNA extension products by incorporating each of the four 3'-*O*-N_3-dNTP-azidomethylbenzoyl-fluorophores and the subsequent cleavage reaction to remove the fluorophores and 3'-azidomethyl group from the DNA extension products. MALDI-TOF MS spectra (Bottom) showing efficient base-specific incorporation of the 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and the subsequent cleavage of the fluorophores from the DNA extension products: (A) Primer extended with 3'-O-N_3-dATP-azidomethylbenzoyl-ROX (1) (peak at 9,216 m/z), (B) its cleavage product (2) (8,437 m/z); (C) primer extended with 3'-O-N_3-dGTP-azidomethylbenzoyl-Cy5 (3) (peak at 9,358 m/z), (D) its cleavage product (4) (8,438 m/z); (E) primer extended with 3'-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-FL-510 (5) (peak at 8,855 m/z), (F) its cleavage product (6) (8,307 m/z); (G) primer extended with 3'-O-N_3-dUTP-azidomethylbenzoyl-R6G (7) (peak at 9,022 m/z), and (H) its cleavage product (8) (8,308 m/z).

4.3.4. Four-color DNA sequencing on a chip using 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (CF-NRTs) and unlabeled 3'-*O*-N_3-dNTPs (NRTs)

In our attempt to perform four-color DNA sequencing on a chip, a mixture of two sets of reversible nucleotide terminators, 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (CF-NRTs) and 3'-O-N_3-dNTPs (NRTs), was used to carry out sequencing by synthesis. The role of the four 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores is to reveal the identity of the incorporated nucleotide by its unique fluorescence emission as well as to extend the DNA strand by one base, whereas the role of unlabelled 3'-O-N_3-dNTPs is only to extend the DNA strand by a single base. The reason for using a mixture of these two sets of nucleotides instead of using only 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores is because 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores leave a trace of modification, a "tail", after cleavage while 3'-O-N_3-dNTPs can turn back into natural nucleotides upon the restoration of 3'-OH and pose no hindrance to the subsequent nucleotide incorporation (Fig. 4.10).

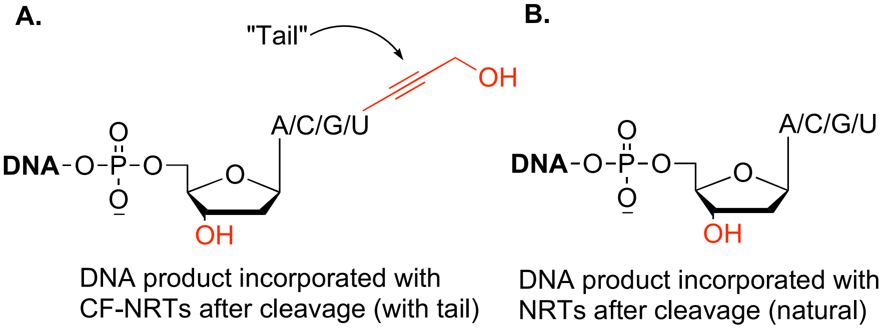

A.
"Tail"
OH
A/C/G/U
DNA-O-P-O
O
O
OH
DNA product incorporated with
CF-NRTs after cleavage (with tail)
B.
A/C/G/U
DNA-O-P-O
O
O
OH
DNA product incorporated with
NRTs after cleavage (natural)

Fig. 4.10. Comparison of structural difference of DNA incorporated A) CF-NRTs and B) NRTs after cleavage reactions

Hence, the amount of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores in the sequencing reaction mixture was kept to a minimum so that they could still produce sufficient fluorescent signals that were above the fluorescence detection system's sensitivity threshold for sequence determination. The expectation was that there would be just enough signal to detect the extension products accurately at each base, but that due to the absence of the "tail" on the majoriy of the incorporated nucleotides, the overall extension would be longer than using 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores alone. Based on this rationale, we performed sequencing by synthesis on a chip-immobilized DNA template (Fig. 4.11A) using a mixture of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3'-O-N_3-dNTPs, and the results are shown in Fig. 4.11.

Each cycle of the *de novo* sequencing by synthesis reaction on the chip was initiated by an incorporation step. During the incorporation step, the self-priming DNA strands were extended using a solution containing four 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, four 3'-O-N_3-dNTPs and the mutant 9ºN DNA polymerase. Only the complementary 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophore and 3'-O-N_3-dNTP could be recognized and incorporated during this step. After the incorporation step, a synchronization step was added to reduce the amount of un-extended DNA strands. The synchronization reaction mixture contained just 3'-O-N_3-dNTPs in relatively high concentration along with 9ºN DNA polymerase. Without this precautionary synchronization procedure, the unextended DNA would participate in the next round of incorporation, yielding mixed fluorescent signals and preventing the identification of the correct nucleotide incorporated. After washing the DNA chip, the fluorescent signal was detected by a four-color fluorescence scanner to identify the incorporated CF-NRT. The surface was then immersed in a TCEP solution to perform the cleavage step. For DNA strands extended with the 3'-O-N_3-dNTP, the 3'-OH capping group was removed and 3'-OH was restored; for DNA strands extended with the 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, which serve as the signal producer, the fluorophore was cleaved and the 3'-OH was restored. The surface of the chip was then washed and only a negligible residual fluorescent signal was detected. The restored 3’-OH on the DNA strand set the stage for the next extension reaction with 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores/3’-O-N_3-dNTPs to identify the incorporated fluorescent nucleotide complementary to the subsequent base on the template. The entire process of incorporation, synchronization, detection, and cleavage was performed multiple times to identify nine successive bases in the DNA templates. The plot of the fluorescence intensity vs. the progress of sequencing extension (raw four-color sequencing data) is shown in Fig. 4.11C. The DNA sequences were unambiguously identified from the four-color raw fluorescent data without any processing.

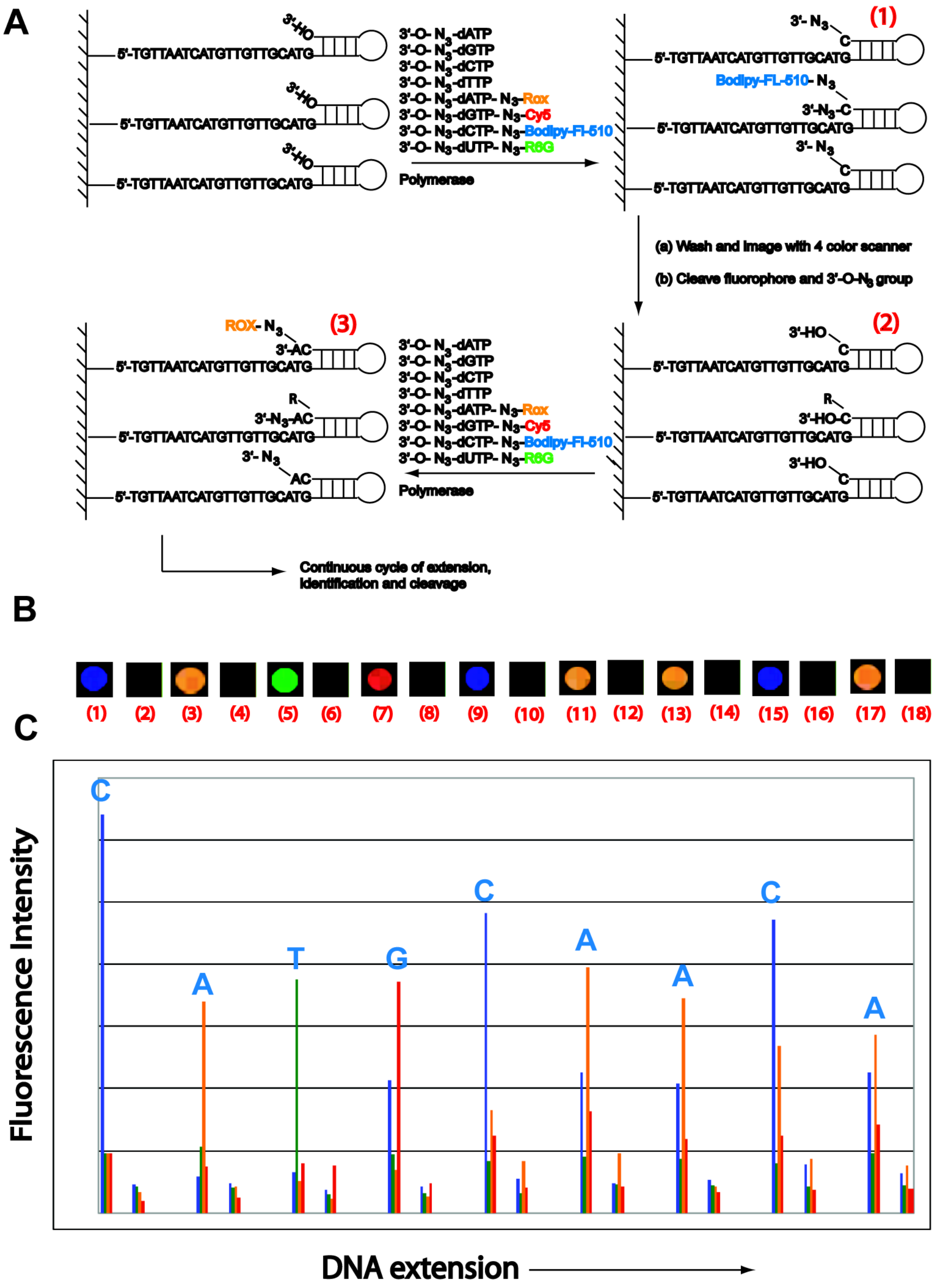


Fig. 4.11. Four-color DNA SBS with 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores. (A) A SBS with CF-NRTs scheme for four-color sequencing on a chip by using four 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3'-O-N_3-dNTPs. (B) Four-color fluorescence images for each step of the SBS: (1) incorporation of 3'-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-Fl-510 and 3'-O-N_3-dCTP; (2) cleavage of azidomethylbenzoyl-Bodipy-Fl-510 and 3'-azidomethyl group; (3)

incorporation of 3'-O-N_3-dATP-azidomethylbenzoyl-Rox and 3'-O-N_3-dATP; (4) cleavage of azidomethylbenzoyl-Rox and 3'-azidomethyl group; images 5-18 were produced similarly. (C) A plot (four-color sequencing data) of raw fluorescence emission intensity obtained by using 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3'-O-N_3-dNTPs. The small groups of peaks between the identified bases are fluorescent background from the DNA chip.

4.4 Materials and Methods

General Information. All solvents and reagents were reagent grade, purchased commercially, and used without further purification. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Oligonucleotides used as primers or templates were synthesized on an Expedite nucleic acid synthesizer (Applied BioSystems) or purchased from Midland. Mass measurement of DNA was performed on a Voyager DE MALDI-TOF mass spectrometer (Applied Biosystems). 9°N polymerase (exo-) A485L/Y409V was obtained from New England Biolabs. Phosphoramidites, reagents, and columns for oligonucleotide synthesis were purchased from Glen Research. The 3'-O-modified nucleotides (NRTs) and fluorescently labeled nucleotides (CF-NRTs) were purified by reverse-phase HPLC on a 150 X 4.6 mm C18 column (Supelco), mobile phase: A, 8.6 mM Et_3N/100 mM 1,1,1,3,3,3-hexafluoro-2-propanol in water (pH 8.1); B, methanol. Elution was performed from 100% A isocratic over 10 min followed by a linear gradient of 0 – 50% B for 20 min and then 50% B isocratic over another 30 min.

4.4.1. Synthesis of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores

The synthesis of the complete set of CF-NRTs, 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, was accomplished by Dr. Huanyan Cao in the organic synthetic chemistry team of our group. The detailed synthesis procedures for

3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores are described in our recent publications.[14]

4.4.2. Polymerase single base extension and subsequent cleavage reactions of 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 in solution and characterization by MALDI-TOF MS

We characterized the dual modified nucleotide reversible terminator, 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2, by first performing a single base extension reaction using a self-priming DNA template (SP26T: 26-mer hairpin DNA with a 4-base 5'-overhang, 5'-GTCAGCGCCGCGCCTTGGCGCGGCGC-3'). The extension reaction (total volume of 20 μL) consisted of 3'-O-azidomethyl-dUTP (180 pmol) along with 60 pmol of the SP26T self-priming DNA template, 1X Thermopol II reaction buffer, 40 nmol of $MnCl_2$ and 1 unit of 9°N DNA polymerase (exo-) A485L/Y409V. The reaction was performed at 65°C for 25 minutes. Subsequently, the extension product was purified by ethanol precipitation and Zip-Tip desalting procedure, and then analyzed using MALDI-TOF MS.

The cleavage of the DNA extension product incorporated with 3'-O-N_3-dUTP-azidomethylbenzoyl-NH_2 was accomplished by resuspending the product in 40 μL of 100mM TCEP solution (pH 5.5) at 65°C for 30 minutes. The cleaved products were then analyzed by MALDI-TOF MS. The same cleavage reaction was repeated at pH9.0 and pH10.0. Again, the cleaved products were analyzed using MALDI-TOF MS without further purification.

4.4.3. Polymerase extension of the complete set of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores as reversible fluorescent nucleotide terminators in solution and characterization by MALDI-TOF MS

We characterized the four cleavable fluorescent nucleotide reversible terminators (CF-NRTs), 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (3'-N_3-O-dATP-azidomethylbenzoyl-ROX, 3'-N_3-O-dGTP-azidomethylbenzoyl-Cy5, 3'-N_3-O-dCTP-azidomethylbenzoyl-Bodipy-FL-510, and 3'-N_3-O-dUTP-azidomethylbenzoyl-R6G) by performing four separate DNA-extension reactions, each with a different self-priming DNA template that allowed the corresponding CF-NRT analogues to be incorporated. The resulting DNA extension products were analyzed by MALDI-TOF MS. The following four self-priming DNA templates (26-mer hairpin DNA with a 4-base 5'-overhang) were used for the extension: 5'-GACTGCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dATP-azidomethylbenzoyl-ROX; 5'-GATCGCGCCGCGC CTTGGCGCGGCGC-3' for 3'-O-N_3-dGTP-azidomethylbenzoyl-Cy5; 5'-ATCGGCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-FL-510; and 5'-GTCAGCGCCGCGCCTTGGCGCGGCGC-3' for 3'-O-N_3-dUTP-azidomethylbenzoyl-R6G. Each extension reaction (total volume of 20μL) contained one of the four 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (180pmol) along with 60pmol of its corresponding self-priming DNA template, 1X Thermopol II reaction buffer, 40nmol $MnCl_{2,}$ and 1 unit 9^{o}N DNA polymerase (exo-) A485L/Y409V. The reaction was performed at 94^{o}C for 5 minutes, 4^{o}C for 5 minutes, and 65^{o}C for 25 minutes. Subsequently, the extension product was purified by reverse-phase HPLC. The fraction containing the desired product was collected and freeze-dried for analysis using MALDI-TOF MS. The cleavage of the DNA extension products incorporated with 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores were accomplished by re-suspending them in 40µL of 100mM TCEP solution (pH=10.0) at 65°C for 25 minutes. The cleaved products were then analyzed using MALDI-TOF MS.

4.4.4. Four-color DNA sequencing on a chip using 3′-O-N_3-dNTP-azidomethylbenzoyl-fluorophores (CF-NRTs) and unlabelled 3′-O-N_3-dNTPs (NRTs)

The DNA chip was constructed by immobilizing a 5’-amino-modified looped oligonucleotide on a CodeLink microarray slide (GE Healthcare) (27). Ten microliters of a solution consisting of 3’-O-N_3-dCTP-azidomethylbenzoyl-Bodipy-FL-510 (2.5pmol), 3’-O-N_3-dUTP-azidomethylbenzoyl-R6G (5pmol), 3’-O-N_3-dATP-azidomethylbenzoyl-ROX (7 pmol), 3’-O-N_3-dGTP-azidomethylbenzoyl-Cy5 (2.5pmol), 3’-O-N_3-dCTP (60pmol), 3’-O-N_3-dTTP (60pmol), 3’-O-N_3-dATP (60pmol), 3’-O-N_3-dGTP (60pmol), 1 unit of 9°N DNA polymerase (exo-) A485L/Y409V, 20nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer was spotted on the surface of the chip, where the self-primed DNA moiety was immobilized. This extension mixture was incubated at 65°C for 15 minutes. Upon the completion of the extension reaction, a capping reaction solution, consisting of 70 pmol each of 3’-O-N_3-dATP, 3’-O-N_3-dGTP, 3’-O-N_3-dCTP, 3’-O-N_3-dTTP, 1 unit of 9°N DNA polymerase (exo-) A485L/Y409V, 20nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer, was spotted on the same spot on the DNA chip and incubated at 65°C for 20 minutes. The chip was then washed with SPSC buffer (0.1% Tween-20) for 2 minutes, followed by rinsing with dH_2O. The blow-dried chip was scanned with a 4-color ScanArray Express scanner (Perkin–Elmer Life Sciences) to detect the fluorescence signal. The four lasers on the 4-color scanner had excitation wavelengths of 488, 543, 594, and 633 nm and emission filters centered at 522, 570, 614, and 670 nm. For the cleavage process, the DNA chip was placed inside a plastic chamber filled with 100mM TCEP (pH 10.0) for 20 minutes of incubation at 65^oC. After washing the surface with dH_2O, the chip was scanned again to compare the intensity of fluorescence after cleavage with the original fluorescence intensity. This process was repeated by carrying out the next round of extension using the 3’-O-N_3-dNTPs-azidomethylbenzoyl-fluorophore/3’-O-N_3-dNTPs solution with subsequent synchronization, washing, fluorescence detection, and cleavage processes performed as described above.

4.5 Conclusion

We have synthesized four novel CF-NRTs: 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores containing substituted 2-azidomethylbenzoyl linkers and used them along with four unlabeled NRTs: 3’-O-N_3-dNTPs to produce four-color *de novo* DNA sequencing data on a chip using our sequencing by synthesis method. After each round of sequencing, the fluorophores linked to the CF-NRTs are removed and the 3’-OH of both 3’-*O*-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3’-O-N_3-dNTPs are restored so it is possible for all the DNA templates to be used in the next round of sequencing. We have experimentally determined the ratio of the 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3’-*O*-N_3-dNTPs to be used in the sequencing mixture. A relatively small amount of CF-NRTs was used to ensure that there are less traces of modification left on the growing DNA stand. Since both 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores and 3'-O-N_3-dNTPs are reversible terminators, which allow the sequencing of each base in a serial manner, they can accurately determine the homopolymeric regions of DNA. In addition, due to the fact that all of the steps of our SBS approach are performed on a DNA chip, there is no need for electrophoretic DNA fragment separation as in the classical Sanger sequencing method since the sequencing products are generated in a base-by-base manner.

The successful design, synthesis, and evaluation of this novel class of CF-NRTs, 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores, have established a foundation for nucleotide modification using azido-based functional groups for sequencing by synthesis (SBS). The fact that this set of CF-NRTs can be recognized by DNA polymerase as substrate and participate in four-color SBS reactions is an excellent indication that azido-modified CF-NRTs have the potential to improve SBS due to the mild and DNA-friendly conditions with which they can be cleaved. However, the implementation of 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores in SBS has its drawbacks as well, the major one being the relatively short sequencing read length (around ten bases). One factor may contribute to the limitation of bases sequenced. When we conducted single base extension reactions in solution, it was noted that a small percentage of the azidomethylbenzoyl linker was not cleaved. This incomplete cleavage of linker translated to residual fluorescent signal interfering with the identification of the next base. Even with effort at synchronization, the accumulation of multiple lagging signals eventually disrupted the base-calling of the downstream sequence. Hence in efforts to advance SBS using CF-NRTs to an ultra high-throughput and robust platform, further optimization will be needed.

4.6 References

1. Ju, J.; Li Z.; Edwards, J.; Itagaki, Y. Massive parallel method for decoding DNA and RNA. United States Patent 6,664,079 USA, **2003**.
2. Ju, J., *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **2006**, *103,* 19635-19640.
3. Li, Z.; Bai, X.; Ruparel, H.; Kim, S.; Turro, N. J.; Ju, J. A photocleavable fluorescent nucleotide for DNA sequencing and analysis. *Proc Natl Acad Sci USA*, **2003**, *100*, 414-419.
4. Ruparel, H., *et al.* Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc Natl Acad Sci U S A* **2005**, *102,* 5932-5937.
5. Seo, T. S.; Bai, X.; Ruparel, H;, Li, Z.; Turro, N. J.; Ju, J. Photocleavable fluoescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci USA*, **2004**, *101*, 5488-5493.
6. Meng, Q., *et al.* Design and synthesis of a photocleavable fluorescent nucleotide 3'-O-allyl-dGTP-PC-bodipy-FL-510 as a reversible terminator for DNA sequencing by synthesis. *J Org Chem,* **2006**, *71,* 3248-3252.
7. Bi, L.; Kim, D. H.; Ju, J. Design and synthesis of a chemically cleavable fluorescent nucleotide, 3’-O-allyl-dGTP-allyl-Bodipy-FL-510, as a reversible terminator for DNA sequencing by synthesis. *J. Am Chem Soc*. **2006**, *128,* 2542-2543.
8. Zavgorodny, S.; Polianski, M.; Besidsky, E.; Kriukov, V.; Sanin, A.; Pokrovskaya, M.; Gurskaya, G.; Lonnberg, H.; Azhayev, A. 1-Alkylthioalkylation of nucleoside hydroxyl functions and its synthetic

applications: a new versatile method in nucleoside chemistry. *Tetrahedron Letters,* **1991**, *32*, 7593-7596.

9. Zavgorodny, S.; Pechenov, A. E.; Shvets, V. I.; Miroshnikov, A. I. S,X-acetals in nucleoside chemistry. III. Synthesis of 2'-and 3'-*O*-azidomethyl derivatives of ribonucleosides. *Nucleosides, Nucleotides & Nucleic Acids,* **2000**, *19*, 1977-1991.

10. Barnes, C.; Balasubramanian, S.; Liu, X.; Swerdlow, H.; Milton, J. Labelled nucleotides. *US Patent 7,057,026* **2006**.

11. Guo, J.; Xu, N.; Li, Z.; Zhang, S.; Wu, J.; Kim, D. H.; Marma, M. S.; Meng, Q.; Cao, H.; Li, X.; Shi, S.; Yu, L.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* **2008**, *105*, 9145-4150.

12. Ju, J.; Cao, H.; Li, Z.; Meng, Q.; Guo, J.; Zhang, S.; Yu, L. Design and synthesis of cleavable fluorescent nucleotides as reversible terminators for DNA sequencing by synthesis. *US Patent WO2009051807A1*, **2009**.

13. Wada, T.; Ohkubo, A.; Mochizuki, A.; Sekine, M. 2-(Azidomethyl)benzoy as a new protection group in nucleosides. *Tetrahedron Lett.*, **2001**, *42*, 1069-1072.

14. Cao, H.; Yu, L.; Meng, Q.; Li, Z.; Ju, J. Design, synthesis and evaluation of a class of fluorescent nucleotides containing substituted 2-azidomethyl benzoic acid linker as reversible terminators for DNA sequencing by synthesis. **2009**, *in preparation*.

Chapter 5: Four-color DNA Sequencing by Synthesis (SBS) Improvements using Cleavable Fluorescent Nucleotide Reversible Terminators

5.1 Introduction

For many years, our lab has been developing four-color DNA sequencing by synthesis technologies by utilizing rationally-designed cleavable fluorescent nucleotide reversible terminators (CF-NRTs).[1-5] Four nucleotides (A, C, G and T) are modified as CF-NRTs by attaching a cleavable fluorophore to a specific location on the nitrogen base and capping the 3'-OH position with a small chemically reversible moiety so that they are still recognized as substrates by DNA polymerase. Self-priming DNA templates consisting of homopolymer regions were accurately sequenced using this approach.[5] A newly developed fluorescent DNA sequencing-by-synthesis system based on a similar design of the CF-NRTs has already found wide applications in genomic science.[6-8] More recently, we have explored the use of 3'-O-modified NRTs, nucleotide reversible terminators *sans* flourophore attachment, for various SBS applications. We reported the successful implementations of these NRTs to solve the homopolymer sequencing problem in conventional pyrosequencing[9] and the establishment of a novel hybrid methodology combining SBS and Sanger sequencing.[10]

The successful implementation of SBS using modified nucleotide reversible terminators has demonstrated its potential as the core sequencing technology in some high-profile genomic studies.[11] However, this methodology has yet to overcome some drawbacks and challenges in order to reach the promised land of ultra high-throughput DNA sequencing. One such drawback, mainly responsible for the relatively short read length associated with SBS, is rooted in the fundamental mechanism of SBS with CF-NRTs that is based on the identification of each base as the DNA strand is extended by simultaneously adding all four CF-NRTs to temporarily pause the DNA synthesis. Upon detection of the current base, the fluorophore and the 3'-OH blocking moiety on the incorporated CF-NRTs are removed to initiate the next cycle of incorporation. Since sequence determination relies exclusively on the fluorescent emission of the current base, any unsynchronized incorporation of nucleotides (de-phasing) would contribute to incorrect signals, which eventually would propagate sufficiently to cause miscalled bases. This de-phasing problem can be traced to two separate sources: 1) lagging behind and 2) reading ahead of the current base. Lagging occurs due to incomplete incorporation of nucleotides during the current cycle, leaving some DNA strands un-extended. Consequently, the fluorescent signal of the previous base appears during the next cycle of sequencing, causing erroneous base-calling (Fig. 5.1).

