Novel Molecular Engineering Approaches for Genotyping and DNA Sequencing

Chunmei Qiu

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

COLUMBIA UNIVERISTY

2011

ABSTRACT

Novel Molecular Engineering Approaches for Genotyping and DNA Sequencing

Chunmei Qiu

The completion of the Human Genome Project has increased the need for investigation of genetic sequences and their biological functions, which will significantly contribute to the advances in biomedical sciences, human genetics and personalized medicine.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) offers an attractive option for DNA analysis due to its high accuracy, sensitivity and speed. In the first part of the thesis, we report the design, synthesis and evaluation of a novel set of mass tagged, cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins) for DNA polymerase extension reaction and its application in DNA sequencing and single nucleotide polymorphism (SNP) genotyping by mass spectrometry. These nucleotide analogs have a biotin moiety attached to the 5 position of the pyrimidines (C and U) or the 7 position of the purines (A and G) via a chemically cleavable azido-based linker, with different length linker arms serving as mass tags that contribute large mass differences among the nucleotides to increase resolution in MS analysis. It has been demonstrated that these modified nucleotides can be efficiently incorporated by DNA polymerase, and the DNA strand bearing biotinylated nucleotides

can be captured by streptavidin coated beads and efficiently released using tris(2-carboxyethyl) phosphine in aqueous solution which is compatible with DNA and downstream procedures. Reversible solid phase capture (SPC) mass spectrometry sequencing using ddNTP-N_3-biotins was performed, and various DNA templates, including biological samples, were accurately sequenced achieving a read-length of 37 bases. In mass spectrometric SNP genotyping, we have successfully exploited our reversible solid phase capture (SPC)-single base extension (SBE) assay and been able to detect as low as 2.5% heteroplasmy in mitochondrial DNA samples, with interrogation of human mitochondrial genome position 8344 which is associated with an important mitochondrial disease (myoclonic epilepsy with ragged red fibers, MERRF); we have also quantified the heteroplasmy level of a real MERRF patient and determined several mitochondrial MERRF mutations in a multiplex approach. These results demonstrated that our improved mass spectrometry genotyping technologies have great potential in DNA analysis, with particular applications in sequencing short-length targets or detecting SNPs with high accuracy and sensitivity requirements, such as DNA fragments with small indels, or SNPs in pooled samples. To truly implement this mass spectrometry-based genotyping method, we further explored the use of a lab-on-a-chip microfluidic device with the potential for high throughput, miniaturization, and automation. The microdevice primarily consists of a micro-reaction chamber for single base extension and cleavage reactions with an integrated micro heater and temperature sensor for on-chip temperature control, a microchannel loaded with streptavidin magnetic

beads for solid phase capture, and a microchannel packed with C18-modified reversed-phase silica particles as a stationary phase for desalting before MALDI-TOF analysis. By performing each functional step, we have demonstrated 100% on-chip single base incorporation, sufficient capture and release of the biotin-ddNTP terminated single base extension products, and high sample recovery from the C18 reverse-phase microchannel with as little as 0.5 pmol DNA molecules. The feasibility of the microdevice has shown its promise to improve mass spectrometric DNA sequencing and SNP genotyping to a new paradigm.

DNA sequencing by synthesis (SBS) appears to be a very promising molecular tool for genome analysis with the potential to achieve the $1000 Genome goal. However, the current short read-length is still a challenge. Therefore, the second part of the thesis focuses on strategies to overcome the short read-length of SBS. We developed a novel primer walking strategy to increase the read-length of SBS with cleavable fluorescent nucleotide reversible terminators (CF-NRTs) and nucleotide reversible terminators (NRTs) or hybrid-SBS with cleavable fluorescent nucleotide permanent terminators and NRTs. The idea of the walking strategy is to recover the initial template after one round of sequencing and re-initiate a second round of sequencing at a downstream base to cover more bases overall. The combination of three natural nucleotides and one NRT effectively regulated the primer walking: the primer extension temporarily paused when the NRT was incorporated, and resumed after removing the 3’ capping group to restore

the 3'-OH group. We have successfully demonstrated the integration of this primer walking strategy into the sequencing by synthesis approach, and were able to obtain a total read-length of 53 bases, nearly doubling the read-length of the previous sequencing. On the other hand, we explored the sequencing bead-on-chip approach to increase the throughput of SBS and hence the total genome coverage per run. The various prerequisite conditions have been optimized, allowing the accurate sequencing of several bases on the bead surface, which demonstrated the feasibility of this approach. Both of these approaches could be integrated into current SBS platforms, allowing increased overall coverage and lowering overall costs.