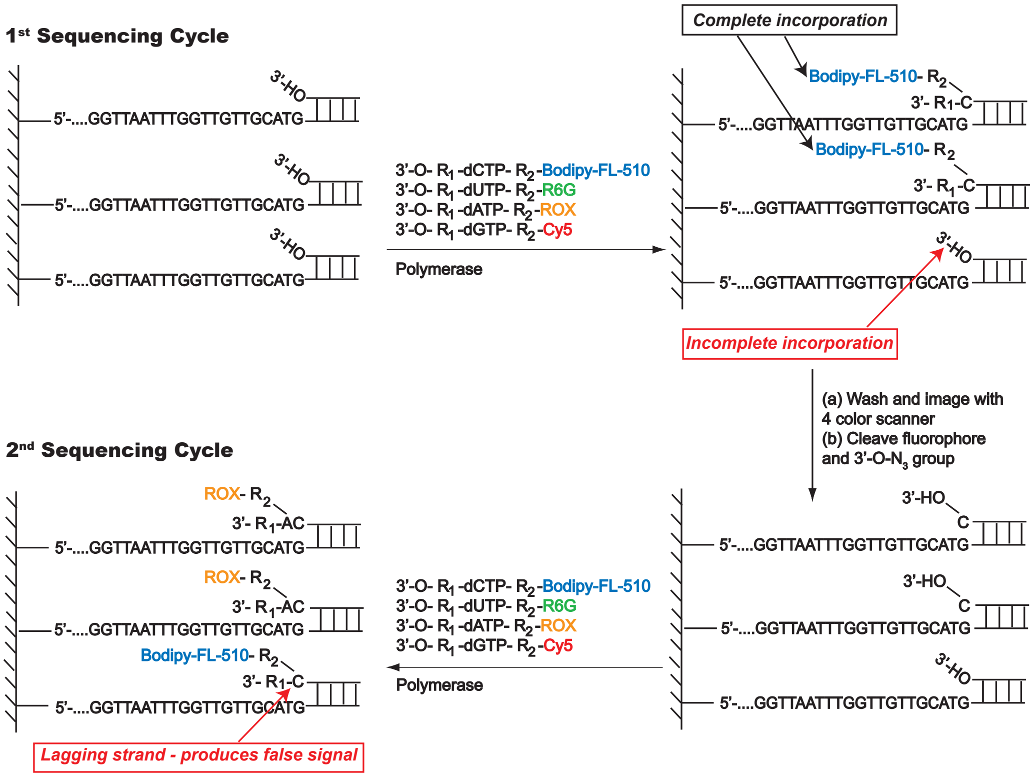


Fig. 5.1. Scheme of "lagging" sequence during SBS due to incomplete incorporation of CF-NRTs

Read ahead often occurs when 3'-OH unblocked nucleotides, mostly impurities from CF-NRT synthesis, are incorporated, resulting in the extension of more than one base during the current cycle (Fig. 5.2). Though these problems can be resolved computationally early on, as the lagging-behind or reading-ahead builds up, it eventually becomes impossible to call bases. In order to increase read length of SBS using CF-NRTs, it is necessary to address the issues of de-phasing and to optimize the entire sequencing process on a molecular level.

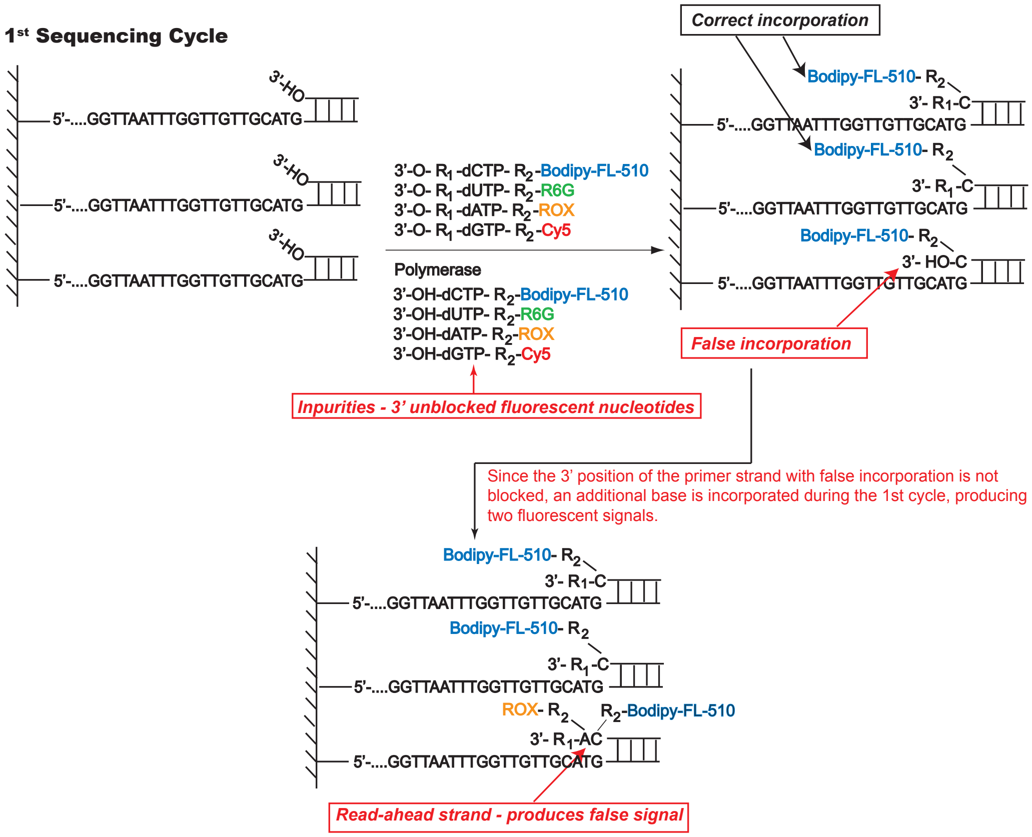

1st Sequencing Cycle
3'-HO
5'-....GGTTAATTTGGTTGTTGCATG
3'-HO
5'-....GGTTAATTTGGTTGTTGCATG
3'-HO
5'-....GGTTAATTTGGTTGTTGCATG
3'-O- R1 -dCTP- R2-Bodipy-FL-510
3'-O- R1 -dUTP- R2-R6G
3'-O- R1 -dATP- R2-ROX
3'-O- R1 -dGTP- R2-Cy5
Polymerase
3'-OH-dCTP- R2-Bodipy-FL-510
3'-OH-dUTP- R2-R6G
3'-OH-dATP- R2-ROX
3'-OH-dGTP- R2-Cy5
Inpurities - 3' unblocked fluorescent nucleotides
Correct incorporation
Bodipy-FL-510- R2
3'- R1-C
5'-....GGTTAATTTGGTTGTTGCATG
Bodipy-FL-510- R2
3'- R1-C
5'-....GGTTAATTTGGTTGTTGCATG
Bodipy-FL-510- R2
3'- HO-C
5'-....GGTTAATTTGGTTGTTGCATG
False incorporation
Since the 3' position of the primer strand with false incorporation is not blocked, an additional base is incorporated during the 1st cycle, producing two fluorescent signals.
Bodipy-FL-510- R2
3'- R1-C
5'-....GGTTAATTTGGTTGTTGCATG
Bodipy-FL-510- R2
3'- R1-C
5'-....GGTTAATTTGGTTGTTGCATG
ROX- R2
R2-Bodipy-FL-510
3'- R1-AC
5'-....GGTTAATTTGGTTGTTGCATG
Read-ahead strand - produces false signal

Fig. 5.2. Scheme of "read ahead" sequence during SBS due to incorporation of 3' unblocked fluorescent nucleotides

5.2 Experimental Rationale and Overview

In this chapter, we report a sequencing method that utilizes cleavable fluorescent nucleotide reversible terminators (CF-NRTs), non-labeled nucleotide reversible terminators (NRTs), and dideoxynucleotides (ddNTPs), and present some of the improvements this approach offers to the existing SBS technology. Instead of the conventional use of only CF-NRTs to extend DNA strands, four NRTs, nucleotides modified by capping the 3'-OH with a small reversible moiety so that they are still recognized as substrates by DNA polymerase, are mixed with CF-NRTs for carrying out the initial sequencing by synthesis reaction. NRTs, smaller and superior substrates for DNA polymerase, are used in large excess to the CF-NRTs in the extension mixture to ensure maximum nucleotide incorporation efficiency. Sequences are determined by the unique fluorescence emission of each fluorophore on the DNA products terminated by the CF-NRTs. Immediately following the detection step, a synchronization reaction is performed using only the NRTs to extend any unextended DNA strands. A capping step with dideoxynucleotides (ddNTPs) is carried out afterwards to completely eliminate the remaining unextended DNA. Upon removing the 3’-OH capping group from the DNA products generated by incorporating both CF-NRTs and NRTs and the fluorophore from the CF-NRTs, the polymerase reaction reinitiates to continue the next cycle of sequence determination (Fig. 5.3). In order to address the reading-ahead issue during each sequencing cycle, ultra-pure NRTs and CF-NRTs are used to minimize the incorporation of impure nucleotide analogues.

Based on our previous experience with CF-NRTs, particularly the 3’-O-N_3-dNTP-azidomethylbenzoyl-fluorophore set, we synthesized a new set of CF-NRTs using an azidomethyl group as a chemically reversible capping moiety in the 3’ position, and an azido-based cleavable linker to tether the fluorophores to the CF-NRTs, After fluorescence detection for sequence determination, the azidomethyl capping moiety on the 3’-OH and the fluorophore attached to the DNA extension product via the azido-based cleavable linker are efficiently removed using Tris(2-carboxyethyl) phosphine (TCEP) in aqueous solution compatible with DNA stability and function. Various DNA templates, including those with homopolymer regions were sequenced using this set of CF-NRTs in our sequencing by synthesis method on a chip and a four-color fluorescent scanner.

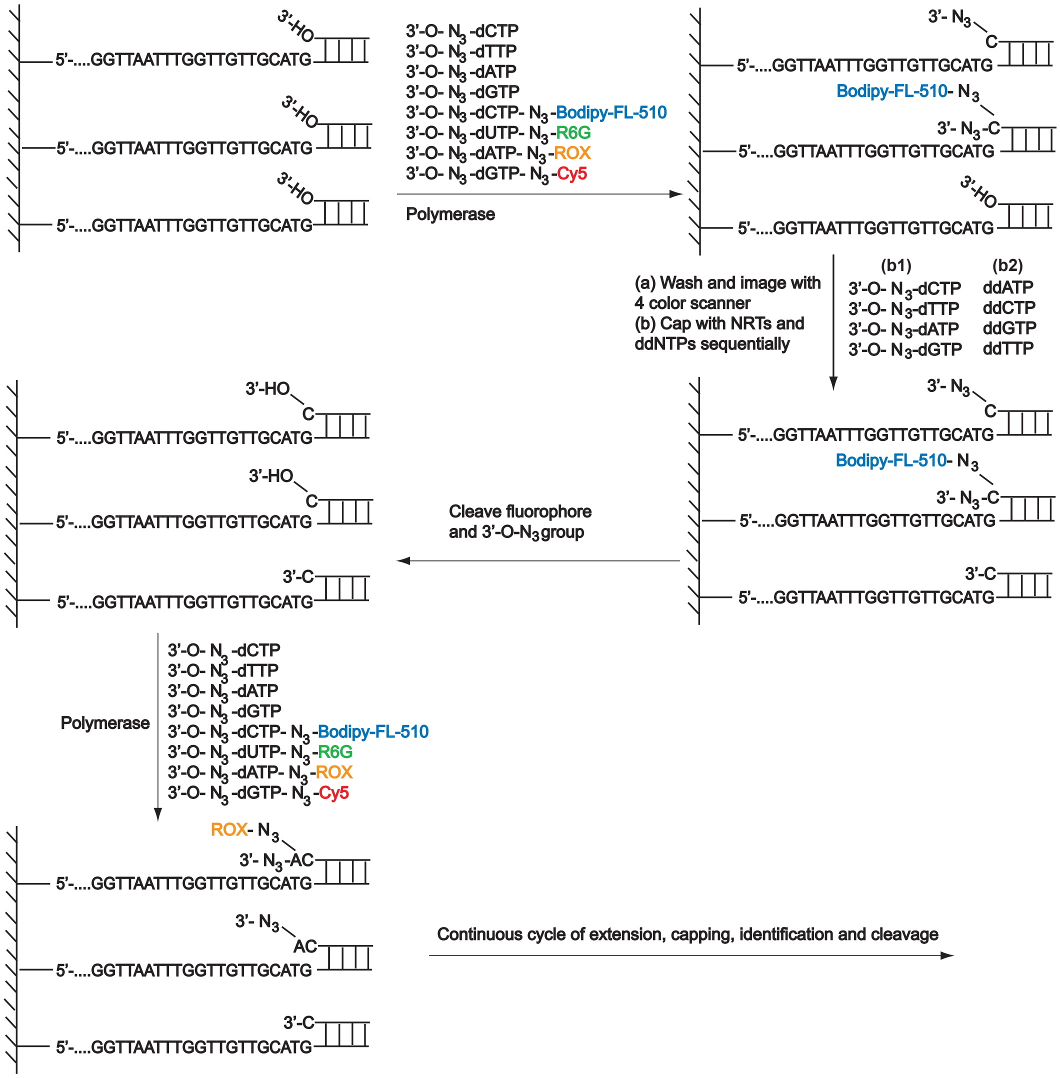


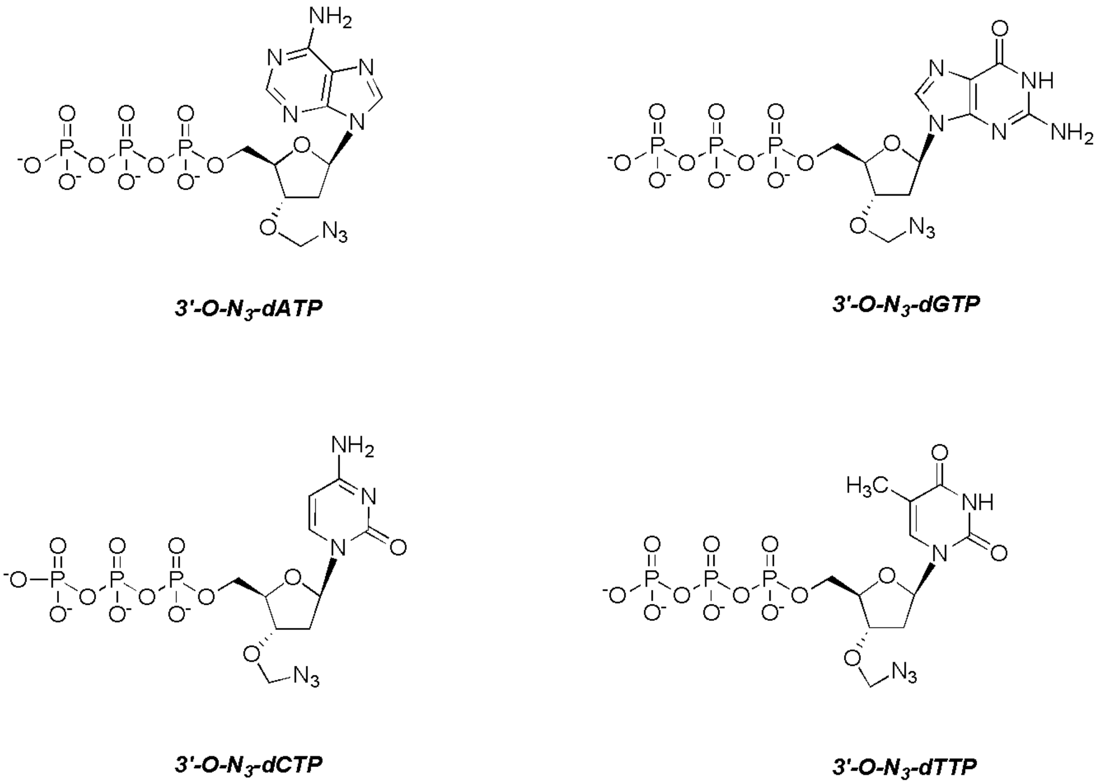
Fig. 5.3. Scheme for sequencing by synthesis using CF-NRTs, NRTs, and ddNTPs

5.3 Results and Discussion

5.3.1. Design and synthesis of cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs for SBS

The successful implementation of sequencing by synthesis is essentially dependent on the modified nucleotides used during the sequencing reactions. The design of NRTs and CF-NRTs must encompass a suitable chemical moiety for capping the 3'-OH of the nucleotide such that it temporarily terminates the polymerase reaction, enabling the unambiguous identification of the incorporated base. Capping the 3’-OH group of the nucleotides with a reversible moiety allows for the simultaneous addition of all four nucleotides, eliminating the necessity of adding each nucleotide sequentially as is done in some next-gen sequencing technologies.[12] This results in increased accuracy and reduction of the number of cycles needed during SBS. Our previous research efforts have firmly established the molecular level strategy to rationally modify the nucleotides by capping the 3’-OH with a small chemically reversible moiety for SBS.[1-5] Building on our successful nucleotide modification strategy, we explored alternative chemically reversible groups for capping the 3’-OH and linking the fluorophore to the nucleotides.

Based on Zavgorodny et al. who reported the capping of the 3’-OH group of the nucleoside with an azidomethyl moiety,[13, 14] we synthesized and evaluated four 3’-O-azidomethyl-modified NRTs (3’-O-N_3-dNTPs) (Fig. 5.4) as described in Chapter 3.



3'-O-N$_3$-dATP *3'-O-N$_3$-dGTP*

3'-O-N$_3$-dCTP *3'-O-N$_3$-dTTP*

Fig. 5.4. Structures of the nucleotide reversible terminators (NRTs), 3'-O-N$_3$-dATP, 3'-O-N$_3$-dCTP, 3'-O-N$_3$-dGTP, 3'-O-N$_3$-dTTP

The 3'-O-azidomethyl group on the DNA extension product generated by incorporating each of the NRTs was able to terminate further elongation of the template. Upon efficient removal of the reversible terminating moiety by the Staudinger reaction using aqueous Tris(2-carboxyethyl) phosphine (TCEP) solution[15, 16] followed by hydrolysis to yield a free 3'-OH group, the DNA strand was able to carry out the next cycles of the SBS.

For CF-NRTs, dual modifications are required so that they can temporarily terminate DNA synthesis and report the nucleotide that is incorporated by the means of fluorescent signaling. The 3'-OH position of the CF-NRTs is modified identically to the NRTs using the same azidomethyl group. An additional modification is needed on the CF-NRTs to tether a fluorophore via a cleavable linker. It had been shown that modified DNA polymerases could be highly tolerant to nucleotide modifications with bulky groups at the 5-position of pyrimidines (C and U) and the 7-position of purines (A and G).[17] Thus, we attached each unique fluorophore to the 5 position of C/U and the 7 position of A/G through a cleavable linker based on an azido modified moiety.[16] This azido moiety must share a cleavage mechanism that is similar to that of the 3'-O-azidomethyl group to allow the simultaneous removal of the capping group and the reporting group (fluorophore) from the DNA extension products. Following this general principle of CF-NRT design, which requires the recognition of the modified nucleotides as substrates by DNA polymerase through the attachment of a cleavable fluorophore to a specific location on the base and capping of the 3'-OH with a small chemically reversible moiety,[1, 5] we reported the synthesis and evaluation of four azido modified CF-NRTs (3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores) in Chapter 4. Due to the relatively short read length obtained using the previous set of CF-NRTs for SBS, we synthesized a new set of CF-NRTs using a different azido-based linker.[18] This set of CF-NRTs, 3'-O-N_3-dNTP-N_3-fluorophores (Fig. 5.5), had the same 3'-azidomethyl capping group as the 3'-O-N_3-dNTPs and the 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores while the cleavable linker was also based on an azido modified moiety[14] as a trigger for cleavage, with a mechanism similar to the removal of the 3'-O-azidomethyl group (shown in Fig. 5.6). This azido-based linker (N_3) can be reduced under milder conditions than the azidomethylbenzoyl linker. Complete cleavage of this azido-based linker is achievable, resulting in efficient removal of the fluorophores on the CF-NRTs.

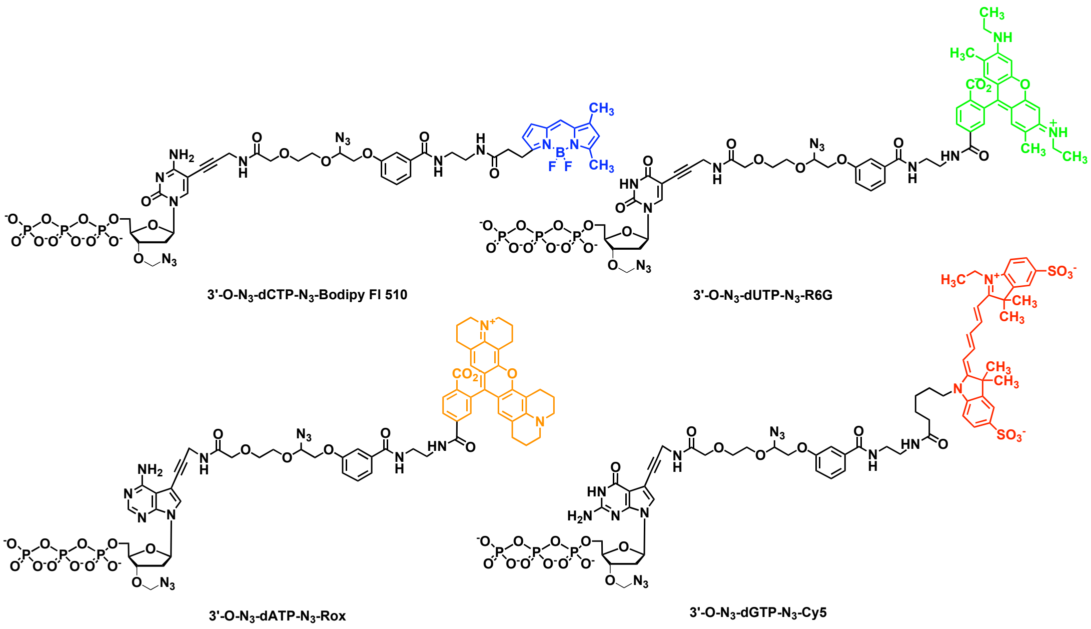

3'-O-N$_3$-dCTP-N$_3$-Bodipy Fl 510
3'-O-N$_3$-dUTP-N$_3$-R6G
3'-O-N$_3$-dATP-N$_3$-Rox
3'-O-N$_3$-dGTP-N$_3$-Cy5

Fig. 5.5. Structures of cleavable fluorescent nucleotide reversible terminators (CF-NRTs: 3’-O-N$_3$-dNTP-N$_3$-fluorophores, with the 4 fluorophores having distinct fluorescent emissions: 3’-O-N$_3$-dCTP-N$_3$-Bodipy-FL-510 ($\lambda_{abs\ (max)}$ = 502 nm; $\lambda_{em\ (max)}$ = 510 nm), 3’-O-N$_3$-dUTP-N$_3$-R6G ($\lambda_{abs\ (max)}$ = 525 nm; $\lambda_{em\ (max)}$ = 550 nm), 3’-O-N$_3$-dATP-N$_3$-ROX ($\lambda_{abs\ (max)}$ = 585 nm; $\lambda_{em\ (max)}$ = 602 nm), and 3’-O-N$_3$-dGTP-N$_3$-Cy5 ($\lambda_{abs\ (max)}$ = 649 nm; $\lambda_{em\ (max)}$ = 670 nm).

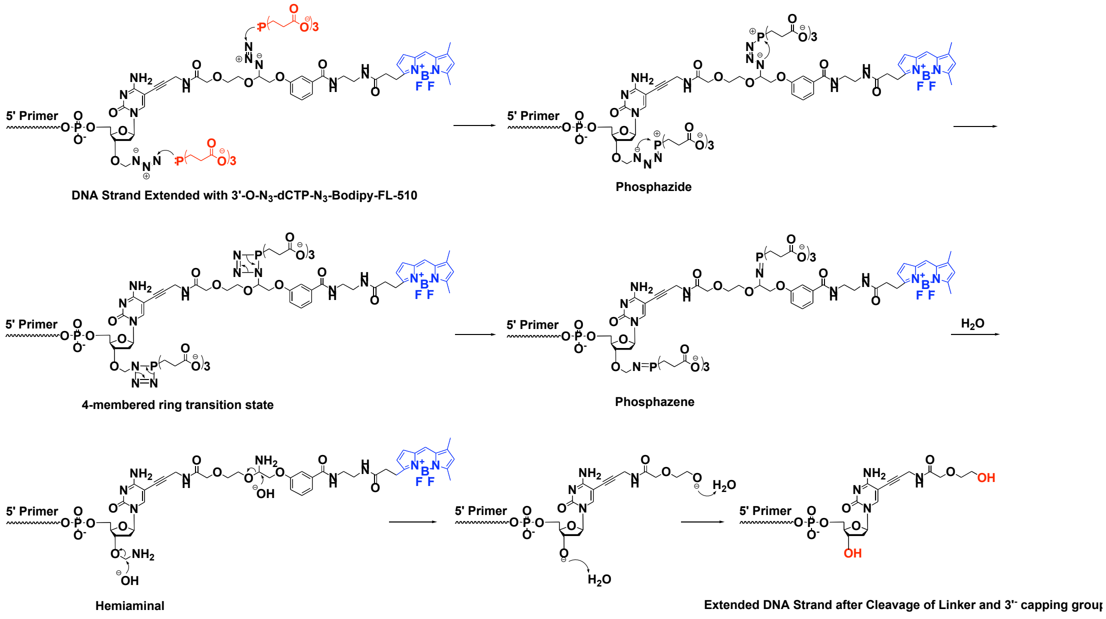

5' Primer
DNA Strand Extended with 3'-O-N$_3$-dCTP-N$_3$-Bodipy-FL-510
Phosphazide
4-membered ring transition state
Phosphazene
H$_2$O
Hemiaminal
Extended DNA Strand after Cleavage of Linker and 3'- capping grou

Fig. 5.6. Scheme of TCEP cleavage mechanism of a DNA strand extended with a CF-NRT (3’-O-N$_3$-dCTP-N$_3$-Bodipy-Fl-510). TCEP is able to cleave both the azido linker to remove the fluorophore and the 3’-azidomethyl capping moiety to restore the 3’-OH group.

5.3.2. Polymerase extension using cleavable fluorescent nucleotide reversible terminators in solution and their characterization by MALDI-TOF MS.