As a step beyond genotyping, the *in vivo* visualization of biomolecules, like DNA and its encoded RNA and proteins, provides further information about their biological functions and mechanisms. The third part of the thesis focuses on the development of a novel quantum dot (QD)-based binary molecular probe, which takes advantage of fluorescent resonance energy transfer (FRET), for detection of nucleic acids, aiming at their eventual use for detection of mRNAs involved in long term memory studies in the model organism *Aplysia californica*. We reported the design, synthesis, and characterization of a binary probe (BP) that consists of carboxylic quantum dot (CdSe/ZnS core shell)-DNA (QD-DNA) conjugated donor and a cyanine-5 (Cy5)-DNA acceptor for the detection of a sensorin mRNA-based synthetic DNA molecule. We have demonstrated that in the absence of target DNA, the QD fluorescence is the main signal observed (605 nm); in the

presence of the complementary target DNA sequence, a decrease of QD emission and an increase of Cy5 emission at 667 nm was observed. We have demonstrated the distance dependence of FRET, with the finding that the target with 16 base separation between the QD and Cy5 after probe hybridization gave the most efficient FRET. Further studies are in progress to evaluate the effectiveness of this QD-based probe inside a cell extract and in living cells.

Table of Contents

List of Figures and Tables x

Acknowledgements xx

Abbreviations and Symbols xxiii

Chapter 1 Introduction to Genomic Analysis Technologies -- DNA sequencing, Genotyping, Nucleic Acid Detection 1

1.1 Introduction to DNA sequencing technology 2

1.1.1 Background and significance 2

1.1.2 DNA sequencing technologies overview 4

1.1.2.1 Conventional Sanger Sequencing 5

1.1.2.2 Mass Spectrometry based sequencing 6

1.1.2.3 Sequencing by hybridization 9

1.1.2.4 Sequencing by synthesis 10

1.1.2.4.1 Pyrosequencing 11

1.1.2.4.2 Sequencing by ligation 13

1.1.2.4.3 Sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators 15

1.1.2.4.4 Single molecule sequencing-by-synthesis 17

1.1.2.5 Sequencing by direct physical recognition of the DNA molecule 22

1.1.2.5.1 Nanopore sequencing.....22

1.1.2.5.2 Tunneling and transmission electron microscopy based single molecule DNA sequencing 26

1.1.3 Conclusion 28

1.2 Introduction to SNP genotyping technology.....29

1.2.1 Background and Significance 29

1.2.2 Overview of SNP genotyping technologies 30

1.2.2.1 SNP genotyping by enzymatic cleavage.....31

1.2.2.2 SNP genotyping by allele specific hybridization.....33

1.2.2.3 SNP genotyping by allele specific ligation.....36

1.2.2.4 SNP genotyping by allele specific primer extension 39

1.2.3 Conclusion 43

1.3 Introduction to the detection of nucleic acids by oligonucleotide probes 44

1.3.1 Background and significance.....44

1.3.2 Overview of oligonucleotide probes for detection of nucleic acids 45

1.3.2.1 Molecular beacons 45

1.3.2.2 Binary oligonucleotide probes 48

1.3.3 Conclusion 49

References.....50

Part I Mass Spectrometric DNA Sequencing and SNP Genotyping with Cleavable Biotinylated Dideoxynucleotides

Chapter 2 Design and Synthesis of Cleavable Biotinylated Dideoxynucleotides for DNA Sequencing and Genotyping by MALDI-TOF Mass Spectrometry 61

2.1 Introduction 61

2.2 Experimental Rationale and Overview 65

2.3 Results and Discussion 67

2.3.1 Design and synthesis of cleavable biotinylated dideoxynucleotides 67

2.3.2 Polymerase extension using ddNTP-N_3-biotins, solid phase capture and cleavage 70

2.3.3 Comparison with non-cleavable biotinylated dideoxynucleotides 75

2.4 Materials and Methods 77

2.4.1 Synthesis of ddNTP-N_3-biotins 77

2.4.2 Polymerase extension using ddNTP-N_3-biotins, solid phase capture and cleavage 82

2.4.3 Comparison of ddATP-N3-biotin and biotin-11-ddATP 83

2.5 Conclusion 84

References 85

Chapter 3 DNA Sequencing by MALDI-TOF Mass Spectrometry using Cleavable Biotinylated Dideoxynucleotides 88