It was of foremost importance to establish that the set of four CF-NRTs, 3'-O-N_3-dNTP-N_3-fluorophores (3'-O-N_3-dATP-N_3-ROX, 3'-O-N_3-dGTP-N_3-Cy5, 3'-O-N_3-dCTP-N_3-Bodipy-FL-510, and 3'-O-N_3-dUTP-N_3-R6G) could be recognized by either natural or mutant DNA polymerase as suitable substrates during polymerase reactions. Hence single base extension reactions of four different linear DNA templates with subsequent complementary bases of AA, GG, CC, or TT were carried out in solution. After the reaction, the 4 different primer extension products were analyzed using MALDI-TOF MS. The MS spectra are shown in Fig. 5.7. As shown, each nucleotide was successfully incorporated into its corresponding primer, leaving none of the starting primers unextended. The single mass peaks at 7369, 7552, 6163 and 6353 (*m/z,* A, G, C, and U respectively) indicated that the single base extension reactions were complete (Fig. 5.7 A, C, E, G). The single mass peak was also a solid confirmation that the 3'-azidomethyl group had successfully terminated DNA synthesis since no 2nd incorporation peaks were observed. To demonstrate that both the fluorophores and the 3' blocking group could be removed simultaneously, cleavage reactions were carried out by incubating the single base extension products in an aqueous TCEP solution (pH9) at 65°C for 15 minutes. As MALDI MS spectra in Fig. 5.7 (B, D, F, H) show, the mass peaks representing the DNA extension products have completely disappeared while single peaks corresponding to the cleavage products appear at 6554, 6609, 5603, and 5612 (*m/z*) respectively. Thus it may be concluded that the CF-NRTs can be successfully incorporated during the DNA polymerase reaction and efficiently cleaved at both the 3' blocking and the fluorophore linkage positions. Therefore, this set of CF-NRTs meets the key requirements necessary for carrying out SBS.

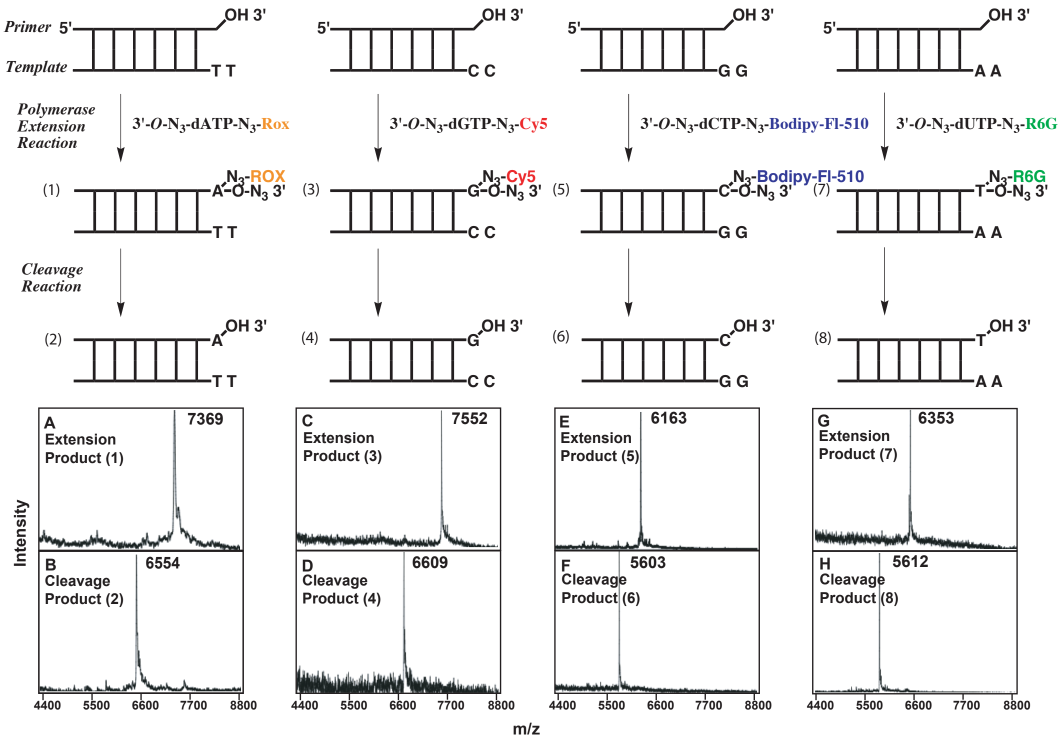


Fig. 5.7. A polymerase reaction scheme (top) to yield DNA extension products by incorporating each of the four 3'-O-N_3-dNTP-N_3-fluorophores and the subsequent cleavage reaction to remove the fluorophore from the DNA extension product. MALDI-TOF MS spectra (bottom) showing efficient base specific incorporation of the 3'-O-N_3-dNTP-N_3-fluorophores and the subsequent cleavage of the fluorophore from the DNA extension products: (A) Primer extended with 3'-O-N_3-dATP-N_3-ROX (1) (peak at 7,369 *m/z*), (B) its cleavage product 2 (6,554 *m/z*); (C) Primer extended with 3'-O-N_3-dGTP-N_3-Bodipy-FL-510 (3) (peak at 7,552 *m/z*), (D) its cleavage product (4) (6,609 *m/z*); (E) Primer extended with 3'-O-N_3-dCTP-N_3-Cy5 (5) (peak at 6,131 *m/z*), (F) its cleavage product (6) (5,603 *m/z*); (G) Primer extended with 3'-O-N_3-dUTP-N_3-R6G (7) (peak at 6,353 *m/z*) and (H) its cleavage product (8) (5,612 *m/z*).

5.3.3. Four-color DNA sequencing by synthesis on a chip using cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs

We have synthesized four CF-NRTs (3'-O-N_3-dNTP-N_3-fluorophores) along with four NRTs (3'-O-N_3-dNTPs) for the implementation of our four-color *de novo*

DNA sequencing by synthesis approach.[18] During the incorporation stage of SBS, a mixture of both sets of nucleotide reversible terminators was used to extend the DNA strand. Only a small percentage of the 3'-O-N_3-dNTP-N_3-fluorophores was present in the mixture so that the majority of the DNA product was extended with the less bulky 3'-O-N_3-dNTPs. This approach led to highly efficient DNA polymerase reaction since the smaller 3'-O-N_3-dNTPs were much more readily incorporated compared to the bulky 3'-O-N_3-dNTP-N_3-fluorophores. In addition, since most of the DNA was extended with 3'-O-N_3-dNTPs, nascent strands of DNA without traces of modification were restored after cleavage of the 3'-azidomethyl capping group on the product. Such DNA would not pose any hindrance for DNA polymerase during the subsequent incorporation step. For DNA extended with the 3'-O-N_3-dNTP-N_3-fluorophores, which serve as the signal producer, the 3'-OH was also restored after the cleavage step so that the next stage of SBS could be carried out. Therefore, theoretically we could utilize nearly all the DNA templates after each round of sequencing, dramatically increasing the potential read length of our SBS methodology.

After fluorescent signal detection, two consecutive capping steps, first with 3'-O-N_3-dNTPs and then with ddNTPs, were performed. The first capping reaction was intended to react with any unextended DNA strands left during the initial incorporation reaction, thus maximizing the amount of extension products and minimizing the loss of templates. In case there were still some unextended products after the first capping step, the second capping with ddNTPs would permanently terminate these DNA strands so that all templates were synchronized. Without these precautionary synchronization steps, mixed fluorescent signals could prevent the identification of the correct nucleotide incorporated due to dephasing. Though the ddNTP capping step would eliminate a few templates from further rounds of sequencing, it was felt that this minor issue was offset by the resulting reduction in lagging strands leading to dephasing. Upon the completion of the dual capping reactions, all the templates were treated with TCEP to simultaneously cleave the fluorophores off DNA products extended with 3'-O-N_3-dNTP-N_3-fluorophores and the 3'-azidomethyl capping moiety of DNA products extended with both 3'-O-N_3-dNTPs and 3'-O-N_3-dNTP-N_3-fluorophores. Since the 3'-OH group was restored on all DNA products after cleavage reaction, the next cycle of sequencing could be carried out to determine the identity of the subsequent base. These steps of incorporation, capping, and cleavage were repeated until all bases on the templates were determined.

The four-color images from a fluorescence scanner for each step of the SBS of a single template on a chip are shown in Fig. 5.8. The first extension of the primer by the complementary cleavable fluorescent nucleotide reversible terminator, 3'-O-N_3-dCTP-N_3-Bodipy-FL-510, was confirmed by observing a blue signal (the emission from Bodipy-FL-510) for all the templates [Fig. 5.8B (1)]. After detection of the fluorescent signal, two separate capping steps, with 3'-O-N_3-dNTPs and then with ddNTPs respectively, were carried out. The DNA chip was then immersed in a TCEP solution to cleave both the fluorophore and the 3'-azidomethyl blocking group from the DNA product extended with 3'-O-N_3-dCTP-N_3-Bodipy-FL-510 and the 3'-azidomethyl blocking group from DNA strands extended with 3'-O-N_3-dCTP. The surface of the chip was washed afterwards, and a negligible residual fluorescent signal was detected, confirming the successful cleavage of the fluorophore [Fig. 5.8B (2)]. This was followed by another extension reaction using the 3’-O-N_3-dNTP-N_3-fluorophores/3’-O-N_3-dNTPs solution to incorporate the next nucleotide complementary to the subsequent base on the template. The entire process of incorporation, detection, synchronization, and cleavage was performed multiple times to identify 30 successive bases in all three DNA templates (an example of one template: 90mer3 is shown in Fig. 5.8). The plot of the raw fluorescence intensity vs. the progress of sequencing extension is shown in Fig. 5.8C. The DNA sequences are identified unambiguously from the 4-color raw fluorescence data without any processing.[18]

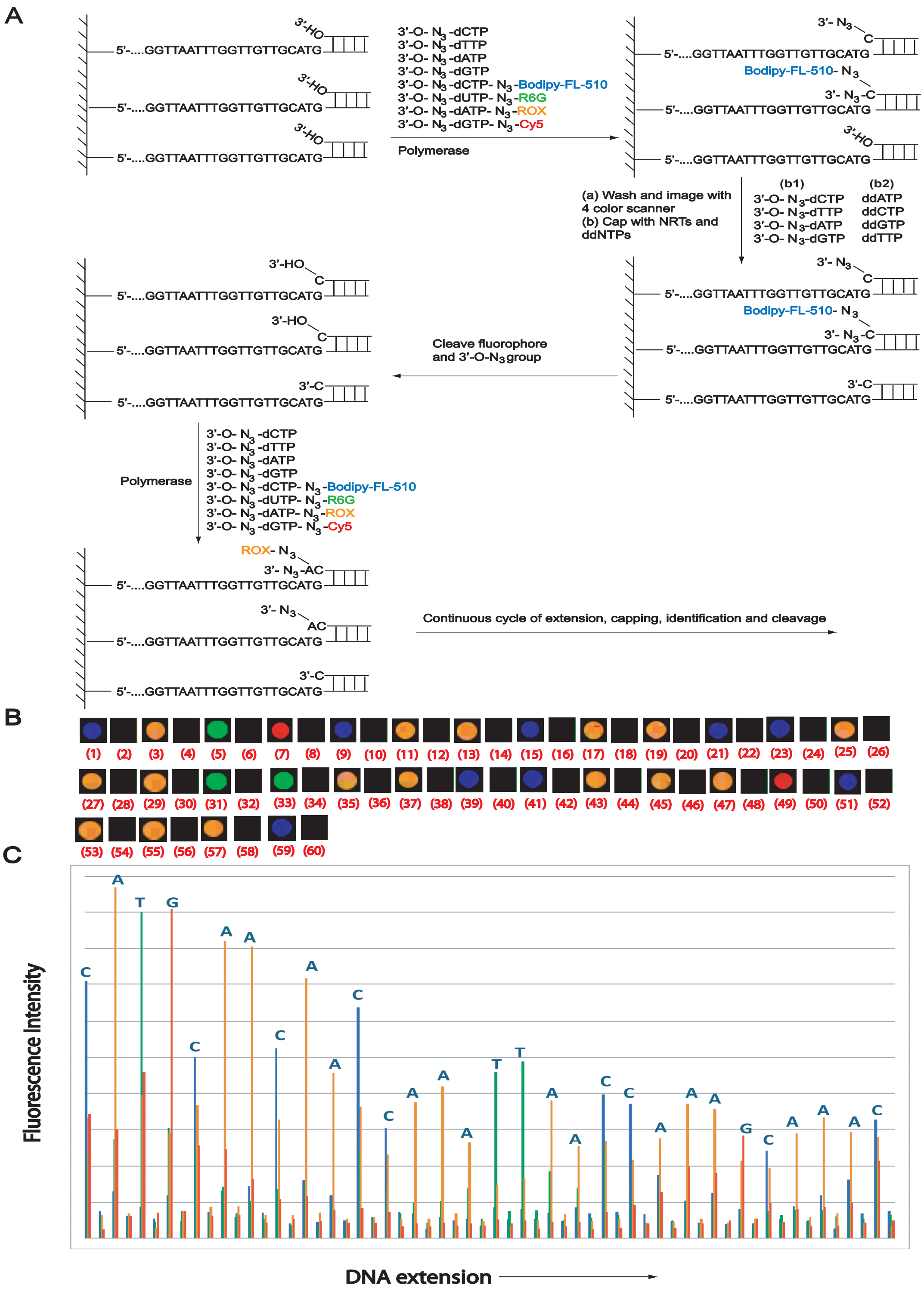


Fig. 5.8. Four-color DNA SBS with 3'-O-N_3-dNTP-N_3-fluorophores (90mer3). (A) A SBS with CFNRTs scheme for four-color sequencing on a chip by using four 3'-O-N_3-dNTP-N_3-fluorophores and 3'-O-N_3-dNTPs with ddNTPs capping. (B) Four-color fluorescence images for each step of the SBS: (1) incorporation of 3'-O-N_3-dCTP-N_3-Bodipy-Fl-510 and 3'-O-N_3-dCTP;

(2) cleavage of N_3-Bodipy-Fl-510 and 3’-CH_2N_3 group; (3) incorporation of 3’-O-N_3-dATP-N_3-Rox and 3’-O-N_3-dATP; (4) cleavage of N_3-Rox and 3’-CH_2N_3 group; images 5-60 were produced similarly. (C) A plot (four-color sequencing data) of raw fluorescence emission intensity obtained by using 3’-O-N_3-dNTP-N_3-fluorophores and 3’-O-N_3-dNTPs. The small groups of peaks between the identified bases are fluorescent background from the DNA chip.

5.4 Materials and Methods

General Information. All solvents and reagents were reagent grade, purchased commercially and used without further purification. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Oligonucleotides used as primers or templates were synthesized on an Expedite nucleic acid synthesizer (Applied BioSystems) or purchased from Midland. 9^oN polymerase (exo-) A485L/Y409V was obtained from New England Biolabs. Phosphoramidite reagents and columns for oligonucleotide synthesis were purchased from Glen Research (Sterling, VA). Both the nucleotide reversible terminators and cleavable fluorescent nucleotide reversible terminators were purified by reverse-phase HPLC on a 150×4.6 mm C18 column (Supelco), mobile phase: A, 8.6 mM Et_3N / 100 mM 1,1,1,3,3,3-hexafluoro-2-propanol in water (pH 8.1); B, methanol. Elution was performed from 100% A isocratic over 10 minutes followed by a linear gradient of 0-50% B for 20 minutes and then 50% B isocratic over another 30 minutes.

5.4.1. Design and synthesis of cleavable fluorescent nucleotide reversible terminators and 3’-O-modified NRTs for SBS

The synthesis of the complete set of CF-NRTs, 3’-O-N_3-dNTP-N_3-fluorophores, was accomplished by Dr. Zengmin Li, Dr. Shenglong Zhang, Dr. Huanyan Cao, Dr. Mong Sang Mama, and Dr. Qinglin Meng of the organic synthetic chemistry team at our group. The detailed synthesis procedures for 3’-O-N_3-dNTP-N_3-fluorophores are described in Dr. Shenglong Zhang’s thesis.

5.4.2. Polymerase extension using cleavable fluorescent nucleotide reversible terminators in solution and their characterization by MALDI-TOF MS.

We characterized the four cleavable fluorescent nucleotide reversible terminators, 3’-O-N_3-dNTP-N_3-fluorophores (3’-O-N_3-dATP-N_3-Rox, 3’-O-N_3-dGTP-N_3-Cy5, 3’-O-N_3-dCTP-N_3-Bodipy-FL-510, and 3’-O-N_3-dUTP-N_3-R6G) by performing four separate DNA single base extension reactions. The resulting DNA extension products were analyzed by MALDI-TOF MS. The following four sets of DNA primer and template were used for the extension:

CF-NRTs	**Template**	**Primer**
3’-*O*-N_3-dATP-N_3-Rox	Exon 8*	6084: 5’-TAGATGACCCTGCCTTGTCG-3’
3’-*O*-N_3-dGTP-N_3-Cy5	Exon 7^	6131: 5’-GTTGATGTACACATTGTCAA-3’
3’-*O*-N_3-dCTP-N_3-Bodipy-FL-510	Exon 8*	5144: 5’-TCTCTGGCCGCGTGTCT-3’
3’-*O*-N_3-dUTP-N_3-R6G	Exon 8*	5163: 5’-GATAGGACTCATCACCA-3’

**Exon 7 complementary sequence:*
3’CCCGGACACAATAGAGGATCCAACCGAGACTGACATGGTGGTAGGTGATGTTGATGTA
CACATTGTCAAGGACGTACCCGCCGTACTTGGCCTCCGGGTA-5’
^Exon 8 complementary sequence:
3’ACGGAGAACGAAGAGAAAAGGATAGGACTCATCACCATTAGATGACCCTGCCTTGTCG
AACTCCACGCACAAACACGGACAGGACCCTCTCTGGCCGCGTGTCTCCTTC-5’

Each of the extension reactions consisted of 100 pmol of 3’-O-N_3-dNTP-N_3-fluorophores along with 20 pmol of the DNA template, 40 pmol of the primer, 1X Thermopol II reaction buffer, 40 nmol of $MnCl_2$ and 1 unit of 9^{o}N DNA polymerase (exo-) A485L/Y409V in a total reaction volume of 20 μL. The reaction was carried out at 94^{o}C for 5 minutes, 4^{o}C for 5 minutes, and 65^{o}C for 20 minutes. Subsequently, the extension product was purified by reverse-phase HPLC using established procedures.[5] The fraction containing the desired product was collected and freeze-dried for analysis by MALDI-TOF MS and cleavage. For cleavage of the DNA extension products bearing the 3’-O-N_3-dNTP-N_3-fluorophores, they were resuspended in 40 μL of 100 mM TCEP solution (pH 9.0) at 65^{o}C for 15 minutes and then analyzed by MALDI-TOF MS.

5.4.3. Construction of a DNA Immobilized Chip of Multiple Linear Templates

Four different 5’-amino-labeled linear DNA templates with the following sequences were purchased from IDT (Coralville, IA).

90mer1: 5’-CCT TTA ATT TTG GCT TTT AAT TGG CTT GCT TTG GTT AAC TTG GTT GTT GCA TG***C CCA TGC GAG TGC GAG TGC ACG TGG CGC AGC AGG TCA***-3’

90mer2: 5’-CCT TTT TGG TTA ACT TTA ATT GGC TTG GCT TTG GTT CAA TTG GTT GTT ACA TG***C CCA TGC GAG TGC GAG TGC ACG TGG CGC AGC AGG TCA***-3’

90mer3: 5’-CCT TTG GTT TTG GCT TTT GGT TGG TTT GCT TTG GTT AAT TTG GTT GTT GCA TG***C CCA TGC GAG TGC GAG TGC ACG TGG CGC AGC AGG TCA***-3’

90mer4: 5’-CCT TTT TGG TTG GCT TTA ATT GGT TTG GCT TTG GTT TAA TTG GTT GTT ACA TG***C CCA TGC GAG TGC GAG TGC ACG TGG CGC AGC AGG TCA***-3’

Each template had the 37 bases of consensus sequence (bold, italicized) for the annealing of a complementary 37mer primer (5’-***TGA CCT GCT GCG CCA CGT GCA CTC GCA CTC GCA TGG G***-3’). The DNA templates were dissolved at 40 μM in 50 mM sodium phosphate buffer, pH 8.5 and spotted using a SpotArray 72 arraying robot (Perkin Elmer) onto high density CodeLink microarray slides (GE Healthcare). After spotting, the slides were incubated at ambient temperature (~ 24°C) for 20 hours in a humid chamber containing saturated sodium chloride solution to allow for 5’-tethering of the spotted amino-modified DNA templates to the slide surface, which was functionalized with succinimide ester groups. After the incubation the slides were removed from the humid chamber and stored in a vacuum desiccator at room temperature. Prior to carrying out SBS, the slides were taken out of storage for the primer annealing process. For a single spot on the chip that contained all four templates, 10 μL of 1X annealing buffer (Thermo Pol buffer) was used for a 10 minute incubation period at 65°C. The buffer was then replaced with an annealing mixture (10 μL: 1X Thermo Pol buffer, 0.5M NaCl and 3.5μM 37mer primer, denatured at 94°C for 1min, then cooled down in ice) for 30 minutes of incubation at 65°C. The annealing process was repeated twice more. The DNA chip was then washed with the annealing buffer 6 times, and ready for SBS.

5.4.4. Four-color DNA sequencing by synthesis on a chip using cleavable fluorescent nucleotide reversible terminators and 3'-O-modified NRTs

The DNA chip was constructed by immobilizing various 5'-amino-modified single strand oligonucleotides (90mers) on a CodeLink microarray slide (GE Healthcare). All linear templates had 37 bases of identical sequence at the 3' end so that a universal primer was used for annealing at 37°C for 30 minutes. Ten microliters of extension solution consisting of 3'-O-N_3-dATP-N_3-ROX (4 pmol), 3'-O-N_3-dGTP-N_3-Cy5 (1 pmol), 3'-O-N_3-dCTP-N_3-Bodipy-FL-510 (5 pmol), 3'-O-N_3-dUTP-N_3-R6G (3 pmol), 3'-O-N_3-dATP (45 pmol), 3'-O-N_3-dGTP (48 pmol), 3'-O-N_3-dCTP (42 pmol), 3'-O-N_3-dTTP (45 pmol), 1 unit of 9°N DNA polymerase(exo-) A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer was spotted on the surface of the chip, at sites where different linear DNA templates with their primers already annealed were immobilized. This extension mixture was incubated at 55°C for 15 minutes. Upon the completion of the extension reaction, the chip was washed with SPSC buffer (0.1M sodium phosphate/0.5M NaCl, pH 7.5, 0.1% Tween-20) for 2 minutes, followed by a brief rinse with dH_2O. The blow-dried chip was scanned with a 4-color ScanArray Express scanner (Perkin–Elmer Life Sciences) to detect the fluorescence signal. The four lasers on the 4-color scanner had excitation wavelengths of 488, 543, 594, and 633 nm and emission filters centered at 522, 570, 614, and 670 nm. After signal detection, a capping reaction solution, consisting of 45 pmol each of 3'-O-N_3-dATP, 3'-O-N_3-dGTP, 3'-O-N_3-dCTP, and 3'-O-N_3-dTTP, 1 unit of 9^{o}N DNA polymerase(exo-) A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer, was spotted on the same position on the DNA chip and incubated at 55^{o}C for 20 minutes. A second capping solution, a mixture of four dideoxynucleotide triphosphates (100 pmol each of ddATP, ddGTP, ddCTP, and ddTTP, Fluka Analytical), 1 unit of 9^{o}N DNA polymerase(exo-) A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer, replaced the first capping solution on the DNA spot for another incubation at 55^{o}C for 5 minutes. For the cleavage process immediately followed the dual capping steps, about 10 μl of 100 mM TCEP (pH 9.0) was spotted on the same area of the DNA chip for 20 minutes of incubation at 55^{o}C. After washing the surface with SPSC buffer, the chip was scanned again to compare the intensity of fluorescence after cleavage with the original fluorescence intensity. This process was repeated by carrying out the next round of extension using the 3'-O-N_3-dNTP-N_3-fluorophore/3'-O-N_3-dNTP solution with subsequent synchronization, washing, fluorescence detection, and cleavage processes performed as described above.

5.5 Conclusion

A complete set of cleavable fluorescent nucleotide reversible terminators (CF-NRTs: 3'-O-N_3-dNTP-N_3-fluorophores) along with four non-fluorescent nucleotide reversible terminators (NRTs: 3'-O-N_3-dNTPs) have been synthesized for the implementation of our four-color *de novo* DNA sequencing by synthesis in order to seek improvements over existing SBS approaches. During the incorporation stage of SBS, a mixture of both sets of modified nucleotide analogues is used to simultaneously extend the primer strand of various target DNA linear templates. Only a small percentage of the 3'-O-N_3-dNTP-N_3-fluorophores is used in the mixture so that the majority of the product is extended with the less bulky 3'-O-N_3-dNTPs. This approach leads to a more efficient DNA polymerase reaction since the smaller 3'-O-N_3-dNTPs are much easier to incorporate. Another advantage of having most of the DNA extended with 3'-O-N_3-dNTPs is the fact that after cleavage of the 3'-OH capping group on the products, nascent strands of DNA that have no traces of modification are restored. Such DNA strands do not have any adverse effect on the DNA polymerase during the incorporation of the next nucleotide. For DNA extended with the 3'-O-N_3-dNTP-N_3-fluorophores, which serve as the signal producer, the 3'-OH is also restored after the cleavage step so that the next stage of SBS can be carried out. Therefore, it is possible to recover nearly all the DNA templates after each round of sequencing, dramatically increasing the potential read length of our SBS methodology. After the incorporation reaction, two separate capping steps, first with 3'-O-N_3-dNTPs and then with ddNTPs, are performed. The rationale behind the first capping reaction is to maximize the amount of extension products and to ensure the minimal loss of templates. In case there is still unextended product after the first capping step, the second capping with ddNTPs will permanently terminate these DNA strands, preventing them from participating in further incorporation reactions.

Therefore essentially all DNA templates are synchronized at the end of the capping stage. Without these precautionary synchronization procedures, mixed fluorescent signals could prevent the identification of the correctly incorporated nucleotide. Hence, we have successfully addressed one of the key problems that has been hindering the progress of SBS, which is the mis-calling of the base due to lagging signals. In addition, since both 3’-O-N_3-dNTP-N_3-fluorophores and 3’-O-N_3-dNTPs are reversible terminators, which allow the sequencing of each base in a serial manner, they can accurately decode the homopolymeric regions of DNA.

It is worth noting that since all of the steps of our SBS approach are performed on a DNA array chip, there is no longer a need for electrophoretic DNA fragment separation as in the classical Sanger sequencing method. Another advantage of using the array chip lies in the deposition of multiple DNA templates on the surface at separate registered sites, allowing parallel sequencing to take place (Fig. 5.9).

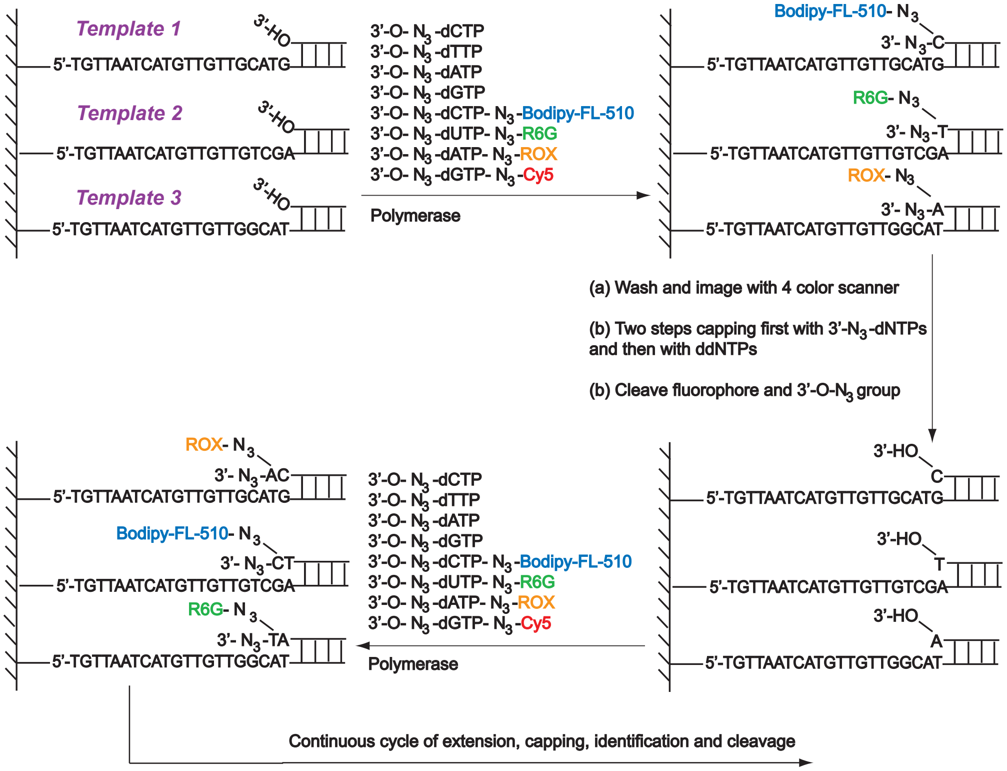


Fig. 5.9. Scheme for parallel sequencing of multiple templates by four-color SBS

As mentioned in the previous sections, we had 3 different linear templates on the chip. By performing a single round of SBS, the current base on each template could be identified simultaneously. As indicated in the sequencing data (Fig. 5.10), by simply comparing the fluorescent intensity emitted from incorporated CF-NRTs (raw data) without further processing, we were able to distinguish sequence differences in DNA, calling the correct base for each template. This is further proof that our SBS approach of combining CF-NRTs and NRTs can perform parallel sequencing, and has the potential to develop into an ultra high-throughput sequencing platform. We purposely immobilized linear templates, which resemble single strand DNAs from a genomic library preparation, instead of self-priming ones, to demonstrate the flexibility of our SBS technology and pave the way for future biological applications. With an already long raw sequencing read around 30 bases, SBS with CF-NRTs and NRTs can achieve even higher read length by applying advanced algorithm software that filters data through minimizing phasing effects. Many discoveries in genome function and regulation have been made recently with ChIP-Seq[6-8] based on sequencing tags of approximately 25 bases and a high-throughput SAGE[19] approach that reaches single copy transcript sensitivity.[20] Therefore, future implementation of SBS with CF-NRTs and NRTs, coupled with a high-density bead array platform based on millions of different PCR templates generated on a solid surface through emulsion PCR or clonal amplification,[8, 21] will provide an ultra-high-throughput and accurate DNA sequencing system with wide applications in genome biology and biomedical research.

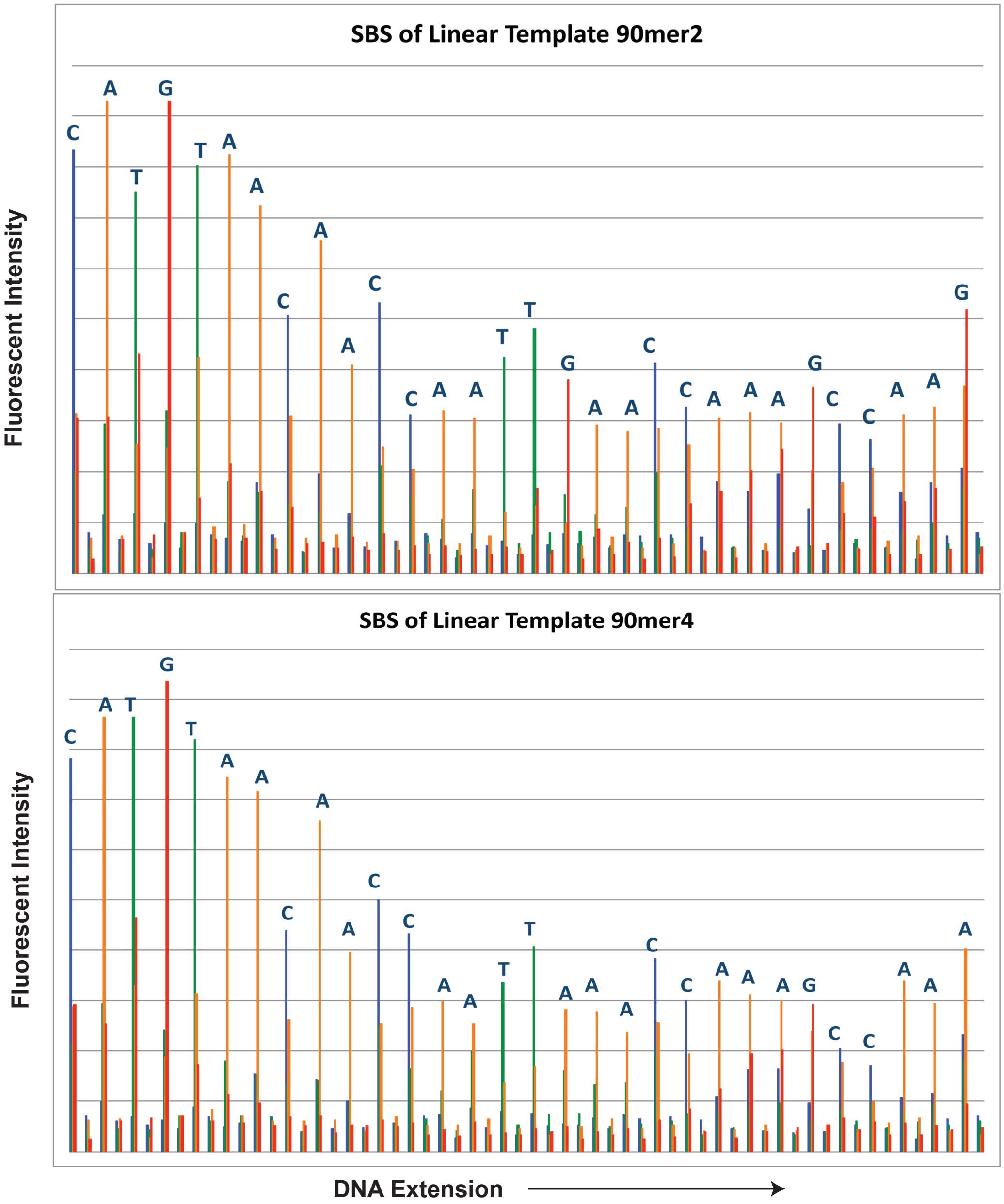


Fig. 5.10. A plot (4-color sequencing data) of raw fluorescence emission intensity of two more templates (90mer2 and 90mer4) sequenced simultaneously with previously presented template (90mer3) using four-color SBS with 3'-O-N_3-dNTPs and 3'-O-N_3-dNTP-N_3-fluorophores.

5.6 References

1. Ju, J.; Li Z.; Edwards, J.; Itagaki, Y. Massive parallel method for decoding DNA and RNA. United States Patent 6,664,079 USA, **2003**.

2. Li, Z.; Bai, X.; Ruparel, H.; Kim, S.; Turro, N. J.; Ju, J. A photocleavable fluorescent nucleotide for DNA sequencing and analysis. *Proc Natl Acad Sci USA*, **2003**, *100*, 414-419.

3. Ruparel, H., *et al.* Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc Natl Acad Sci U S A* **2005**, *102,* 5932-5937.

4. Seo, T. S.; Bai, X.; Kim, D. H.; Meng, Q.; Shi, S.; Ruparel, H.; Li, Z.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci USA,* **2005**, *102*, 5926-5931.

5. Ju, J., *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **2006**, *103,* 19635-19640.

6. Mikkelsen, T. S.; Ku, M.; Jaffe, D. B.; Issac, B.; Lieberman, E.; Giannoukos, G.; Alvarez, P.; Brockman, W.; Kim, T. K.; Koche, R. P. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature,* **2007**, *448*, 553-560.

7. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions**.** *Science,* **2007**, *316*, 1497-1502.

8. Barski, A.; Cuddapah, S.; Cui, K.; Roh, T. Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao,K. High-resolution profiling of histone methylations in the human genome. *Cell,* **2007**, *129*, 823-837.

9. Wu, J., *et al.* 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* **2007**, *104,* 16462-16467.

10. Guo, J.; Xu, N.; Li, Z.; Zhang, S.; Wu, J.; Kim, D. H.; Marma, M. S.; Meng, Q.; Cao, H.; Li, X.; Shi, S.; Yu, L.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* **2008**, *105*, 9145-4150.

11. Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008,** *456,* 53-59.

12. Harris, T. D.; Buzby,P. R.; Babcock, H.; Beer, E.; Bowers, J.; Braslavsky, I.; Casey, M.; Colonell, J.; DiMeo, J.; Efcavitch, J. W.; *et al*. Single-molecule DNA sequencing of a viral genome. *Science,* **2008**, *320*, 106-109.

13. Zavgorodny, S.; Polianski, M.; Besidsky, E.; Kriukov, V.; Sanin, A.; Pokrovskaya, M.; Gurskaya, G.; Lonnberg, H.; Azhayev, A. 1-Alkylthioalkylation of nucleoside hydroxyl functions and its synthetic applications: a new versatile method in nucleoside chemistry. *Tetrahedron Letters,* **1991**, *32*, 7593-7596.

14. Zavgorodny, S.; Pechenov, A. E.; Shvets, V. I.; Miroshnikov, A. I. S,X-acetals in nucleoside chemistry. III. Synthesis of 2'-and 3'-*O*-azidomethyl derivatives of ribonucleosides. *Nucleosides, Nucleotides & Nucleic Acids,* **2000**, *19*, 1977-1991.

15. Saxon, E.; Bertozzi, C. R. Cell surface engineering by a modified Staudinger reaction. *Science,* **2000**, *287*, 2007-2010.

16. Milton, J.; Ruediger, S.; Liu, X. Nucleosides/nucleotides conjugated to labels via cleavable linkages and their use in nucleic acid sequencing. *United States Patent Application US20060160081A1*, **2006**.

17. Rosenblum, B. B.; Lee, L. G.; Spurgeon, S. L.; Khan, S. H.; Menchen, S. M.; Heiner, C. R.; Chen, S. M. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.,* **1997**, *25*, 4500-4504.

18. Yu, L.; Guo, J.; Qui, C.; Li, Z.; Kim, D. H.; Cao, H.; Zhang, S.; Meng, Q.; Marma, M. S.; Wu, J.; Xu, N.; Li, X.; Shi, S.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis (SBS) improvements with cleavable fluorescent nucleotide reversible terminators. **2009**, *submitted, under review.*

19. Velculescu, V. E.; Zhang, L.; Vogelstein, B.; Kinzler, K. W. Serial analysis of gene expression. *Science,* **1995**, *270*, 484-487.

20. Kim, J. B.; Porreca, G. J.; Song, L.; Greenway, S. C.; Gorham, J. M.; Church, G. M.; Seidman, C. E.; Seidman, J. G. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science,* **2007**, *316*, 1481-1484.

21. Margulies, M.; Egholm, M.; Altman, W. E.; Atiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376-380.

Chapter 6: Exploration of Novel Primer Resetting Strategies to Extend Read-length for DNA Sequencing by Synthesis

6.1 Introduction

There is an ever-growing need in the research community for low-cost, high-throughput sequencing approaches. In our laboratory, we have developed a four-color sequencing by synthesis (SBS) system for decoding DNA using a combination of CF-NRTs and NRTs.[1, 2] We are also collaborating with commercial partners to produce new generations of polymerases that are able to incorporate our modified nucleotides at high efficiency and specificity, and developing a series of continually improving cleavable linkers for fluorophore attachment. We are fortunate to have obtained a prototype machine designed around our SBS system for further refining the system. We also can design experiments using our SBS system on commercial platforms which utilize similar SBS design. By incorporating these advances, it is possible to begin conducting pilot studies on real sequencing projects. However, we would like to increase read length to the point where not only tag based sequencing projects, but projects that require continuous stretches of DNA approaching 100 nucleotides, become feasible. Therefore, the successful implementation of SBS is essentially dependent on the read length of the target DNA templates.

One of the major factors that determines the read length of SBS is the number of available templates. The total number of sequenceable templates decreases after each cycle of SBS reaction due to such causes as DNA strands being extended with permanent terminators (ddNTPs) and loss of primers due to vigorous washing. Various means can be employed to minimize this rate of template reduction. A powerful method termed “primer resetting” can potentially diminish the negative effect of template loss and increase the read length of SBS several-fold.[3] The fundamental rationale behind primer resetting is to remove the already-extended “sequenced” primer and attach a new primer to continue the sequencing. In general, three steps are involved with this approach: 1) annealing of the first primer, 2) performing SBS, 3) denaturing the sequenced section of the template to recover a single-stranded DNA for the second primer annealing. These steps are carried out repeatedly until the target DNA is sequenced in its entirety. The advantage of primers resetting lies in its ability to restore all the templates after the denaturation step, including those that are terminated with ddNTPs, so the next cycle of SBS can restart with potentially the same amount of sequenceable DNA as the previous round.

6.2 Experimental Rationale and Overview

We propose three general approaches for our template resetting strategy to achieve long read lengths for SBS. In the first approach, the DNA sequencing primer is reset by replacing the sequenced strand with the original primer, extending this primer with natural or minimally modified nucleotides to approximately the end of the first round sequence, and then sequencing from that point. The second strategy relies on annealing of a second round primer that is longer than the first, containing at its 5’ end the same sequence as the original primer, followed by a run of 20 universal nucleotides such as inosines, from which the second round of sequencing can be primed. If the duplex stability of this highly degenerate primer with DNA templates is found to be low, a number of locked nucleotides can be added at either end of the primer to increase the stability of the primer-template complex. In the third strategy, extra priming sites are inserted within a template strand via Type IIS or Type III restriction-recircularization. Each of these approaches has distinct advantages and some difficulties that need to be overcome. None of the three aforementioned strategies are sensitive to the type of library (genomic, cDNA or other), the method of amplification prior to sequencing (spotting of clones, ePCR, polony PCR), or the mode of sequencing. Hence they are all platform agnostic, thus greatly increasing their range of applications in sequencing technologies.

6.2.1. Strategy 1: template "walking" by unlabeled nucleotides

The fundamental rationale behind this template "walking" strategy is the removal of the sequenced strand and reattachment of the original primer to allow the extension, or "walking", of the template with a combination of natural and modified nucleotides to approximately the end of the first round sequence so that SBS can be carried on from that point. Since the original sequenced strand is stripped away, including those terminated with ddNTPs, all the templates become available for "walking". Given that "walking" is carried out with either natural or 3'-modifed nucleotides, the subsequent round of SBS is performed on nascent DNA strands for maximum read length. In general, three steps are involved with this "walking" strategy: 1) annealing of the primer and performing SBS, 2) denaturing of the sequenced strand, reattaching the original primer, and "walking" of the template with natural or modified nucleotides over the sequenced region, and 3) repeating SBS from the new starting point. These steps are repeated until the target DNA is sequenced in its entirety (Fig. 6.1).

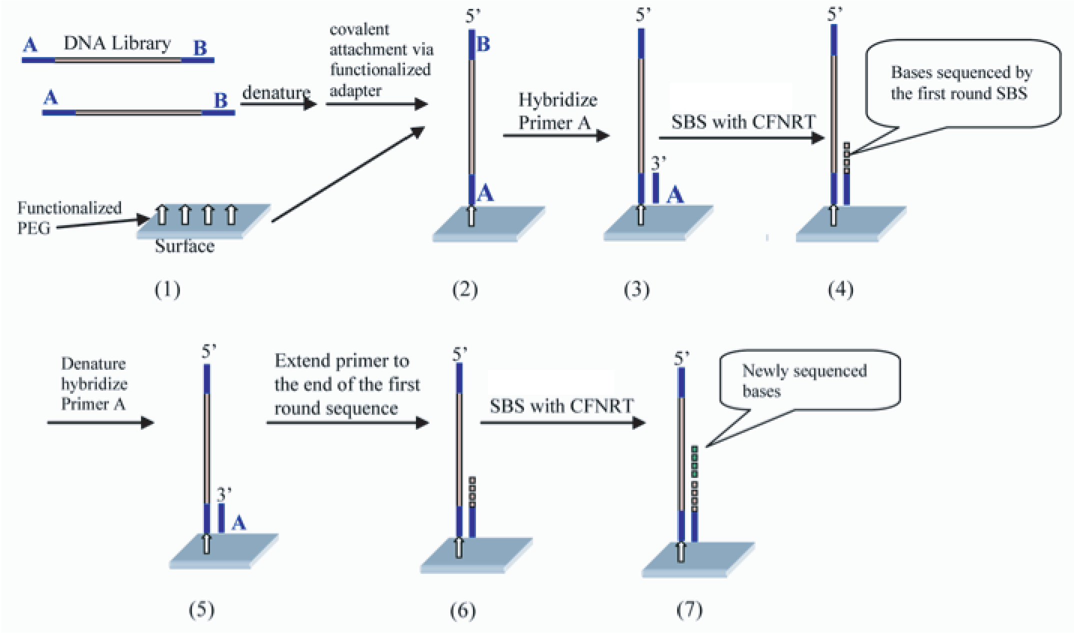

A DNA Library B
A B
denature
covalent attachment via functionalized adapter
5'
B
A
Functionalized PEG
Surface
Hybridize Primer A
3'
A
SBS with CFNRT
Bases sequenced by the first round SBS
(1)
(2)
(3)
(4)
Denature hybridize Primer A
5'
3'
A
Extend primer to the end of the first round sequence
SBS with CFNRT
Newly sequenced bases
(5)
(6)
(7)

Fig. 6.1. Scheme for primer resetting strategy 1: template walking

Even though theoretically SBS with CF-NRTs, as described in the previous chapters, can be executed without losing templates, the utilization of ddNTP capping does reduce the number of available templates during the actual sequencing reaction. In addition, DNA strands incorporated with CF-NRTs leave a tail on the modified nucleotides after cleavage that can potentially reduce the incorporation efficiency of the subsequent base. The advantage of template "walking" is its ability to restore all the templates after the denaturing step so that the next cycle of SBS can restart with potentially the same amount of nascent DNA as the previous round. The "walking" methodology is applicable to our four-color SBS using CF-NRTs, and has the potential to dramatically increase the read length of the SBS platform.

In order to implement the template "walking" strategy in our SBS platform, we first carry out a round of SBS using CF-NRTs (as described in Chapter 5) to its maximum read length.[4] Immediately after the first round of SBS, DNA templates are denatured by heat or mild alkali conditions to remove the extended primer. The initial primer is annealed to the template again and enzymatic incorporations of either natural or modified nucleotides are conducted to fill the gap between the first and second stages of SBS. Three methods are proposed to achieve the "walking". Each approach has its advantages and shortcomings, which are summarized in the following.

Method 1. Nucleotide reversible terminators (3'-*O*-R_1-dNTPs) are used as substrates to perform enzymatic incorporation (Fig. 6.2) to extend the primer. Similar to regular SBS with CF-NRTs, after incorporation, a specific chemical reaction is applied to regenerate the 3'-OH to ensure the subsequent incorporation. The number of repeated cycles of such incorporation and cleavage will exactly match the actual read length in the first stage of SBS, so that this "filling gap" incorporation stops at the same point that the first round of SBS attained.

Method 2. Enzymatic incorporation is conducted using two sets of nucleotides as substrates. For example, the first set of natural nucleotides composed of dCTP, dATP, and dGTP (*sans* dTTP) was used to perform incorporation, so that the polymerase reaction stops once it reaches a base "A" in the template. Then enzymatic incorporation is resumed with the second set of nucleotides composed of dTTP, dATP, and dGTP (*sans* dCTP), resulting in a polymerase reaction that stops at the base "G" in the template. The repeated cycles of such incorporations fill the gap between the first and second stages of SBS, but in the case of unknown templates, cannot be assured of reaching the precise point of the first round termination (Fig. 6.3).

Method 3. Enzymatic incorporation is conducted using three natural dNTPs (e.g. dATP, dCTP, and dTTP) and one NRT (3'-O-R_1-dGTP) as substrates (Fig. 6.4). Primer elongation will only be stopped once it incorporates a NRT. After incorporation, a specific chemical reaction is applied to regenerate the 3'-OH which ensures consecutive incorporation in the next round. Repeated cycles of such incorporation and cleavage will fill the gap between the first and second stages of SBS, again with the provision mentioned for the second method.

At the conclusion of the "walking" process, the primer is extended to approximately the end of the previous round of SBS. The second round of SBS with CF-NRTs is carried out to identify the subsequent bases. If the process is repeated more times, it is theoretically possible to achieve substantially long read length of more than 200 bases.