3. 1 Introduction 88

3.2 Experimental Rationale and Overview 94

3.3 Results and Discussion 96

3.3.1 DNA sequencing on synthetic template 96

3.3.2 DNA sequencing on biological template 97

3.4 Materials and Methods 99

3.4.1 Sanger DNA sequencing reaction 99

3.4.2 PCR reactions for generating biological template 100

3.4.3 DNA sequencing on biological template 101

3.4.4 Solid-phase purification of DNA-sequencing products for mass spectrometry measurements 101

3.5 Conclusion 102

References 103

Chapter 4 SNP Genotyping of Mitochondrial DNA by MALDI-TOF MS using Cleavable Biotinylated Dideoxynucleotides 104

4.1 Introduction 104

4.2 Experimental Rationale and Overview 108

4.3 Results and Discussion 111

4.3.1 PCR amplification of targeted region in mitochondrial DNA 111

4.3.2 A8344G uniplex genotyping 113

4.3.3 5-plex genotyping 116

4.3.4 Quantitative mutation analysis for mitochondrial sample 118

4.3.5 Sensitivity of SPC-SBE MALDI-TOF MS for detecting low heteroplasmy of mitochondrial DNA 121

4.4 Materials and Methods....125
4.4.1 PCR amplification....125
4.4.2 Quantification of mtDNA in the samples....126
4.4.3 Single base extension using cleavable biotinylated dideoxynucleotides for MALDI-TOF MS detection....128
4.4.4 Direct Sanger DNA Sequencing and PCR-RFLP Assay....129
4.5 Conclusion....130
References....131
Chapter 5 Exploration of the Integrated Microdevice for SNP Genotyping by MALDI-TOF Mass Spectrometry....135
5.1 Introduction....135
5.2 Experiment Rationale and Overview....137
5.3 Results and Discussion....141
5.3.1 Temperature Sensor Calibration....141
5.3.2 On-chip testing....142
5.3.2.1 On-chip single base extension....143
5.3.2.2 On-chip solid phase capture....144
5.3.2.3 On-chip desalting....145
5.4 Materials and Methods....146
5.4.1 Microfluidic Device fabrication....147
5.4.2 Experimental setup....149

5.4.3 Microdevice performance testing 151

5.4.3.1 Temperature Sensor Calibration.....151

5.4.3.2 Single base extension.....151

5.4.3.3 Solid phase capture and cleavage test.....152

5.4.3.4 C18 reversed phase desalting.....153

5.5 Conclusion 153

References.....154

Part II Strategies to Improve Sequencing by Synthesis with Cleavable Fluorescent Nucleotide Reversible Terminators

Chapter 6 Development of Primer "Walking" Strategy to Increase the Read-Length of Sequencing by Synthesis.....157

6.1 Introduction.....157

6.2 Experimental Rationale and Overview 168

6.3 Results and Discussion 171

6.3.1 Optimization of primer annealing to DNA template on solid surface 171

6.3.2 Sequencing by synthesis on linear DNA template.....173

6.3.2 Primer resetting and walking for extending read-length.....175

6.3.2.1 Primer walking using three natural nucleotides.....176

6.3.2.2 Primer walking using three natural nucleotides and one reversible nucleotide terminator 182

6.4 Materials and Methods.....187

6.4.1 Primer hybridization 189

6.3.2 Primer resetting and walking 191

6.3.2.1 Primer walking using three natural nucleotides 192

6.3.2.2 Primer walking using three natural nucleotides and one reversible nucleotide terminator 193

6.4 Conclusion 195

References 196

Chapter 7 Exploration of an "Emulsion PCR-Bead-on-Chip" Approach to Improve the Throughput of Sequencing by Synthesis 198