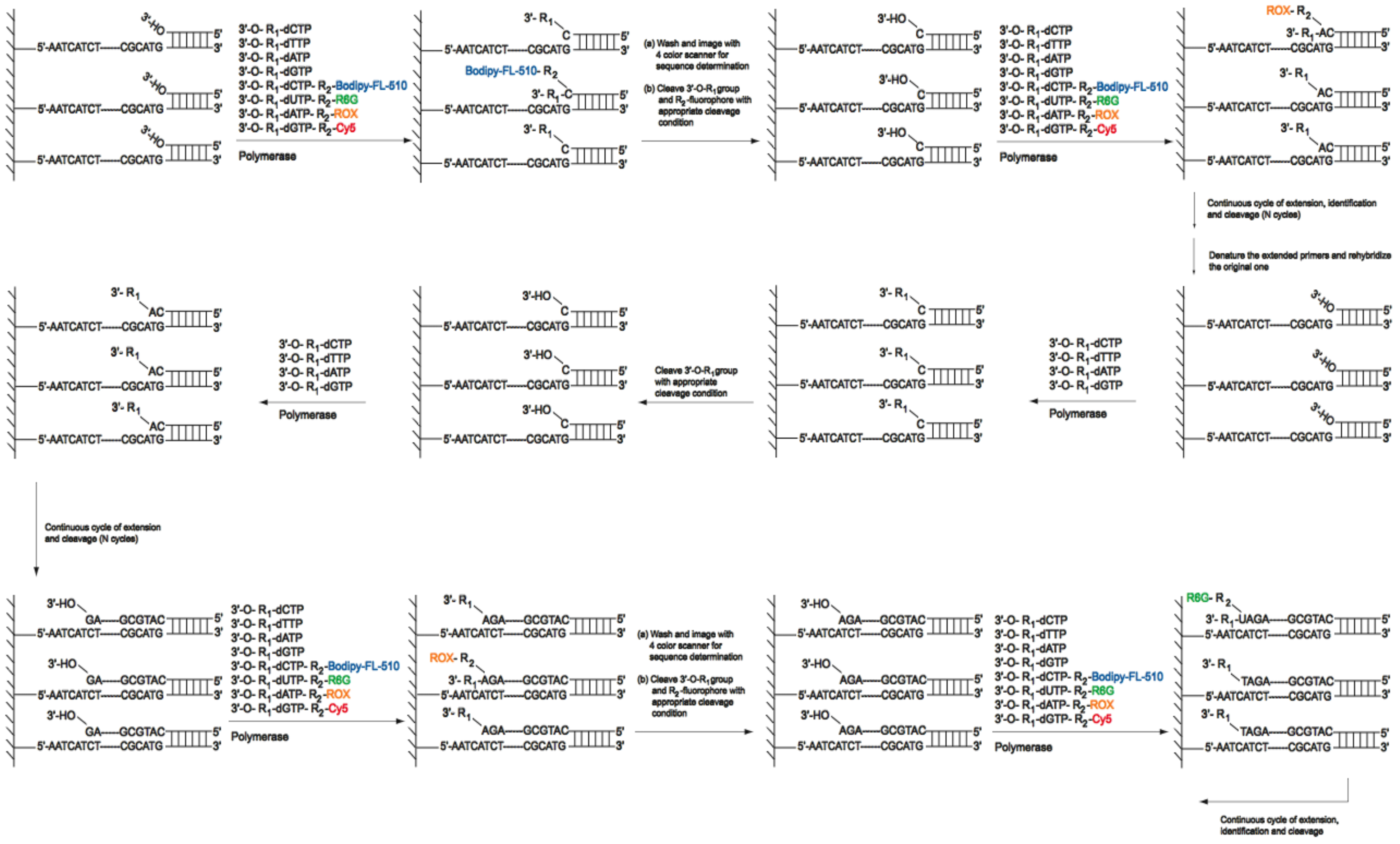

5'-AATCATCT------CGCATG
3'-HO
5'
3'
3'-O- R$_1$-dCTP
3'-O- R$_1$-dTTP
3'-O- R$_1$-dATP
3'-O- R$_1$-dGTP
3'-O- R$_1$-dCTP- R$_2$-Bodipy-FL-510
3'-O- R$_1$-dUTP- R$_2$-R6G
3'-O- R$_1$-dATP- R$_2$-ROX
3'-O- R$_1$-dGTP- R$_2$-Cy5
Polymerase
3'- R$_1$
C
Bodipy-FL-510- R$_2$
(a) Wash and image with 4 color scanner for sequence determination
(b) Cleave 3'-O-R$_1$ group and R$_2$-fluorophore with appropriate cleavage condition
ROX- R$_2$
3'- R$_1$-AC
AC
Continuous cycle of extension, identification and cleavage (N cycles)
Denature the extended primers and rehybridize the original one
Cleave 3'-O-R$_1$ group with appropriate cleavage condition
Continuous cycle of extension and cleavage (N cycles)
GA------GCGTAC
AGA------GCGTAC
3'- R$_1$-AGA------GCGTAC
R6G- R$_2$
3'- R$_1$-UAGA------GCGTAC
TAGA------GCGTAC
Continuous cycle of extension, identification and cleavage

Fig. 6.2. Template “walking” method 1 for SBS with CF-NRTs

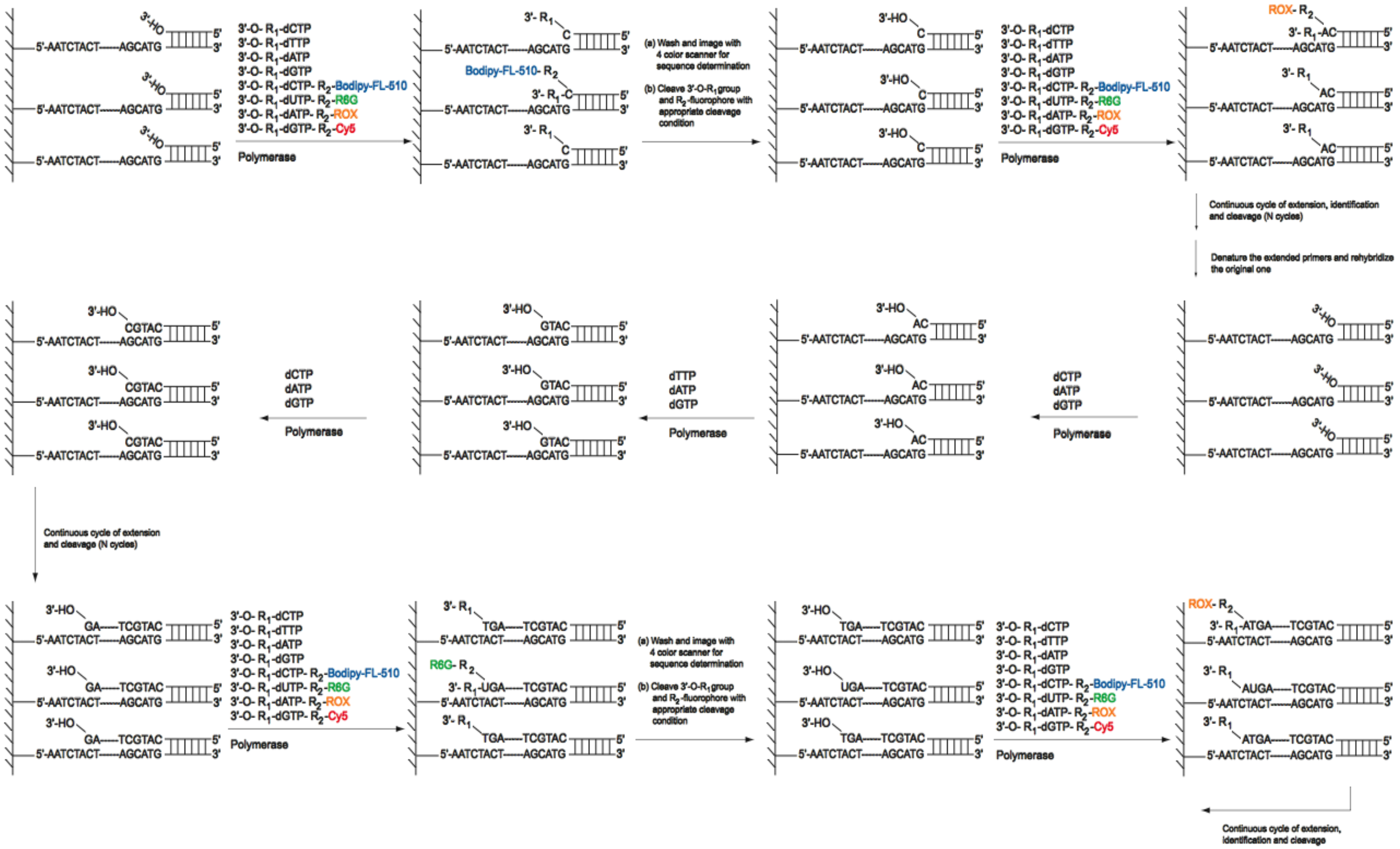

3'-HO
5'-AATCTACT------AGCATG
5'
3'
3'-O- R1-dCTP
3'-O- R1-dTTP
3'-O- R1-dATP
3'-O- R1-dGTP
3'-O- R1-dCTP- R2-Bodipy-FL-510
3'-O- R1-dUTP- R2-R6G
3'-O- R1-dATP- R2-ROX
3'-O- R1-dGTP- R2-Cy5
Polymerase
3'- R1
C
Bodipy-FL-510- R2
3'- R1-C
(a) Wash and image with 4 color scanner for sequence determination
(b) Cleave 3'-O-R1 group and R2-fluorophore with appropriate cleavage condition
ROX- R2
3'- R1-AC
AC
Continuous cycle of extension, identification and cleavage (N cycles)
Denature the extended primers and rehybridize the original one
dCTP
dATP
dGTP
dTTP
GTAC
CGTAC
Continuous cycle of extension and cleavage (N cycles)
GA------TCGTAC
TGA------TCGTAC
R6G- R2
3'- R1-UGA------TCGTAC
UGA------TCGTAC
3'- R1-ATGA------TCGTAC
AUGA------TCGTAC
ATGA------TCGTAC
Continuous cycle of extension, identification and cleavage

Fig. 6.3. Template “walking” method 2 for SBS with CF-NRTs

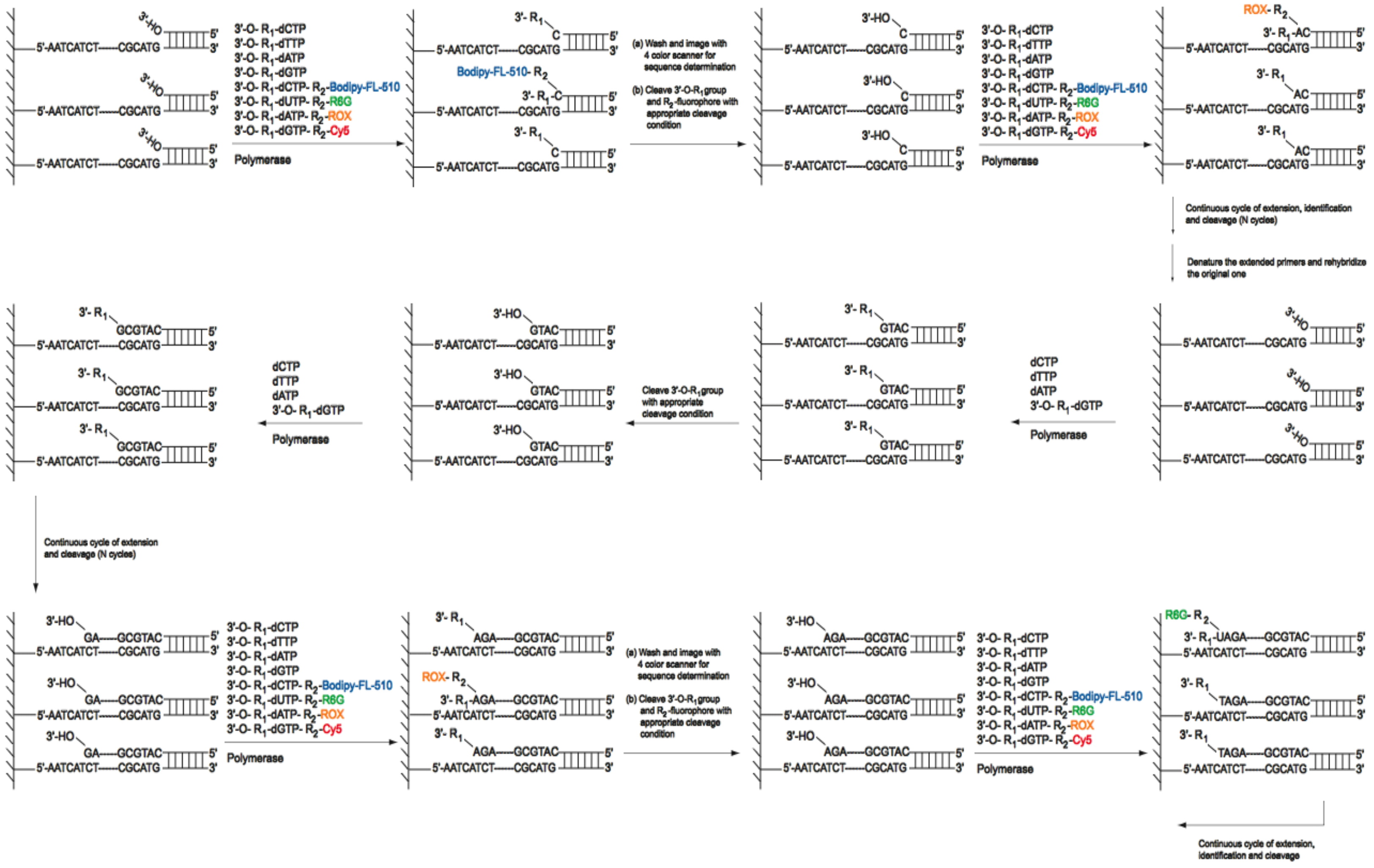

3'-HO
5'-AATCATCT-----CGCATG
5'
3'
3'-O-R$_1$-dCTP
3'-O-R$_1$-dTTP
3'-O-R$_1$-dATP
3'-O-R$_1$-dGTP
3'-O-R$_1$-dCTP-R$_2$-Bodipy-FL-510
3'-O-R$_1$-dUTP-R$_2$-R6G
3'-O-R$_1$-dATP-R$_2$-ROX
3'-O-R$_1$-dGTP-R$_2$-Cy5
Polymerase
3'-R$_1$
Bodipy-FL-510-R$_2$
(a) Wash and image with 4 color scanner for sequence determination
(b) Cleave 3'-O-R$_1$ group and R$_2$-fluorophore with appropriate cleavage condition
ROX-R$_2$
3'-R$_1$-AC
Continuous cycle of extension, identification and cleavage (N cycles)
Denature the extended primers and rehybridize the original one
dCTP
dTTP
dATP
3'-O-R$_1$-dGTP
GTAC
Cleave 3'-O-R$_1$ group with appropriate cleavage condition
GCGTAC
Continuous cycle of extension and cleavage (N cycles)
GA-----GCGTAC
AGA-----GCGTAC
3'-R$_1$-AGA-----GCGTAC
R6G-R$_2$
3'-R$_1$-UAGA-----GCGTAC
TAGA-----GCGTAC
Continuous cycle of extension, identification and cleavage

Fig. 6.4. Template “walking” method 3 for SBS with CF-NRTs

6.2.2. Strategy 2: template "walking" with universal bases

In this variation on Strategy 1, the template reset is achieved not with nucleotide walking, but with the use of a longer primer partially consisting of universal nucleotides for the second round. Attachment of the template DNA to the surface and the first few steps of the procedure are identical to the first method. However, after stripping the first extended primer for the initial sequence readout, a long primer with the following features will be hybridized to the template: (a) the first half is identical to the initial primer; (b) the second half is composed almost entirely of universal bases. One possible candidate for the universal base is inosine (Fig. 6.5), which, in its deoxynucleoside form, can base pair with all four nucleotides,[5, 6] though its affinity for C and A is significantly higher than for G and T; (c) the last one or two anchoring bases of the long primers are degenerate with each of the four possible bases being represented.

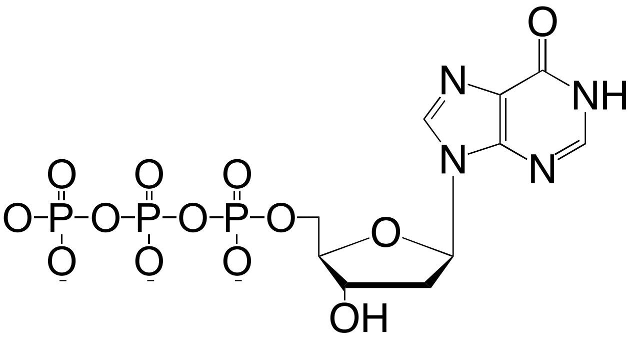


Fig. 6.5. Structure of 2'deoxyinosine-5'-triphosphate in deoxynucleotide form

Since the universal bases can form hydrogen bonds with any of the other four bases with some efficiency, they have the capacity to bind to the first 20 or so bases of the sequence. However, because the melting temperature of the ensuing hybridization is reduced substantially by the run of inosines,[6] a few of the bases in the first half and the two 3'-anchoring bases can be substituted with locked nucleotides. Locked nucleic acids have a chemical bond between the 2' and 4' carbons of the ribose.[7] While slower to associate with their complementary base, once hybridized, they tend not to dissociate. Thus, they provide a nice solution to ensure that the long primer remains attached appropriately to the template. In addition, the percentage of locked nucleosides in the primer can be manipulated to achieve higher hybridization strength. After hybridization of the new long primer, a second round of SBS with CF-NRTs can be performed (Fig. 6.6).

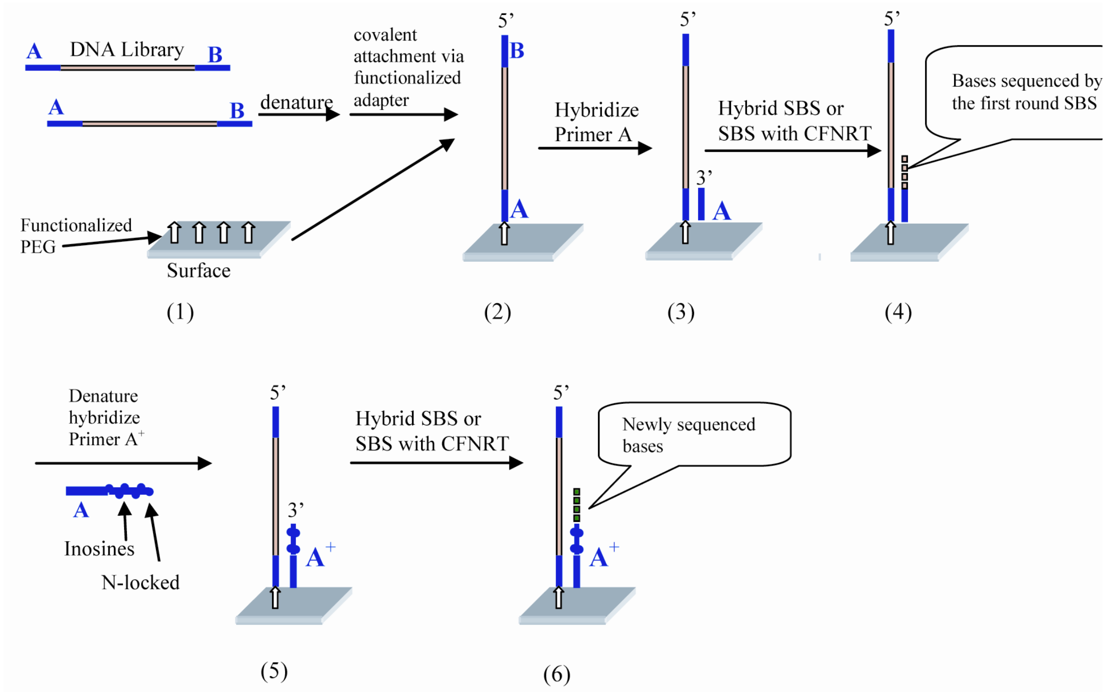

A
DNA Library
B
A
B
denature
covalent attachment via functionalized adapter
Functionalized PEG
Surface
(1)
5'
B
A
(2)
Hybridize Primer A
5'
3'
A
(3)
Hybrid SBS or SBS with CFNRT
5'
Bases sequenced by the first round SBS
(4)
Denature hybridize Primer A+
A
Inosines
N-locked
5'
3'
A+
(5)
Hybrid SBS or SBS with CFNRT
5'
Newly sequenced bases
A+
(6)

Fig. 6.6. Scheme for template resetting strategy 2: universal base walking

6.2.3. Strategy 3: multiple primer hybridization

In this third strategy, one or two additional primer annealing sites are introduced into the DNA to be sequenced at a distance just about equal to the number of bases that can be sequenced from the first primer.

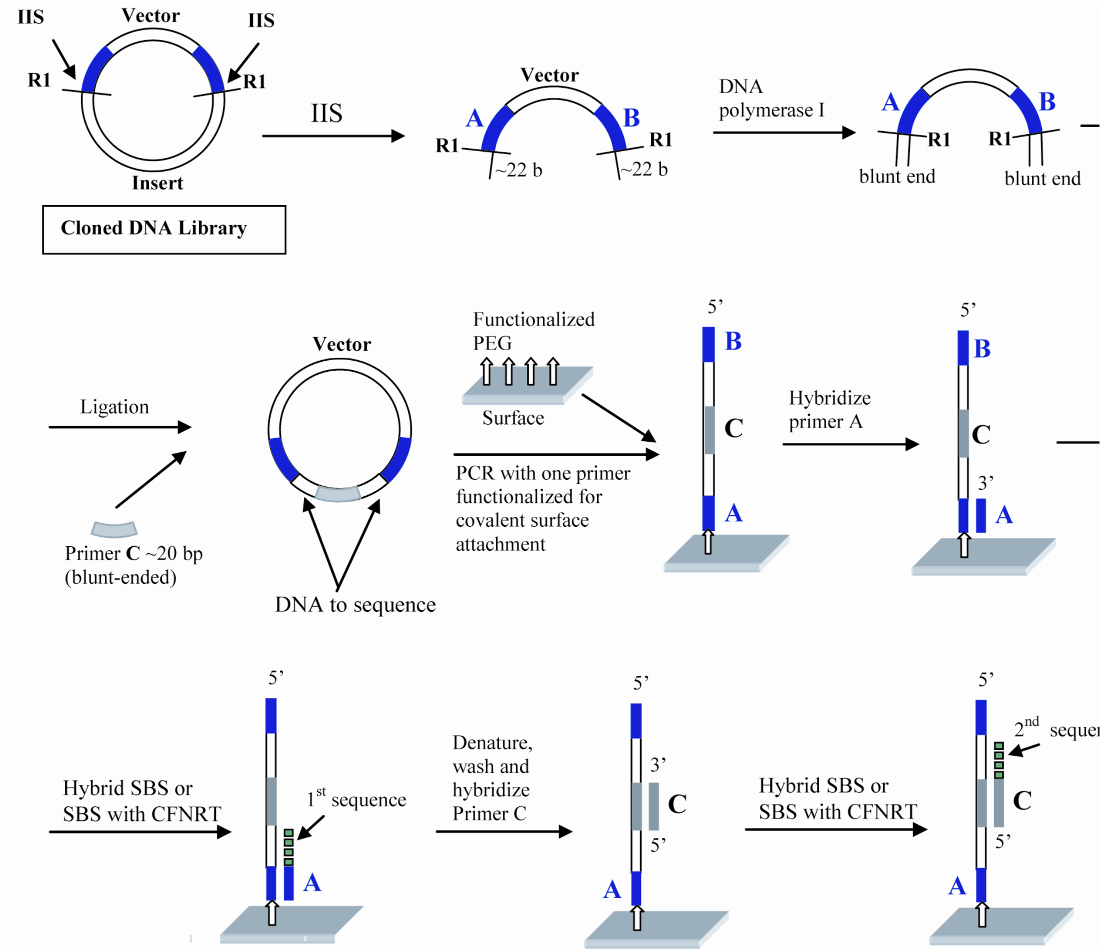


Fig. 6.7. Scheme for template resetting strategy 3: multiple primer hybridization

As illustrated in Fig. 6.7, template preparation for SBS utilizes the cloning of genomic DNA into a specially designed vector containing type IIS or III restriction sites (*MmeI* and *EcoP15 I*) flanking the genomic DNA cloning site. In this procedure size fractionated DNA (minimal length 100 bp) is inserted into the cloning vector using blunt-end ligation. Upon cloning, the resulting recombinant plasmids will be re-cut at one of the type IIS/III sites and the sticky ends will be filled in with Klenow enzyme. Next, specific sequencing primers will be introduced via ligation inside the genomic DNA inserts, 22 bases distant from the first primer in the case of *Mme*I or 27 bases away in the case of *EcoP15* I. After insertion of the internal priming sites, the constructs will be re-cloned in *E. coli*, the recombinant plasmids isolated and the inserts re-amplified by PCR at vector-insert junctions and attached to the beads for sequencing. Alternatively, emulsion or polony PCR strategies can be used to accomplish attachment of single molecules to individual beads or slide locations and their subsequent amplification at a much lower cost than cloning. In any case, once the DNA is immobilized, the first round of SBS with CF-NRTs will be primed from the flanking primer (P1), then after stripping these extended primers, the second set of sequencing reactions will be initiated at the internal primer (P2). It should be noted that with this scheme, the two sequenced portions come from opposite ends of the initial DNA, and are in essence paired end reads.

Several novel modifications of this approach can address the desire of many investigators to sequence an entire 100-base stretch of DNA, the length of a typical exon including surrounding intronic bases adjacent to the splice site. For instance, one can prepare a construct with two internal primers. In this case, the initial vector is designed with *Mme*I at one flank and *EcoP15*I on the other; using two consecutive restriction, cloning and circularization steps; the final construct consists of four alternative priming sites (two on the insert flanks and two internal), which in the case of 100 bp segments of genomic DNA will guarantee their complete sequencing with 25-30 cycles of SBS and three primer resets. The extra cycles would enable some of the sequence reads to run into the next primer, which would help to confirm the direction (*e.g.*, the last sequence might end with the *Mme*I or *EcoP15*I site). Other tricks would include modifying the ends of the primers to allow looping and reverse direction sequencing, incorporation of one or two decoding bases in the internal primers to confirm directions, and deconvoluting the results after all the data is generated. One would want to have a single set of primers for sequencing, regardless of which strand is attached. In order to achieve this, and to overcome the non-directional nature of their insertion, the internal primer or primers will be designed as palindromes so that sequencing can be initiated in either direction.

6.3 Results and Discussion

We have proposed a number of primer resetting (walking) strategies to increase the read length of SBS. The fundamental rationale behind primer resetting is to recover the initial template after one round of sequencing and start the next round anew at a downstream base to cover more bases. Among the three strategies, we focused on Strategy 1, which was template "walking" by unlabeled nucleotides, due to its simplicity and compatibility for implementation into our SBS platform. In general, three steps are involved in this strategy: (1) annealing of the sequencing primer and performing a 1st round of 4-color SBS, (2) denaturing the sequenced section of the template to recover a single-stranded DNA for annealing of the second primer. After the fresh primer annealing, which is identical to the first primer, the template is reset by extending the primer with unlabeled nucleotides (natural or modified) to the approximate end of the first round sequence, and then (3) sequencing from that point on with 4-color SBS.

We introduced three different methods to accomplish the template walking part of the strategy. Method 1 used the complete set of 3'-capped NRTs to walk over the template. Since each cycle of walking reaction could only extend the primer by one base, the entire process was deemed to be laborious and inefficient. We explored

Method 2 (walking with three natural nucleotides) and Method 3 (walking with three natural nucleotides and one NRT) to seek the most accurate and efficient approach for template walking.

6.3.1. Template walking using three natural nucleotides and MALDI-TOF MS characterization of walking products

The rationale behind template walking using three natural nucleotides is the halting of the DNA polymerase reaction when the incoming nucleotide corresponding to the next base in the template is absent. For example, if the downstream sequence after the primer is 5’-AGCTAGCT, then by using three natural nucleotides (dATP, dCTP, and dGTP) in the incorporation mixture, the primer is only extended to 5’-AGC. The polymerase reaction stops at this point due to the missing nucleotide (dTTP). In order to further extend the DNA strand, a different three nucleotide mixture (dTTP, dATP, and dGTP) is used so that the reaction stops at the C position this time. By alternating the two incorporation mixtures, the primer is extended to the end of the first round of sequencing so the next round can be initiated.

In order to carry out template walking with three natural nucleotides, it is crucial to verify that DNA polymerase can incorporate the nucleotides with high fidelity and accuracy. More importantly, the polymerase must stop its activity at the position of the template where the incorporated nucleotide to be is missing. We tested various DNA polymerases to find the most suitable one for our template walking method. For each polymerase, a single base extension reaction was performed using a self-priming template (SP26T, m.w. = 7966) and the natural nucleotide dTTP. Since the bases adjacent to the primer site is 5’-TGAC, the polymerase reaction was expected to stop after incorporating only one base. Four different commercially available DNA polymerases were tested: 9^{o}N Therminator II (New England Biolab), Pfu DNA Polymerase (with 3'-5' exonuclease activity, Agilent Technologies), Tgo DNA Polymerase (Roche Applied Science), and Thermo Sequenase (GE Healthcare), and the products were analyzed using MALDI-TOF MS. As shown in Fig. 6.8, the four polymerases behaved differently during the single base extension reaction. Therminator II (Fig. 6.8A) was unable to stop after incorporating the first dTTP. Instead, it mis-incorporated an extra dTTP to yield a DNA product extended with two Ts instead of one (m.w. = 8606). Both the Pfu (Fig. 6.8B) and Tgo polymerases (Fig. 6.8C) failed to incorporate the single nucleotide (dTTP) due to the lack of the other three nucleotide substrates (dATP, dGTP, and dCTP). Instead, the looped primer was cleaved by the polymerases due to their 3'-5' exonuclease activity. Tgo polymerase digested five bases off the primer, as demonstrated by the corresponding peaks in the MALDI-TOF MS spectrum. Thermosequenase was the only polymerase to correctly incorporate a single dTTP and promptly stopped ahead of a mis-match at the second base, displaying a single peak at m/z = 8286 (Fig. 6.8D). Hence, we decided to use Thermosequenase for further testing.

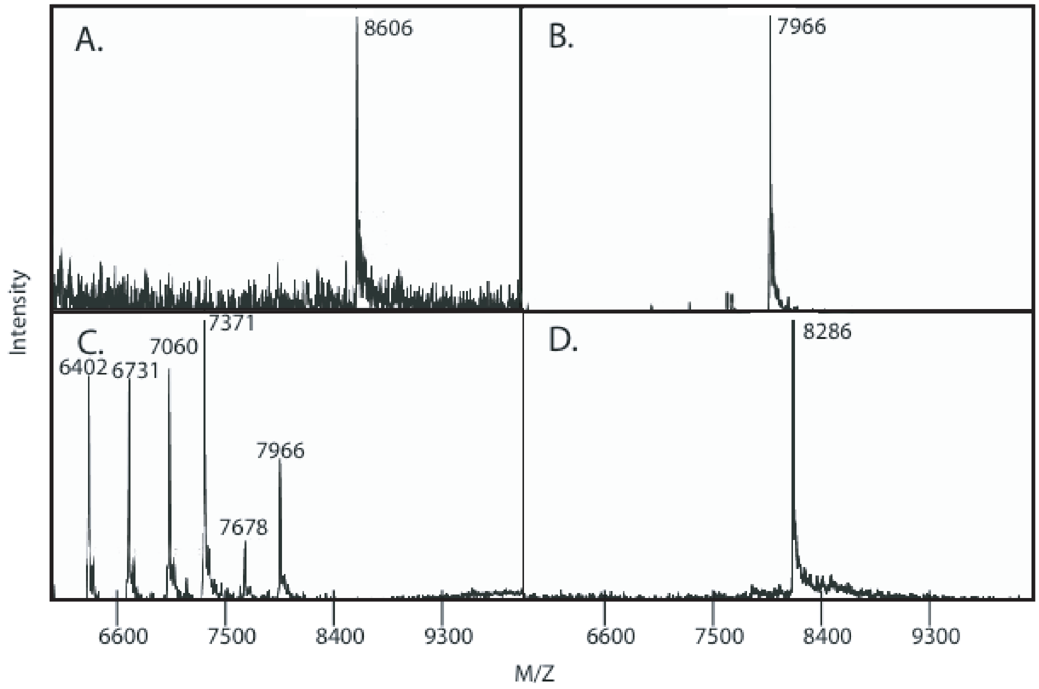


Fig. 6.8. MALDI-TOF MS spectrum of single base extension reaction using self-priming SP26T template and dTTP with DNA polymerase: A) Therminator II, B) Pfu DNA polymerase, C) Tgo DNA polymerase, and D) Thermosequenase

Our next task was to verify whether Thermosequenase could incorporate three natural nucleotides with high fidelity and accuracy since this approach for template walking relied on the use of all but one natural nucleotide to control the walking rate. The same self-priming template SP26T was used with dATP, dGTP, and dTTP. Since the sequence next to the primer was 5'-TGAC, we expected the polymerase to stop at the third base (A). However, the MALDI-TOF MS spectrum (Fig. 6.9) revealed that while a large portion of the extension products had the correct three nucleotides incorporated (m/z = 8908), the remaining products had a mis-incorporation of the forth base (m/z = 9214).

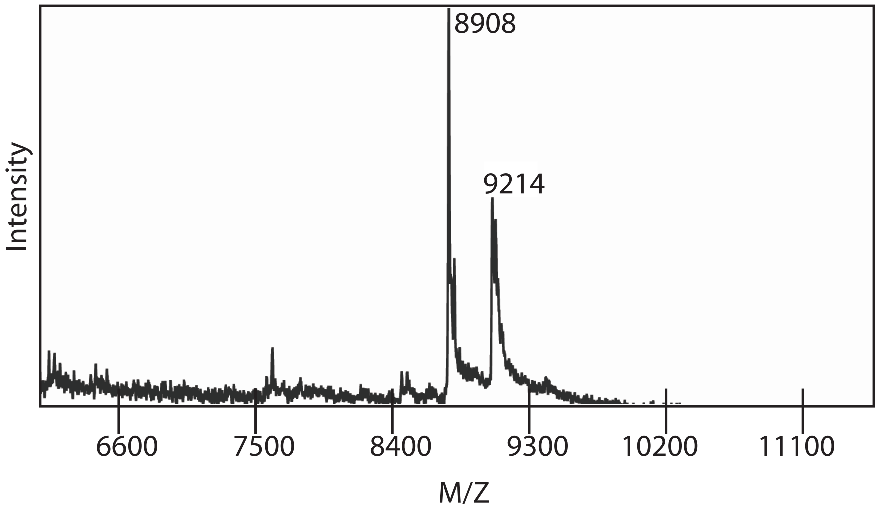


Fig. 6.9. MALDI-TOF MS spectrum of self-priming SP26T template extended with dATP, dGTP, and dTTP using Thermosequenase. Both the expected product (m/z = 8908) and product with extra base mis-incorporation (m/z = 9214) were present.

We hypothesized that the presence of excessive dNTPs in the reaction mixture during the extension reaction lowered the fidelity of Thermosequenase, resulting in the incorporation of the extra base. Hence, we added a PyroE solution (containing Apyrase, an enzyme used to eliminate un-reacted dNTPs during pyrosequencing) to the extension mixture to digest the excess dNTPs. The result, as shown in Fig. 6.10, was the correct incorporation of the three nucleotides without any mis-incorporation.

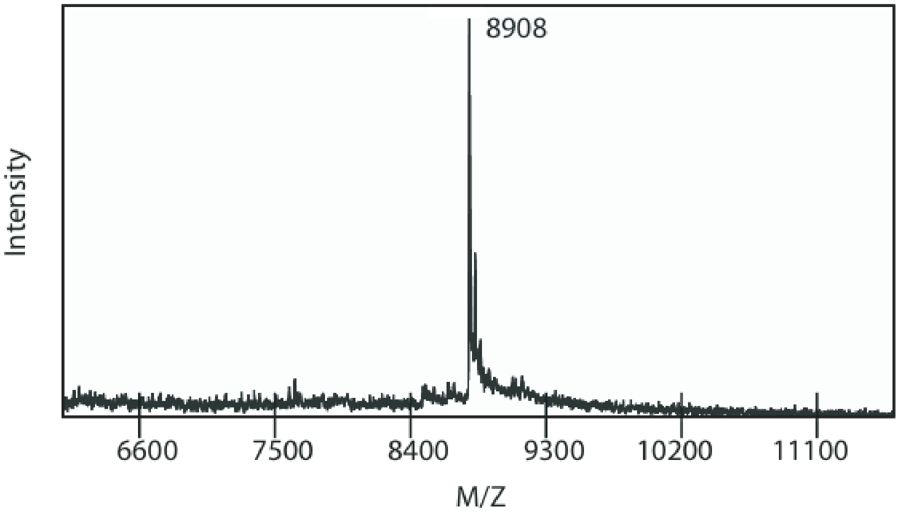


Fig. 6.10. MALDI-TOF MS spectrum of self-priming SP26T template extended with dATP, dGTP, and dTTP using Thermosequenase and PyroE solution. Only the expected product (m/z = 8908) was present, demonstrating the successful removal of excessive dNTPs in the incorporation mixture.

To truly integrate template walking with natural nucleotides into our SBS platform, we needed to carry out the extension reactions using linear templates instead of self-priming looped template. Primers extended on linear templates could be removed to allow the next round of sequencing, which was essential to our walking strategy. When we attempted to adapt our self-priming template walking conditions to the linear templates, mis-incorporations of more than two bases occurred. Even after adjusting various reaction conditions such as temperature, Thermosequenase concentration, nucleotide concentration, PyroE concentration, and time, we were still not able to obtain the correct number of incorporated nucleotides. Hence, we decided to explore our next option, which was template walking with three natural nucleotides and one NRT.

6.3.2. Template walking using three natural nucleotides and one NRT for four-color DNA Sequencing by Synthesis

The rationale behind template walking using three natural nucleotides (dATP, dCTP, and dTTP) in combination with one nucleotide reversible terminator (NRT, 3'-O-N_3-dGTPs) is the halting of the walking reaction whenever the NRT is incorporated. Upon cleavage of the 3'-azidomethyl capping moiety on the 3'-O-N_3-dGTPs to restore the OH group at the 3' position, the next round of walking can be initiated (Fig. 6.11). Unlike template walking with three natural nucleotides, this walking method includes the use of all four nucleotides (A, C, T, and 3'-reversibly terminated G). It greatly increases the accuracy of the DNA polymerase and reduces the chance of mis-incorporation. In addition, all walking templates are synchronized at the G nucleotide after the incorporation of the NRT, which later facilitates bioinformatic analysis of the sequencing information.

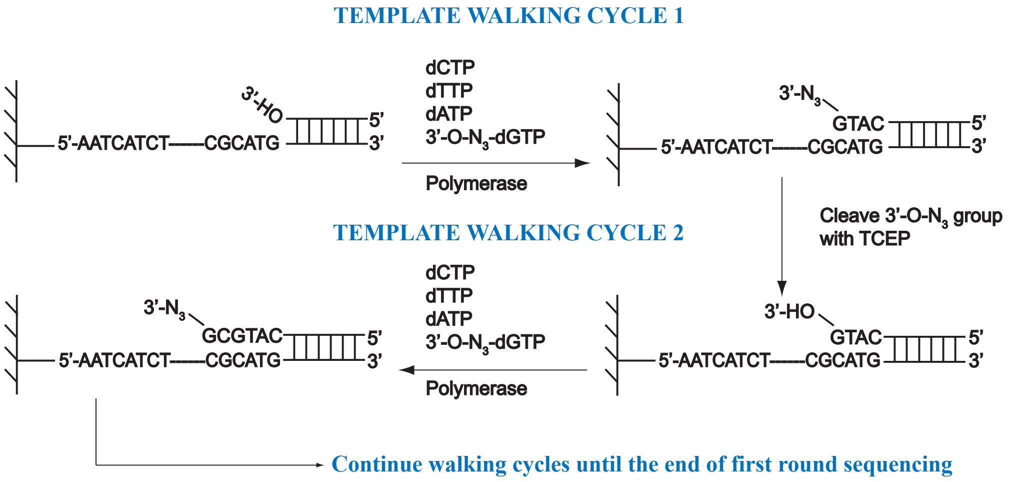


Fig. 6.11. Scheme for template walking using three natural nucleotides (dATP, dCTP, and dTTP) and one NRT (3'-O-N_3-dGTP)

We integrated this template walking technique into our four-color SBS using CF-NRTs approach, hoping to increase the overall read length of the sequencing platform. We first carried out a round of SBS using CF-NRTs (3'-O-N_3-dNTP-N_3-fluorophores) on a linear template to its maximum read length according to the process described in Chapter 5. Immediately after the first round of SBS, the extended sequencing primers were removed by denaturing in mild alkaline conditions. The initial primer was annealed to the template and the first template walking cycle with three dNTPs (dATP, dCTP, dTTP) and one NRT (3'-O-N_3-dGTP) was performed. The walking process stopped after the G. To synchronize any un-extended primers, a capping step with 3'-O-N_3-dGTP was performed. In case there were still un-extended primers, which would interfere with the base-calling process during SBS, we carried out another capping step, using ddNTPs. Any primer terminated with ddNTPs could no longer participate in the subsequent extension reactions. This step ensured the uniformity of all templates and minimized de-phasing of the signals during actual sequencing detection. The capping step was followed by cleavage of the 3'-O-azidomethyl group to regenerate the 3'-OH position of the extended primer so that it was ready for next cycle of walking. After several cycles of walking, a second round SBS using CF-NRTs was carried out to identify additional bases following the ones sequenced in the first round of SBS.

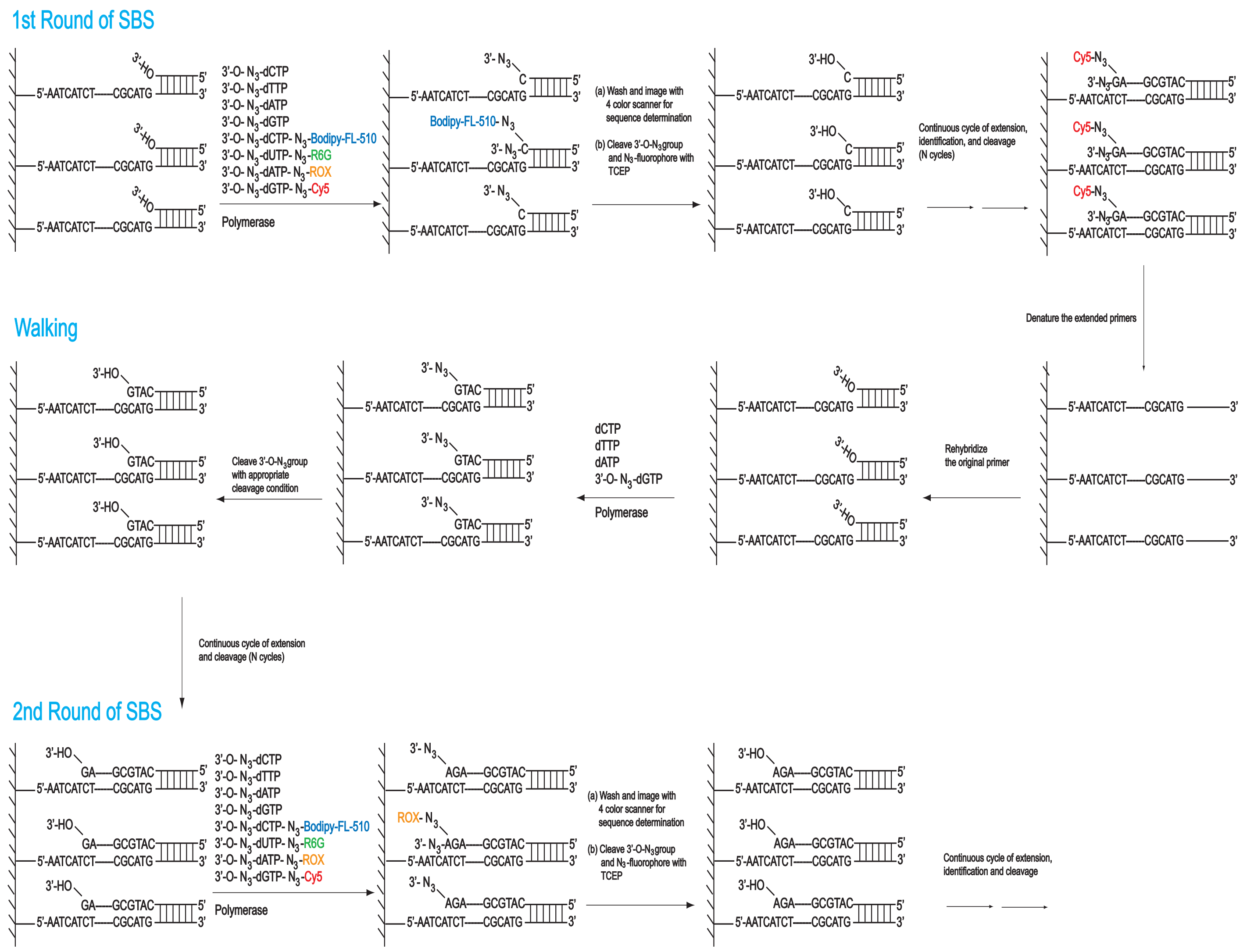

1st Round of SBS
3'-HO
5'-AATCATCT—CGCATG
5'
3'
3'-O- N_3-dCTP
3'-O- N_3-dTTP
3'-O- N_3-dATP
3'-O- N_3-dGTP
3'-O- N_3-dCTP- N_3-Bodipy-FL-510
3'-O- N_3-dUTP- N_3-R6G
3'-O- N_3-dATP- N_3-ROX
3'-O- N_3-dGTP- N_3-Cy5
Polymerase
3'- N_3
C
Bodipy-FL-510- N_3
3'- N_3-C
(a) Wash and image with 4 color scanner for sequence determination
(b) Cleave 3'-O-N_3group and N_3-fluorophore with TCEP
Continuous cycle of extension, identification, and cleavage (N cycles)
Cy5-N_3
3'-N_3GA—GCGTAC
Denature the extended primers
Walking
GTAC
Cleave 3'-O-N_3group with appropriate cleavage condition
dCTP
dTTP
dATP
3'-O- N_3-dGTP
Rehybridize the original primer
Continuous cycle of extension and cleavage (N cycles)
2nd Round of SBS
GA—GCGTAC
AGA—GCGTAC
ROX- N_3
3'- N_3-AGA—GCGTAC
Continuous cycle of extension, identification and cleavage

Fig. 6.12. General scheme of SBS integrated with template walking

Fig. 6.12 illustrates the general scheme for template walking integrated SBS, which is composed of repeated cycles of 1) SBS, 2) extended primer removal, and 3) template walking. We used a DNA chip immobilized with linear 90mer templates on the surface to carry out template walking in SBS. After annealing the 37mer primer and completing the first round of SBS using CF-NRTs, we correctly identified 32 bases of the template (Fig. 6.13). The extended sequencing primers were removed by treating the DNA chip with a mild alkali solution. The disappearance of the fluorescent signal from the chip was a good indication that denaturing of the primer was achieved (Fig. 6.13). The same 37mer primer was then annealed for template walking using dATP, dCTP, dTTP, and 3'-O-N_3-dGTP. After four cycles of walking, which were each composed of extension, the two capping steps, and cleavage, the primer was extended by 30 bases. A second round of SBS using CF-NRTs was performed to identify an additional 23 bases (Fig. 6.14). By combining template walking and SBS, we were able to sequence the target template in its entirety, pushing the read length to 53 bases (Fig 6.15).

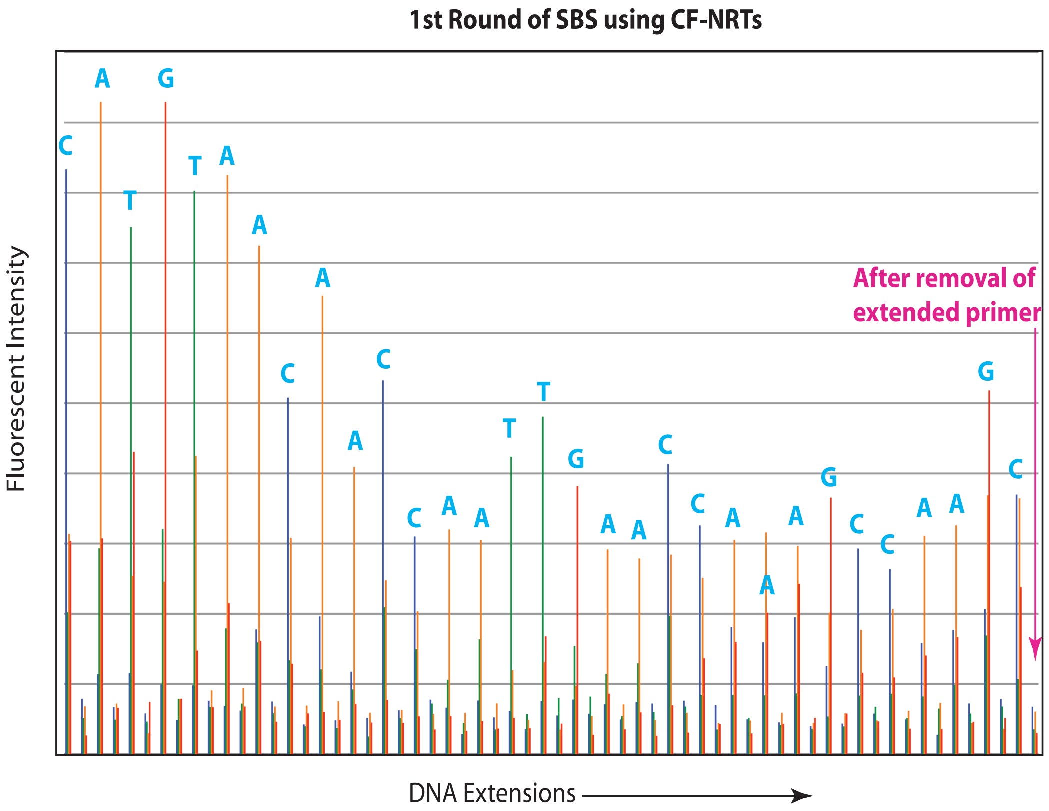


Fig. 6.13. Four-color sequencing data plot of raw fluorescence emission intensity obtained by 1st round SBS using CF-NRTs of a 90mer linear template. 31 consecutive bases were correctly identified. The small groups of peaks between the identified bases represent fluorescent background from the DNA chip after cleavage. The last small group of peaks, following base 31, revealed the removal of the sequencing primer extended with CF-NRTs after the denaturing process.

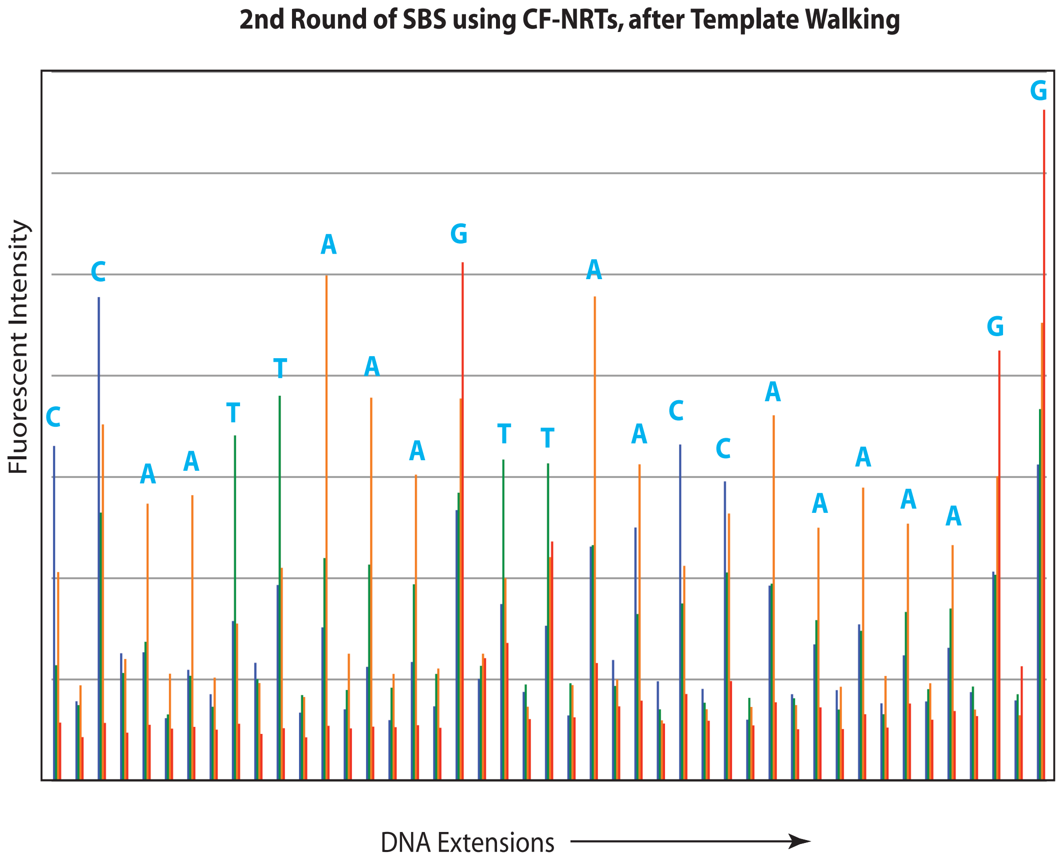


Fig. 6.14. Four-color sequencing data plot of raw fluorescence emission intensity obtained by 2nd round SBS using CF-NRTs after four cycles of template walking. 23 consecutive bases were correctly identified. The small groups of peaks between the identified bases represent fluorescent background from the DNA chip after cleavage.

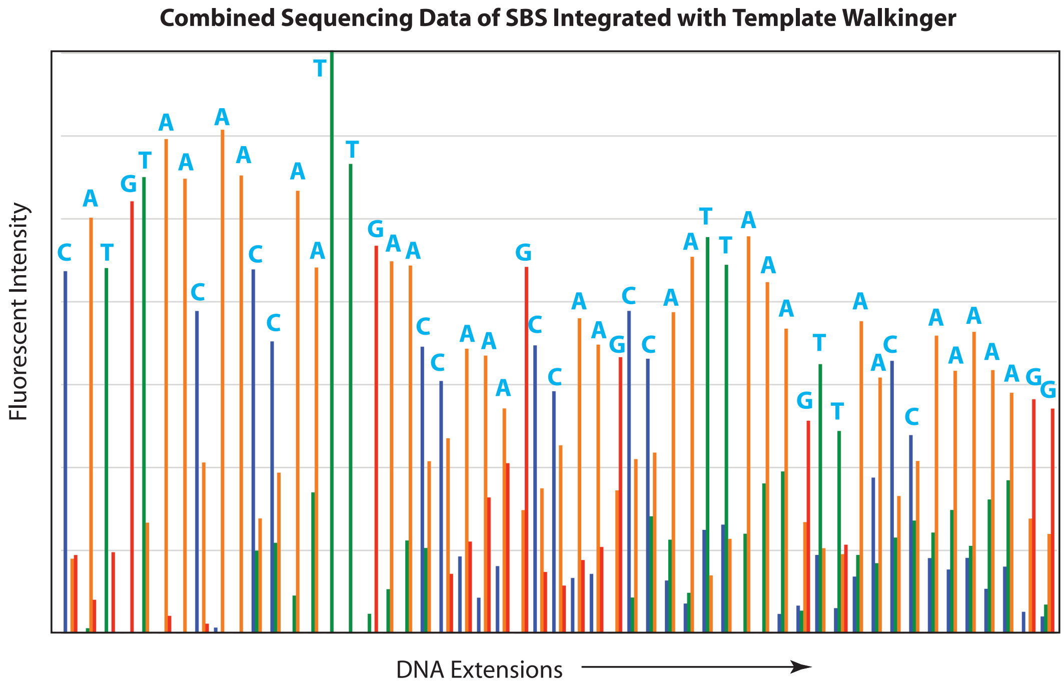


Fig. 6.15. Four-color sequencing data plot of fluorescence emission intensity obtained by combining 1st and 2nd round SBS using CF-NRTs. The first 30 bases were sequenced during the 1st round of SBS. After stripping away the sequencing primer extended with CF-NRTs, the original primer was reannealed to the template for four cycles of template walking using a mixture of three natural dNTPs (dATP, dCTP, and dTTP) and one NRT (3'-O-N_3-dGTP). Upon reaching the base at the end of the previous sequencing round, SBS was re-initiated to correctly identify the next 23 consecutive bases. By using SBS integrated with template walking, the target template of 53 bases was sequenced in its entirety.

6.4 Materials and Methods

General Information. All solvents and reagents were reagent grade, purchased commercially and used without further purification. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Oligonucleotides used as primers or templates were synthesized on an Expedite nucleic acid synthesizer (Applied

BioSystems) or purchased from Midland. 9^oN Therminator II polymerase (exo-) A485L/Y409V was obtained from New England Biolabs. Pfu DNA Polymerase with 3'-5' exonuclease activity was obtained from Agilent Technologies. Tgo DNA Polymerase was purchased from Roche Applied Science, and Thermosequenase was ordered from GE Healthcare. CodeLink HD Activated Slides for DNA immobilization were obtained from GE Healthcare as well. Phosphoramidite reagents and columns for oligonucleotide synthesis were purchased from Glen Research (Sterling, VA). Both the nucleotide reversible terminators and cleavable fluorescent nucleotide reversible terminators were purified by reverse-phase HPLC on a 150×4.6 mm C18 column (Supelco), mobile phase: A, 8.6 mM Et_3N / 100 mM 1,1,1,3,3,3-hexafluoro-2-propanol in water (pH 8.1); B, methanol. Elution was performed from 100% A isocratic over 10 minutes followed by a linear gradient of 0-50% B for 20 minutes and then 50% B isocratic over another 30 minutes.

6.4.1. Template walking using three natural nucleotides and MALDI-TOF MS characterization of walking products

We tested various commercially available DNA polymerases (9^oN Therminator II, Pfu, Tgo, and Thermoseqenase) for their fidelity and accuracy in the single base extension reaction. Each extension reaction was composed of a self-priming DNA template SP26T (5'-**GTCA**GCGCCGCGCCTTGGCGCGGCGC-3', 40pmol), 200pmol of dTTP, 1X enzyme reaction buffer, and 1 unit of the DNA polymerase. In the case of 9^oN Therminator II, 20nmol of Mn^{2+} was used as the cofactor. The reaction mixture was incubated at 65^oC for 2 minutes. The extension products were analyzed using MALDI-TOF MS.

After assessing the results from the single base extension reactions, we chose Thermosequenase to carry out template walking with three natural nucleotides due to the polymerase’s fidelity and accuracy. A three-base extension reaction was carried out using the same self-priming template (SP26T, 40pmol), three natural nucleotides (dTTP, dGTP, and dATP, 200pmol each), 1X Thermosequenase reaction buffer, and 1 unit of Thermosequenase (diluted from 36 unit/μL using the manufacturer’s enzyme dilution buffer). The reaction underwent incubation at 65°C for 15 minutes. The resulting extension products were analyzed using MALDI-TOF MS after ethanol precipitation and ZipTip desalting procedure. Due to the presence of mis-incorporation products, as shown in MALDI-TOF MS spectra, we repeated the same extension reaction with the addition of 2μL PyroE solution (Pyrosequencing enzyme mixture, Roche). The extension products were again analyzed using MALDI-TOF MS after ethanol precipitation and ZipTip desalting.

We also attempted three-base extension reactions using linear templates instead of self-priming (looped) ones. 20pmol of the linear template (3’ACGGAGAACGAAGAGAAAAGGATAGGACTCATCACCATTAGATG ACCCTGCCTTGTCGAACTCCACGCACAAACACGGACAGGACCCTCTCTGGCCGCGTGTCTCCTTC-5’), annealed with 60pmol of a 17mer primer (5’-GATAGGACTCATCACCA-3’), were mixed with dT/C/GTP (200pmol each), 1X Thermosequenase reaction buffer, and 1 unit of Thermosequenase for the extension reactions. Various amounts of PyroE solution were added, ranging from 1-5μL volumes. The reactions were incubated at 55°C for 15 minutes and the products were analyzed using MALDI-TOF MS. The resulting MALDI-TOF MS spectra all showed mis-incorporation products. Hence, we decided to pursue another template walking method, walking with three natural nucleotides and one NRT.

6.4.2. Template walking using three natural nucleotides and one NRT for four-color DNA Sequencing by Synthesis

We prepared a DNA chip for our template walking integrated SBS. The DNA chip was constructed by immobilizing 5'-amino-modified single strand oligonucleotide (90mer) on a CodeLink HD microarray slide. A 37 base primer was annealed at 37°C for 30 minutes. We then carried out the first round of SBS using CF-NRTs. Similar to the SBS process described in Chapter 5, 10 μL of extension solution consisting of 3'-O-N_3-dATP-N_3-ROX (4 pmol), 3'-O-N_3-dGTP-N_3-Cy5 (1 pmol), 3'-O-N_3-dCTP-N_3-Bodipy-FL-510 (5 pmol), 3'-O-N_3-dUTP-N_3-R6G (3pmol), 3'-O-N_3-dATP (45pmol), 3'-O-N_3-dGTP (48 pmol), 3'-O-N_3-dCTP (42 pmol), 3'-O-N_3-dTTP (45 pmol), 1 unit of 9°N DNA polymerase(exo-) A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer was spotted on the surface of the chip. This extension mixture was incubated at 55°C for 15 minutes. Upon the completion of the extension reaction, the chip was washed with SPSC buffer (0.1M sodium phosphate/0.5M NaCl, pH 7.5, 0.1% Tween-20) for 2 minutes, followed by a brief rinse with dH_2O. The blow-dried chip was scanned with a 4-color ScanArray Express scanner (Perkin–Elmer Life Sciences) to detect the fluorescence signal. The four lasers on the 4-color scanner had excitation wavelengths of 488, 543, 594, and 633 nm and emission filters centered at 522, 570, 614, and 670 nm. After signal detection, a capping reaction solution, consisting of 45 pmol each of 3'-O-N_3-dATP, 3'-O-N_3-dGTP, 3'-O-N_3-dCTP, 3'-O-N_3-dTTP, 1 unit of 9°N DNA polymerase(exo-)

A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer, was spotted on the same spot on the DNA chip and incubated at 55°C for 20 minutes. After washing the spot with SPSC buffer, a second capping solution consisting of the four dideoxynucleotide triphosphates (100 pmol each of ddATP, ddGTP, ddCTP, and ddTTP, Fluka Analytical), 1 unit of 9°N DNA polymerase(exo-) A485L/Y409V, 20 nmol of $MnCl_{2,}$ and 1X Thermopol II reaction buffer, was placed on the DNA spot for another incubation at 55°C for 5 minutes. For the cleavage process immediately followed the dual capping steps, about 10 μl of 100 mM TCEP (pH 9.0) was spotted on the same area of the DNA chip for 20 minutes of incubation at 55°C. After washing the surface with SPSC buffer, the chip was scanned again to compare the intensity of fluorescence after cleavage with the original fluorescence intensity. This process was repeated by carrying out the extension with subsequent synchronization, washing, fluorescence detection, and cleavage steps to identify the first 31 bases of the template.

After the 31st extension reaction, the DNA chip was incubated twice with 10μL of a denaturing solution (consisting of 80% formamide, 0.1% SDS, and 50mM Tris HCl, pH7.5) at 65°C for 30 minutes. The chip was rinsed with distilled water and air-dried. After fluorescent signal scanning to confirm the removal of the extended sequencing primer, the initial 37mer primer was annealed to the template at 37°C for 30 minutes. Template walking was carried out by adding 10 μL walking solution consisting of 10 pmol of each dNTP (dATP, dCTP, dTTP), 20 pmol 3'-O-N_3-dGTP, 1 unit of 9°N polymerase, 20 nmol of Mn^{2+} and 1X ThermoPolII Buffer to the DNA spot and incubating at 65°C for 15 min. To synchronize the unextended templates, two capping steps were carried out. First, a 10 μL solution consisting of 10 pmol 3'-O-N_3-dGTP, 1unit of 9^oN polymerase, 20 nmol of Mn^{2+} and 1X ThermoPolII buffer was added to the same spot and incubated at 65^oC for 25 min. Then the second capping solution (10 μL) composed of 10 pmol of each ddNTP (ddATP, ddCTP, ddGTP, ddTTP), 2 unit of 9^oN enzyme, 20 nmol of Mn^{2+} and 1X ThermoPolII buffer was added and incubated at 65^oC for 5 mins. The DNA chip was washed with SPSC buffer three times followed by the addition of TCEP cleavage solution (100 mM, pH = 9) for 15 mins incubation at 65^oC. Upon the completion of the cleavage reaction to recover the 3'-OH group, the next walking cycle was initiated. A total of four walking cycles were performed to extend the original sequencing primer to the end of the first round of SBS.

Once the walking process was completed, the second round of SBS was carried out using the procedure described above. However, the amount of CF-NRTs (3'-O-N_3-dNTP-N_3-fluorophores) was doubled in the extension mixture to ensure the strength of fluorescent signals. After another 23 cycles of extension, capping, detection, and cleavage, the second portion of the template sequence was correctly identified.

6.5 Conclusion

In our attempts to increase read length for DNA Sequencing by Synthesis, we developed a novel template walking strategy that allowed us to carry out multiple rounds of SBS. The template walking method involved extending the sequencing primer with three natural nucleotides and one NRT so that the polymerase reaction was temporarily paused when the NRT was incorporated. Upon restoring the 3'-OH group of the NRT incorporated into the primer via cleavage, the next cycle of walking could be carried out. The advantage of using the complete set of nucleotides (three natural and one modified) was that the DNA polymerase could function with high fidelity and accuracy. In addition, the incorporation of the NRT served as a measure to ensure that all templates were synchronized. We have successfully demonstrated the integration of this template walking strategy into our four-color DNA SBS platform by performing one round of SBS, replacing the extended sequencing primer with the initial one, carrying out four cycles of template walking, and then completing a second round of SBS. Through this effort, we were able to sequence a linear DNA template in its entirety, nearly doubling the read length of our previous sequencing results (from 30 to 53 bases).

We are currently in the process of integrating our template walking strategy into commercial sequencers. Our focus so far has been on Illumina's Genome Analyzer II (IGA), the next-gen sequencer that is based on the four-color SBS technology we first reported. We have utilized the Cluster Generation Station (CGS), which is part of the IGA, for automated template walking in flow cells. The .xml recipe files of the CGS are modified so that the template walking protocol described in this chapter can be implemented. Several preliminary trials of automated template walking, with various numbers of walking cycles, followed by SBS in IGA have been carried out. The results are very promising, although much optimization is still needed. By combining the throughput of commercial sequencers, including the Intelligent Bio-System's which uses our lab's sequencing by synthesis approach, and our novel template walking strategy, we can truly advance the sequencing field into the much-coveted $1000 genome territory.

6.6 References

1. Ju, J.; Li Z.; Edwards, J.; Itagaki, Y. Massive parallel method for decoding DNA and RNA. United States Patent 6,664,079 USA, **2003**.

2. Ju, J., *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **2006**, *103,* 19635-19640.

3. Yu, L.; Qui, C.; Guo, J.; Kalachikov, S.; Li, Z.; Xu, N.; Li, X.; Shi, S.; Russo, J. J.; Turro, N. J.; Ju, J. Novel template walking strategy for read length increment in DNA sequencing by synthesis. **2009**, *in preparation.*

4. Yu, L.; Guo, J.; Qui, C.; Li, Z.; Kim, D. H.; Cao, H.; Zhang, S.; Meng, Q.; Marma, M. S.; Wu, J.; Xu, N.; Li, X.; Shi, S.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis (SBS) improvements with cleavable fluorescent nucleotide reversible terminators. **2009**, *submitted, under review.*

5. Loakes, D. The applications of universal DNA base analogues. *Nucleic Acid Research*, **2001**, *29*, 2437-2447.

6. Ueno, Y.; Shibata, A.; Matsuda, A.; Kitade, Y. Thermal stability of triple helical DNAs containing 2'-deoxyinosine and 2'-deoxyanthosine. *Bioorg and Med Chem*, **2004**, *12*, 6581-6586.

7. Babu, B. R.; Wengel, J. Universal hybridization using LNA (locked nucleic acid) containing a novel pyrene LNA nucleotide monomer. *Chem Commum*, **2001**, 2114-2115.

Chapter 7: Massively Parallel Monitoring of Gene Expression in *Aplysia* Central Nervous System (CNS) using Four-color DNA Sequencing by Synthesis

7.1 Introduction

Through research efforts in the past several years, our lab has developed a novel paradigm for massive parallel DNA Sequencing by Synthesis (SBS) using cleavable fluorescent nucleotide reversible terminators (CF-NRTs). This SBS platform offers a high-throughput accurate sequencing technology that has the potential to become a vital tool for tackling many biological and genetic problems. Recently, the SBS principle based on CF-NRTs that we first reported has been successfully implemented in the Illumina Genome Analyzer (IGA), which has found wide range of applications in genome biology.[1, 2] This new sequencing technology has enabled us to start several high profile neurobiology projects. One of the projects, an ongoing collaboration with Dr. Eric Kandel of Columbia University College of Physicians and Surgeons and Dr. Leonid Moroz of University of Florida, is the study of the molecular mechanism of memory formation using *Aplysia* as a model organism. With the aid of the state-of-the-art sequencing technology, we hope to shed some light on one of the most complex mechanisms in living organisms.

Memory is typically defined as an organism's ability to store and recall information. The nervous system, composed of an intricate network of neurons, is responsible for both long and short-term memory formation. The neuron is an integrated dynamic molecular system, finely regulated at many levels. Most neurons are highly polarized, indeed among most compartmentalized cells in an organism.

Subcellular compartmentalization and localized mRNA processing in neurons lead to local protein synthesis. This asymmetric distribution of mRNAs characteristic of neurons has emerged as a general mechanism used by eukaryotic cells to distribute and regulate the levels of selected proteins in defined subcellular domains. Asymmetric mRNA distribution followed by translation has been found to be a mechanism for altering the strength of preexisting interconnections between neurons underlying learning and memory in the human brain.[3,4] It is also likely important in *de novo* synapse formation. Hence, it is important to map out the molecular differences, in particular the pattern of expression of specific transcripts in different neurons and their compartments. Characterizing cDNA libraries and measuring gene expression in isolated neurites and cell bodies can map these differences, leading to the possibility of actually identifying the mRNAs and their associated proteins while in transit from the cell body to the compartments. Our collaborators have recently delineated components of the cargo associated with the molecular motor kinesin retrieved following exposure of neurons to agents that modify neural structures using the ChIP-on-Chip approach with kinesin antibodies.[5] We are now in a position to identify specific mRNAs that are actively transported to the synapses or other regions of the cells following *de novo* synapse formation and learning related growth following synaptic stimulation by facilitatory neurotransmitters such as serotonin (5HT)[4] and inhibitory neurotransmitters such as FMRFamide.[10] With the aid of high-throughput sequencing, it is possible now to identify the passengers of this molecular trafficking that are recruited by different molecular signals important for learning using the gene expression approach based on SBS.

Aplysia californica is an ideal experimental model for this study due to the ease of monitoring its global transcriptome at the subcellullar level[7] from a single neuron up to the neural circuitry underlying complex behaviors. The nervous system of Aplysia contains only about 20,000 nerve cells, in about 9 central ganglia. Since a given ganglion controls a family of behaviors, the number of cells committed to a single behavior may be as small as 100 cells. Many of the nerve cells are gigantic (up to 1000 µm in diameter) so they can easily be dissected out for biochemical studies. From a single cell or even a subcellular compartment, one can obtain sufficient mRNA to make a cDNA library.

Fig. 7.1. Photo of *Aplysia californica*, courtesy of Moroz lab.

To provide a basis for systematic and comprehensive understanding of gene expression during long-term facilitation in *Aplysia*, we will start from isolating and characterizing steady-state mRNA transcripts, transcripts from whole ganglia treated with 5HT and FMRFamide. This will allow us to identify subsets of translationally regulated genes responsible for long-term facilitation and depression in *Aplysia*. By analyzing differential expression patterns of mRNAs with and without 5HT or FMRFamide treatment, we can identify target genes that might be responsible for memory formation and inhibition. These target genes may be candidate drug targets for developing new therapies for memory related diseases. In order to achieve these goals, we need a comprehensive gene expression platform to characterize mRNAs and their abundances in neurons and neuronal compartments. The next-generation IGA sequencing platform, developed by Illumina based on the four-color SBS chemistry we reported, is an integrated part of this gene expression platform, allowing the identification of the important mRNAs that play key roles in learning and memory.

7.2 Experimental Rationale and Overview

A hallmark of long-term memory storage is modulation of gene expression induced by experience. In *Aplysia*, application of 5HT, a modulatory transmitter induced by learning, alters gene expression in sensory and motor neurons of the gill-withdrawal reflex.[4] Previously, several genes such as C/EBP, CPEB, Reductase-related gene, sensorin, EF1alpha and Kinesin Heavy Chain were shown to be regulated by 5HT in *Aplysia*.[5, 7-9] Our goal is to identify and study the full repertoire of 5HT-regulated or learning-induced genes. Previously, our lab performed an initial genomic analysis of genes regulated by 5-HT in sensory neurons using microarrays. The initial screening of differentially expressed transcripts suggested that 5-HT modulates the expression of more than 600 genes. Since the hybridization-based gene expression analysis is limited in its power to identify rare gene transcripts important for the neuronal functions, we conducted digital gene expression studies implemented on the IGA sequencing platform. Such a collection will help us elucidate signaling networks underlying memory formation and storage.

IGA, Illumina's Genome Analyzer, is a massively parallel high-throughput sequencer based on DNA SBS technology using CF-NRTs first reported by our lab.

The entire sequencing process is composed of three general steps. First, target DNA is fragmented and ligated with specific adaptors, the process known as library preparation (Fig. 7.2).

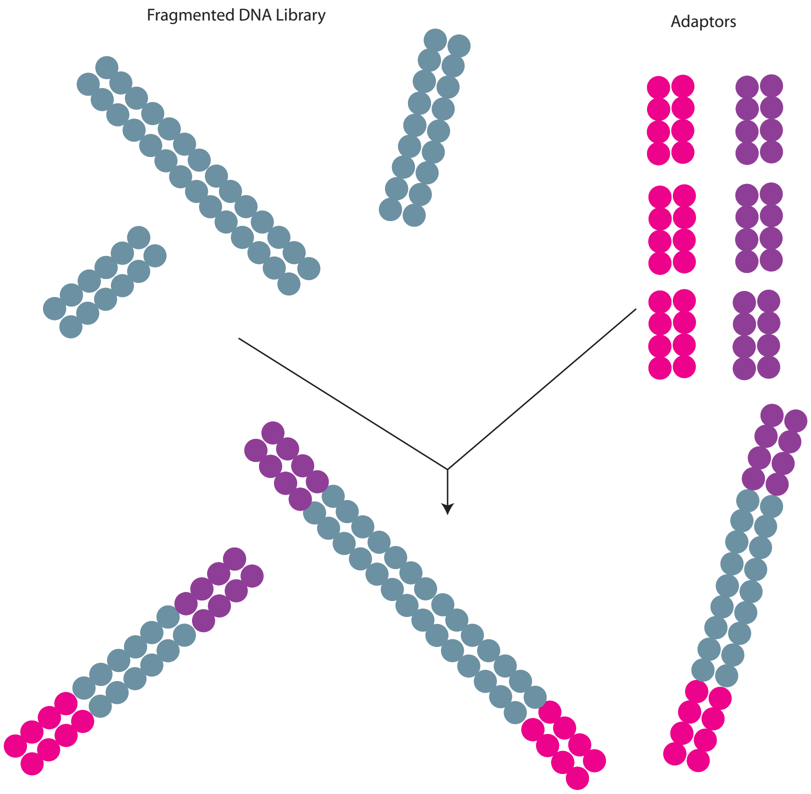


Fig. 7.2. Scheme of adaptor ligation for genomic library DNA in Illumina Genome Analyzer.

The resulting samples are then immobilized onto an eight-lane flow cell and amplified into clusters by bridge PCR[1] using a cluster generating station (Fig. 7.3). The third and last step is the actual sequencing process involving automated DNA SBS using CF-NRTs as described in previous chapters. Since up to 20 million templates can be attached to a single lane in the flow cell, each run will produce at least 1 gigabase of sequences assuming a 50 bases read length for each template. Such high throughput, combined with reasonable cost of sequencing consumables and short sequencing time, will enable us to conduct massively parallel monitoring of gene expression in *Aplysia* nervous system to investigate the molecular mechanism of memory formation.

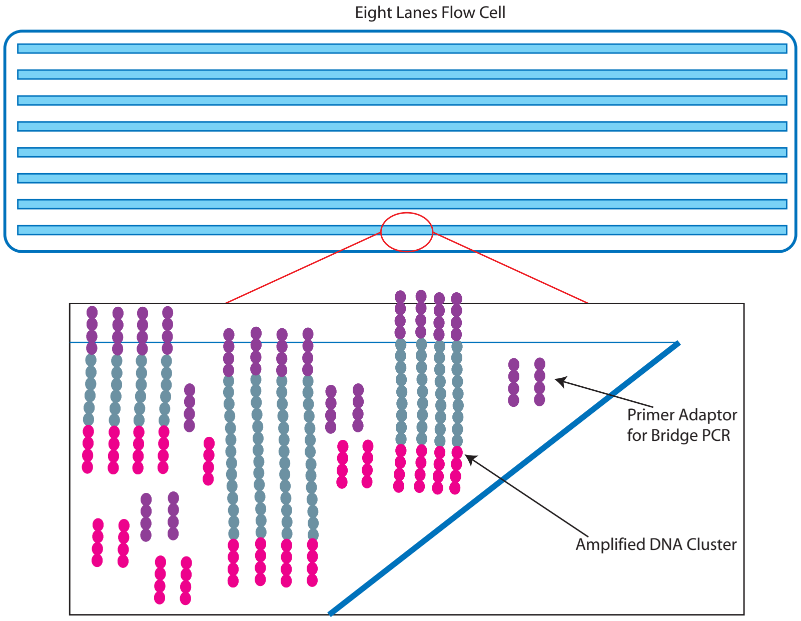


Fig. 7.3. Scheme of cluster generation in flow cell prior to sequencing in Illumina Genome Analyzer.

7.3 Results and Discussion

In order to evaluate the feasibility of using IGA for gene transcript analysis in neuronal cells, we conducted a pilot experiment for transcriptome sequencing of *Aplysia* neurons. It is known that in *Aplysia*, application of 5HT (serotonin), a modulatory transmitter induced by learning, alters gene expression in sensory and motor neurons of the gill-withdrawal reflex.[4] Previous studies have shown that several genes such as C/EBP and CPEB are regulated by 5HT in *Aplysia*.[7,8] In contrast, the neuropeptide FMRFamide released from a subset of inhibitory interneurons is able to suppress learning and memory consolidation and induce long-term depression.[10] In an effort spearheaded by Drs. Sergey Kalachikov and Irina

Morozova from our lab and Dr. Sathyanarayan Puthanveettil from the Kandel lab, we attempt to identify and study the full repertoire of 5HT-regulated and/or FMRFamide-regulated genes through digital gene expression studies enabled by the sequencing-by-synthesis approach we developed. We focused on a single neuron type, the sensory cluster of the Aplysia pleural ganglia, the cells of which are capable of forming functional connections with gill motor neurons. A total 36 clusters were exposed to either 5HT or FMRFamide. Total RNA was isolated following treatment, and the corresponding cDNA libraries were prepared. A total of 7 libraries, three treated with 5HT, three treated with FMRFamide, and one untreated sample, were allocated to seven lanes of the IGA flow cell with the eighth lane used for a bacterio phage genome as an internal control. The control lane was necessary as an internal standard for matrix and phasing estimation. After performing cluster generation, 36 cycles of SBS were carried out on the IGA. We evaluated several key statistics of the sequencing process to validate our results. Since IGA utilizes a CDC camera for fluorescent image capturing, we first examed the focus quality for each of the four fluorescent dye as sequencing cycle progressed. Based on the plot shown in Fig. 7.4, the majority of the focus quality was above the standard of 70, indicating successful capture of fluorescent images.

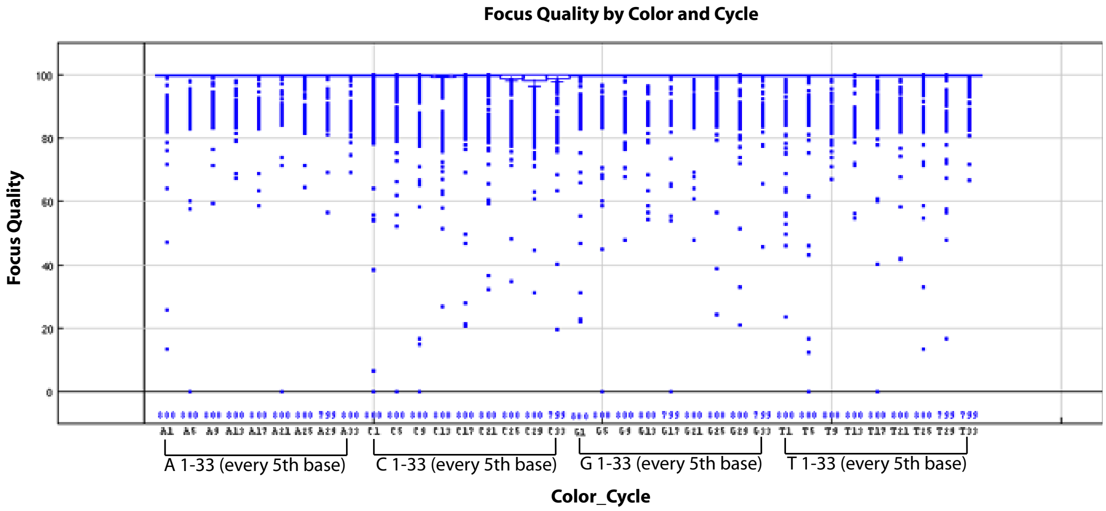


Fig. 7.4. Illumina Genome Analyzer focus quality by fluorescent emission and cycle.

It is equally important to check the fluorescent signal intensity produced by sequencing templates incorporated with CF-NRTs to assess the quality of the sequencing run. A plot of fluorescent intensity versus base during every 5th cycle was produced below for our analysis (Fig. 7.5). For each CF-NRT (A, C, G, T), the fluorescent signal intensity remained relatively constant and uniform throughout the entire 36 cycles of the sequencing process. The expected drop of intensity for the later cycles due to signal de-phasing (Chapter 5) was also evident. However, the signal drops were minimal across all four bases, indicating a high-quality SBS process.

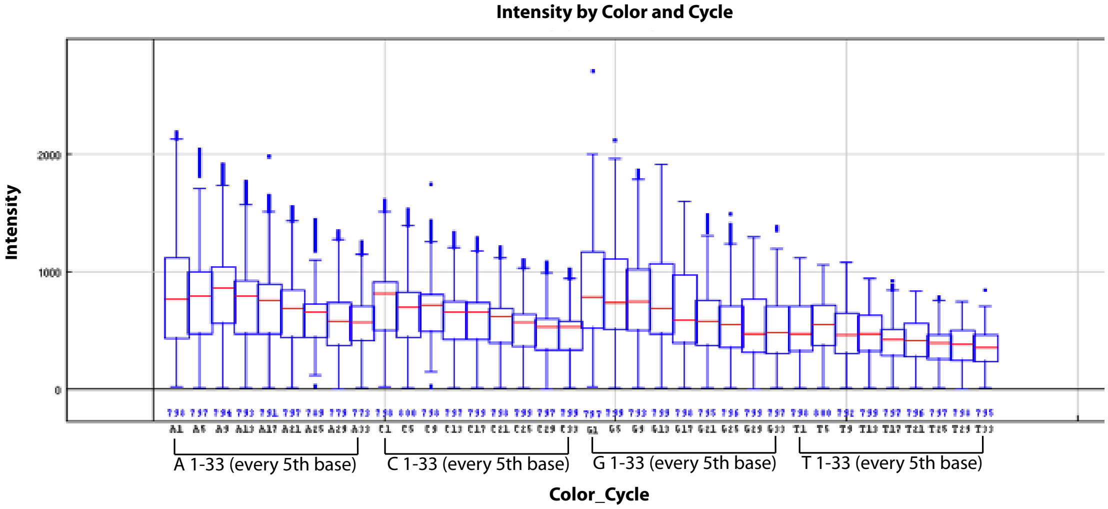


Fig. 7.5. Illumina Genome Analyzer fluorescent intensity of each base for selected sequencing cycles.

Overall, each lane generated an average of 10 million clusters. With a read length of 36 bases, a total of 360,000 kb (kilobases) of raw sequence was produced from each lane. Out of all the bases sequenced, ~70% passed the internal filter for data quality control. Hence, approximately 1.6 Gb (gigabases) of reliable sequencing data were produced from a single run in the IGA. We conducted several runs varying conditions for 5HT and FMRFamide treatments, and organized the sequencing data for analysis. We selected 247 different mRNAs displaying the largest differences in frequencies (counts) between treated and control samples. This differentially expressed gene list revealed many new genes that could be targeted by serotonin in addition to previously identified 5HT-dependent genes. The list includes aPKC (atypical protein kinase C), which is an evolutionarily conserved regulator of cell polarity and axon determinant,[11] and UCH-L1 (Ubiquitin carboxy-terminal hydrolase L1), a transcription factor that is highly specific to neurons. Another differentially expressed gene, HCH-L1, has been associated with Huntington and Alzheimer disease, and is required for normal synaptic and cognitive function. A point mutation in this gene is implicated as the cause of Parkinson's disease.

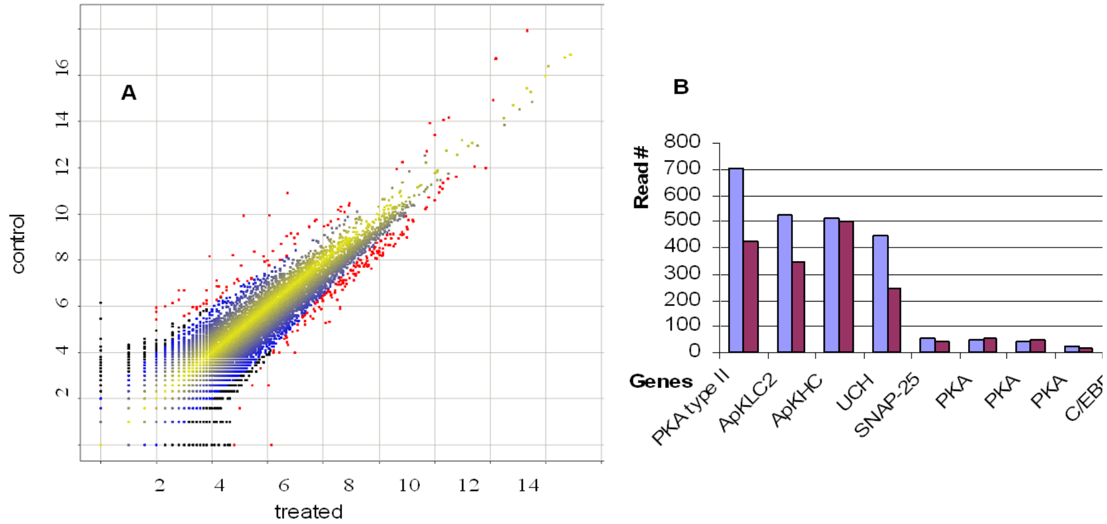


Fig. 7.6. (A) Scatter-plot reflecting differences in gene expression levels in 5HT-treated and intact (control) ganglia. Shown on the plot are the frequencies of gene tags which were log-transformed at base 2 and normalized using *Lowess* normalization. In yellow are tags corresponding to genes with lower than 2-fold changes in expression level. Shown in red are 247 gene tags with maximum expression level changes (LogRatio ranging from +4 to -6) that were chosen for further study. (B) Examples of differentially expressed genes: gene tag count in treated (blue) *vs* control cells (red), courtesy of Drs. Sergey Kalachikov and Irina Morozova.

7.4 Materials and Methods

General Information. All solvents and reagents were reagent grade, purchased commercially and used without further purification. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Sequencing consumables for the Illumina Genome Analyzer II, such as Cluster Generation Kit V2 and SBS Sequencing Kit V3, were perchased from Illumina. Sequencing data analysis, including base calling, quality filter, and sequence alignment, was carried out using Illumina's Pipeline V1.2 software.

7.4.1. Gene transcript analysis in *Aplysia* neuronal cells using Illumina Genome Analyzer

The cDNA library preparation of the sensory clusters of *Aplysia* pleural ganglia was accomplished in the Kandel lab by Dr. Sathyanarayan Puthanveettil. 36 sensory clusters were treated with either 5HT or FMRFamide. Total RNA was isolated for cDNA library generation. The libraries were then purified and quantified by Dr. Sergey Kalachikov prior to sequencing in the IGA. Seven *Aplysia* libraries, three treated with 5HT, three treated with FMRFamide, and one untreated, were each immobilized in one lane of the sequencing flow cell that was included in the Illumina Cluster Generation Kit V2. The flow cell was loaded onto the cluster generation station for template amplification, linearization, and sequencing primer hybridization. Upon completion of the cluster generation process, the flow cell was placed into the IGA for 36 cycles of sequencing using the SBS Sequencing Kit V3. The resulting sequencing data were analyzed using Illumina's Pipeline V1.2 software under the direction of Dr. Irina Morozova. The Pipeline software is composed of three algorithms for data processing. The Firecrest algorithm performs image analysis by extracting and quantifying clusters. Then the Bustard program uses a normalized matrix based on the control lane to call the bases. Once the sequences of the templates were determined, Gerald, the third component of the Pipeline, completes the analysis by aligning these sequences to target genomes using either PhageAlign or ELAND algorithms.

7.5 Conclusion

The four-color SBS using CF-NRTs platform, first developed and reported in our lab, has been successfully used in biological applications. The IGA sequencer, based on this SBS CF-NRT chemistry, has already seen its applications in a wide range of genomic and genetic projects. Equipped with this powerful technology, we embarked on tackling a real biological problem using next generation sequencing (NGS). Most existing technologies can only measure the properties of a population of cells and not those of individual cells. The development of novel approaches is necessary to use single cells to generate data to understand the molecular phenotype of a particular cell type and the role it plays in tissue and organ function. In the study of the memory formation mechanism, no large-scale neuron-specific or any other transcriptome project at the level of characterized neural circuits has been performed so far. We took advantage of the massive throughput of NGS, and successfully tested the applicability of IGA for single neuron types. The resulting sequencing data from our digital gene expression studies using IGA will enable large-scale comparisons of gene expression profiles between individual neurons representing operational circuits as a function of learning and memory formation. In the near future, we will be able to sequence all RNAs from a given single cell and even its subcompartments. This will lead to the most direct and unbiased way to characterize a neuronal transcriptome (i.e. simultaneously identify and count different RNA species in a single cell).

7.6 References

1. Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; Smith, G. P. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456,* 53-59.
2. Freeman, J. D.; Warren, R. L.; Webb, J. R.; Nelson, B. H.; Holt, R. A. Profiling the T-cell reseptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **2009**, *19*, 1817-1824.
3. Martin, K. C.; Casadio, A.; Zhu, H.; Yaping, E.; Rose, J. C.; Chen, M.; Bailey,

C. H.; Kandel, E. R. Synapse-specific, long-term facilitation of aplysia sensory to motor synapses: a function for local protein synthesis in memory storage. *Cell* **1997**, *91*, 927-938.

4. Kandel, E. R. The molecular biology of memory storage: a dialogue between genes and synapses. *Science* **2001**, *294*, 1030-1038.
5. Puthanveettil, S. V.; Monje, F. J.; Miniaci, M. C.; Choi, Y. B.; Karl, K. A.; Khandros, E.; Gawinowicz, M. A.; Sheetz, M.P.; Kandel, E.R. (2008) A new component in synaptic plasticity: upregulation of kinesin in the neurons of the gill-withdrawal reflex. *Cell* **2008**, *135*, 960-73.
6. Moroz, L. L.; Edwards, J. R.; Puthanveettil, S. V.; Kohn, A. B.; *et al.* Neuronal transcriptome of aplysia: neuronal compartments and circuitry. *Cell* **2006**, *127*, 1453-1467.
7. Alberini, C. M.; Ghirardi, M.; Metz, R.; Kandel, E. R. (1994) C/EBP is an immediate-early gene required for the consolidation of long-term facilitation in Aplysia. *Cell* **1994**, *76*, 1099-114.
8. Si, K.; Giustetto, M.; Etkin, A.; Hsu, R.; Janisiewicz, A. M.; Miniaci, M. C.; Kim, J. H.; Zhu, H.; Kandel, E. R. A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in aplysia. *Cell* **2003**, *115*, 893-904.
9. Sun, Z. Y.; Wu, F.; Schacher, S. J. Rapid bidirectional modulation of mRNA expression and export accompany long-term facilitation and depression of Aplysia synapses. *Neurobiol.* **2001**, *46*, 41-47.
10. Guan, Z.; Giustetto, M.; Lomvardas, S.; Kim, J. H.; Miniaci, M. C.; Schwartz, J. H.; Thanos, D.; Kandel, E. R. (2002) Integration of long-term-memory-related synaptic plasticity involves bidirectional regulation of gene expression and chromatin structure. *Cell* **2002**, *111*, 483-493.
11. Lee, Y. S.; Choi, S. L.; Kim, T. H.; Lee, J. A.; *et al.* Transcriptome analysis and identification of regulators for long-term plasticity in Aplysia kurodai. *Proc Natl Acad Sci U S A* **2008**, *105*, 18602-18607.

Chapter 8: Summary and Future Outlook

The ultimate goal of this thesis was to contribute to the development and improvement of a chip-based four-color DNA sequencing by synthesis (SBS) platform using fluorescent nucleotide reversible terminators (CF-NRTs) for genome sequencing. The following specific goals, which were set to test and implement key components of the SBS platform, were successfully achieved: (1) design and evaluation of a new chemical moiety for nucleotide modification; (2) design, synthesis, and analysis of two complete sets (A, C, G, and T) of CF-NRTs for the polymerase reaction in SBS; (3) optimization of conditions and processes for extension and cleavage reactions during SBS; (4) design and integration of a novel template walking strategy for SBS to increase sequence read length.

8.1 Exploration of a New Chemical Moiety for Nucleotide Reversible Terminator Modification in DNA Sequencing by Synthesis [1]

We explored the potential of using an azido group as a chemical moiety for nucleotide modification. Based on our established rationale for nucleotide reversible terminator (NRT) design, we synthesized a complete set of NRTs capped at the 3' position with an azidomethyl group (3'-O-N_3-dATP, 3'-O-N_3-dCTP, 3'-O-N_3-dGTP, 3'-O-N_3-dTTP). Through testing and optimization, it was apparent that these NRTs were good substrates of a DNA polymerase. We worked out a chemical cleavage condition to remove the azidomethyl group capping the 3'-OH of the nucleotide analogues under conditions that were compatible with DNA. We also performed four continuous cycles of DNA polymerase reactions incorporating each NRT in sequential order in order to validate the use of the NRTs in SBS.

8.2 Design, Synthesis, and Evaluation of a Novel Class of Cleavable Fluorescent Nucleotide Reversible Terminators Containing Substituted 2-Azidomethyl Benzoic Acid Linker for DNA Sequencing by Synthesis [2]

We have designed and synthesized four novel CF-NRTs: 3'-N_3-*O*-dATP-azidomethylbenzoyl-ROX, 3'-N_3-*O*-dGTP-azidomethylbenzoyl-Cy5, 3'-N_3-*O*-dCTP-azidomethylbenzoyl-Bodipy-FL-510, and 3'-N_3-*O*-dUTP-azidomethylbenzoyl-R6G for applications in SBS. These CF-NRTs were capped at the 3'-OH with an azidomethyl group identical to the NRTs and contained a substituted 2-azidomethylbenzoyl linker to tether a fluorophore. These CF-NRTs were used to produce four-color *de novo* DNA sequencing data on a chip based on our sequencing by synthesis approach. After each round of sequencing, both the fluorophores linked to the CF-NRTs and the 3'-azidomethyl group on the DNA extension products generated by incorporating 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores were removed using a TCEP cleavage solution. This one step dual cleavage process for reinitiating the polymerase reaction increased the overall SBS efficiency. Since 3'-O-N_3-dNTP-azidomethylbenzoyl-fluorophores are reversible terminators, which allow the sequencing of each base in a serial manner, they can accurately determine the homopolymeric regions of DNA. In addition, due to the fact that all of the steps of our SBS approach are performed on a DNA chip, there is no longer a need for electrophoretic DNA fragment separation as in the classical Sanger sequencing method.

8.3 Four-color DNA Sequencing by Synthesis (SBS) Improvements Using Cleavable Fluorescent Nucleotide Reversible Terminators [3]

Another complete set of CF-NRTs (3’-O-N_3-dNTP-N_3-fluorophores) along with four non-fluorescent NRTs (3’-O-N_3-dNTPs) have been synthesized for the implementation of our four-color *de novo* DNA sequencing by synthesis in order to seek improvements over existing SBS approaches. During the incorporation stage of SBS, a mixture of both sets of modified nucleotide analogues was used to simultaneously extend the primer strand of various target DNA linear templates. This approach led to a more efficient DNA polymerase reaction since the smaller 3’-O-N_3-dNTPs were much easier to incorporate. Moreover primers extended with NRTs resembled nascent strands of DNA that had no traces of modification after cleavage of the 3’-azidomethyl capping group. Such DNA strands would not have any adverse effect on the DNA polymerase during later cycles of sequencing reactions. After the incorporation reaction, two separate capping steps, first with 3’-O-N_3-dNTPs and then with ddNTPs, were performed to synchronize all the templates on the surface. Without these precautionary synchronization procedures, mixed fluorescent signals would prevent the identification of the correctly incorporated nucleotide. Hence, we have successfully addressed one of the key drawbacks of SBS, which was the mis-calling of the base due to lagging signals. In addition, since both 3’-O-N_3-dNTP-N_3-fluorophores and 3’-O-N_3-dNTPs were reversible terminators, which allow the sequencing of each base in a serial manner, they could accurately determine the homopolymeric regions of DNA.

8.4 Exploration of Novel Primer Resetting Strategies to Extend Read Length for DNA Sequencing by Synthesis [4]

We developed a novel template walking strategy to increase read length for DNA Sequencing by Synthesis. The template walking method involved extending the sequencing primer with three natural nucleotides and one NRT so that the polymerase reaction was temporarily paused when the NRT was incorporated. Upon restoring the 3'-OH group of the NRT incorporated into the primer via a cleavage reaction, further cycles of walking could be carried out until the entire previously sequenced portion of the template was skipped. We have successfully demonstrated the integration of this template walking strategy into our four-color DNA SBS platform by performing one round of SBS, four cycles of template walking reactions, and then a second round of SBS. Through this effort, we were able to sequence a linear DNA template in its entirety, nearly doubling the read length of our previous sequencing results.

8.5 Future Outlook for 4-color DNA Sequencing by Synthesis using CF-NRTs [5, 6]

The field of DNA sequencing technology has been rapidly advancing recently due to its important role in genetic and genomic studies. In terms of read length, Sanger sequencing remains the gold standard but is limited in its throughput and cost. The major advantage of our four-color SBS approach compared to the Sanger sequencing method is the high throughput and simplicity of the SBS platform in which all sequencing is conducted on a chip without any separation steps. The ultimate read length of fluorescent SBS depends on three factors: the number of starting DNA molecules on each spot of a DNA chip, the nucleotide incorporation and cleavage efficiency, and the sensitivity of the detection system. The read length with the Sanger sequencing method commonly reaches more than 700 base pairs. The fluorescent SBS system (with current read length at ~50 base pairs) has the potential to approach this read length if stepwise yield can reach over 99% and the sensitivity of the fluorescent detection system can be further improved (Fig. 8.1).

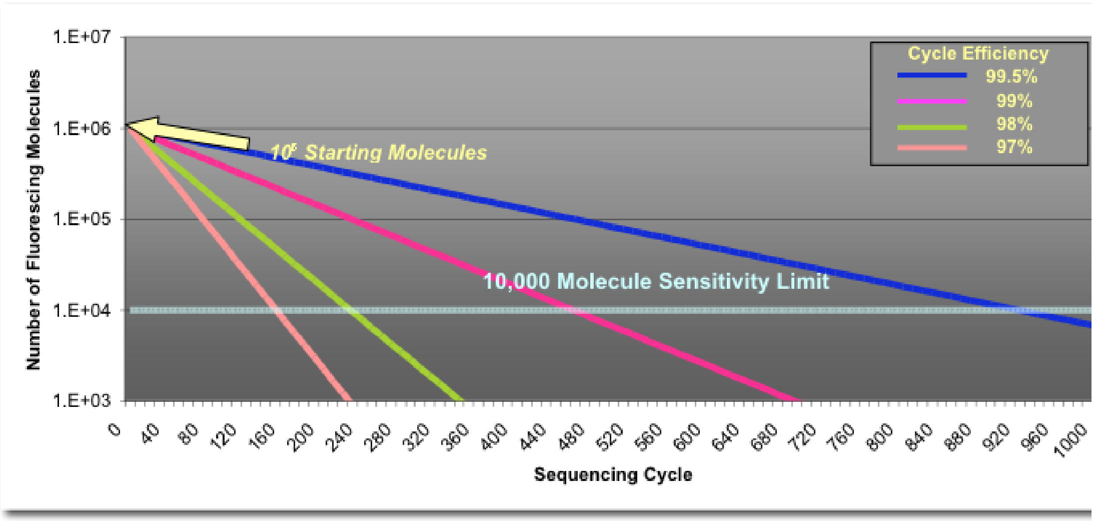


Fig. 8.1. Theoretical SBS read length based on sequencing cycle efficiency

At present, three platforms are in widespread use for massively parallel DNA sequencing: the Roche/454 FLX Pyrosequencer,[7] the Illumina Genome Analyzer,[8] and the Applied Biosystems SOLiD^{TM} System.[9] The Roche/454 Pyrosequencer and the Applied Biosystems SOLiD sequencer use emulsion PCR for DNA template preparation,[10] while bridge PCR amplification is used by the Illumina Genome Analyzer.[8, 11] Recently, two platforms for single molecule DNA sequencing by synthesis have been reported. For the Helicos Heliscope sequencer,[12] single-molecule DNA templates are directly immobilized on a solid surface to allow SBS to take place. Pacific Biosciences SMRT single molecule sequencing system[13] captures the polymerase rather than the DNA library on the surface to carry out real-time sequencing determination. With these next generation DNA sequencing systems, massively parallel digital gene expression analogous to a high-throughput SAGE approach has been reported reaching single copy transcript sensitivity,[14] and CHIP-Seq[15-17] based on sequencing tags of around 25 bases has led to many new discoveries in genome function and regulation. In an ongoing project to explore the molecular mechanism of long-term memory formation, we have already taken advantage of the massive throughput of the Illumina Genome Analyzer that is based on our SBS technology to conduct digital gene expression study of *Aplysia* central nervous system. It is well established that millions of different DNA templates can be generated on a solid surface to construct massively parallel DNA sequencing platforms. Therefore, future implementation of the molecular level SBS approaches on a high-density DNA array will provide a high-throughput and accurate DNA sequencing system with wide applications in genome biology and biomedical research.

8.6 References

1. Guo, J.; Xu, N.; Li, Z.; Zhang, S.; Wu, J.; Kim, D. H.; Marma, M. S.; Meng, Q.; Cao, H.; Li, X.; Shi, S.; Yu, L.; Kalachikov, S.; Russo, J. J.; Turro, N. J.; Ju, J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* **2008**, *105*, 9145-4150.

2. Cao, H.; Yu, L.; Meng, Q.; Li, Z.; Ju, J. Design, synthesis and evaluation of a class of fluorescent nucleotides containing substituted 2-azidomethyl benzoic acid linker as reversible terminators for DNA sequencing by synthesis. (2009) *In preparation*.

3. Yu, L.; Guo, J.; Qui, C.; Li, Z.; Kim, D. H.; Cao, H.; Zhang, S.; Meng, Q.; Marma, M. S.; Wu, J.; Xu, N.; Li, X.; Shi, S.; Kalachikov, S.; Russo, J. J.;

Turro, N. J.; Ju, J. Four-color DNA sequencing by synthesis (SBS) improvements with cleavable fluorescent nucleotide reversible terminators. (2009) *Submitted, under review*.

4. Yu, L.; Qui, C.; Guo, J.; Kalachikov, S.; Li, Z.; Xu, N.; Li, X.; Shi, S.; Russo, J. J.; Turro, N. J.; Ju, J. Novel template walking strategy for read length increment in DNA sequencing by synthesis (2009) *In preparation*.

5. Guo, J.; Yu, L.; Turro, N. J.; Ju, J. An integrated system for DNA sequencing by synthesis using novel nucleotide analogues. *Accounts for Chemical Research*. *Accepted*.

6. Yu, L.; Guo, J.; Xu, N.; Li, Z.; Ju, J. DNA sequencing by synthesis using novel nucleotide analogues. *Submitted*.

7. Wheeler, D. A.; Srinivasan, M.; Egholm, M.; *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **2008**, *452,* 872-877.

8. Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008,** *456,* 53-59.

9. McKernan, K. J.; Peckham, H. E.; Costa, G. L.; *et al*. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Research* **2009,** *19*, 1527-1541.

10. Margulies, M.; Egholm, M.; Altman, W. E.; *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005,** *437,* 376-380.

11. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum.Genet*. **2008**, *9*, 387-402.

12. Harris, T. D.; Buzby, P. R.; Babcok, H.; Beer, E.; Bowers, J.; Braslavsky, I.; Causey, M.; Colonell, J.; DiMeo, J.; Efcavitch, J. W.; Giladi, E.; Gill, J.; Healy, J.; Jarosz, M.; Lapen, D.; Moulton, K.; Quake, S. R.; Steinmann, K.; Thayer, E.;

Tyurina, A.; Ward, R.; Weiss, H.; Xie, Z. Single-molecule DNA sequencing of a viral genome. *Science* **2008**, *320*, 106-109.

13. Eid, J.; Fehr, A.; Gray, J.; *et al*. Real-time DNA sequencing from single polymerase molecule. *Science* **2008**, *323*, 133-138.

14. Kim, J. B.; Porreca, G. J.; Song, L.; Greenway, S. C.; Gorham, J. M.; Church, G. M.; Seidman, C. E.; Seidman, J. G. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **2007**, *316*, 1481-1484.

15. Mikkelson, T. S.; Ku, M.; Jaffe, D. B.; Issac, B.; Lieberman, E.; Giannoukos, G.; Alvarez, P.; Brockman, W.; Kim, T.; Koche, R. P.; Lee, W.; Mendenhall, E.; O'Donovan, A.; Presser, A.; Russ, C.; Xie, X.; Meissner, A.; Wernig, M.; Jaenisch, R.; Nusbaum, C.; Lander, E. S.; Bernstein, B. E. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **2007**, *448*, 553-560.

16. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*,1497-1502.

17. Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.; Schones, D.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **2007**, *129*, 823-837.