7.1 Introduction 198

7.2 Experiment Rationale and Overview 201

7.3 Results and Discussion 203

7.3.1 Covalent attachment of DNA onto the beads 203

7.3.2 Biological affinity attachment of DNA onto the beads 204

7.3.3 Nucleotide incorporation studies on bead surfaces 206

7.3.3.1 Pyrosequencing 206

7.3.3.2 Incorporation of fluorescence labeled ddNTPs and natural dNTPs 207

7.3.3.3 Sequencing by synthesis on beads using CF-NRTs/NRTs 208

7.3.4 PCR on beads in solution and beads in emulsion 210

7.4 Materials and Methods 212

7.4.1 DNA attachment to beads 212

7.4.2 Nucleotide incorporation test on beads.....215

7.4.3 Sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators on beads 216

7.4.4 PCR on beads and emulsion-beads in droplet.....218

7.5 Conclusion 219

References.....220

Part III Detection of Nucleic Acids with Molecular Probes

Chapter 8 Quantum Dot based FRET Binary Probes for Detection of Nucleic Acids223

8.1 Introduction.....223

8.2 Experimental Rationale and Overview 226

8.3 Results and Discussion 228

8.3.1 The synthesis of QD-DNA conjugates.....228

8.3.2 Hybridization kinetics studies.....230

8.3.3 Comparison of FRET efficiency between carboxyl-QD-DNA conjugate and streptavidin-QD-DNA conjugate based binary probes 232

8.3.4 Distance dependent FRET studies with carboxyl-QD binary probes .233

8.4 Materials and Methods.....236

8.4.1 Synthesis of QD-DNA conjugates 237

8.4.2 Hybridization of QD-DNA and Cy5-DNA with different targets.....238

8.4.3 Steady-state fluorescence and time-resolved fluorescence measurement239

8.5 Conclusion 239

References.....240

Chapter 9 Summary and Future Outlook.....242

9.1 Mass spectrometric DNA sequencing and SNP genotyping with cleavable biotinylated dideoxynucleotides[1, 2] 242

9.2 Strategies to improve sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators[3].....243

9.3 Detection of nucleic acids with molecular probes[4].....244

9.4 Future outlook.....245

9.4.1 Mass spectrometric DNA sequencing and SNP genotyping.....245

9.4.2 DNA sequencing by synthesis 246

9.4.3 In vivo visualization of nucleic acids.....247

References……..…….……248

List of Figures and Tables

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

Fig. 1.2. Principle of Sanger sequencing. 6

Fig. 1.3. Schematic process of mass spectrometry assisted sequencing. (A) Sanger sequencing reaction based MS sequencing; (B) Enzymatic digestion based MS sequencing 8

Fig.1.4. Schematic view of pyrosequencing.[25] 11

Fig. 1.5. Principle of sequencing by ligation. (A) DNA template molecule with two mate-pair tags of unique sequence which are flanked and separated by universal sequences complementary to amplification or sequencing primers. (B) Steps of sequencing by ligation. 13

Fig. 1.6. Principle of sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators 17

Fig. 1.7 Schematic of the single molecule sequencing-by-synthesis approach with virtual terminators.[49] 19

Fig. 1.8 Principle of single-molecule, real-time DNA (SMRT) sequencing. [31] 21

Fig. 1.9. Schematic of various Nanopore sequencing method. (A) Fundamental principle of nanopore sequencing; (B) Nanopore sequencing approach by Oxford Technology;[65] (C) duplexes halting translocation sequencing by MspA nanopore;[66] (D) Nanopore single-molecule optical detection approach;[69] (E) Schematic of the DNA transistor. The light blue regions with voltage labeled are the conductors, while the light green regions are insulators.[67] 23

Fig. 1.10 (A) Reveo STM-based sequencing.[49] (B) DNA sequencing by direct inspection of DNA using electron microscopy [9] 27

Fig. 1.11. Schematic diagram of SNP enzymatic cleavage assay. (A)PCR-RFLP assay; (B) Invasive cleavage assay. 32

Fig. 1.12 Approaches of SNP genotyping by allele specific hybridization. (A) DNA microarray hybridization; (B) TaqMan assay.....35

Fig. 1.13 Methods for SNP genotyping by ligation. (A) Fluorescence based ligation assay; (B) Ligation-rolling circle amplification assay. 39

Fig. 1.14 Allele specific primer extension. (A) Fluorescence based detection; (B) Mass spectrometry based detection. 41

Fig. 1.15. Principle of oligonucleotide probes for detection of nucleic acid. (A) molecular beacons; (B) binary probes. 47

Fig. 2.1. Principle of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS).[7] 62

Fig. 2.2. Scheme of single base extension, solid phase capture and cleavage using chemically cleavable dideoxynucleotides.....66

Fig. 2.3. Structures of cleavable dideoxynucleotides (ddNTP-N_3-biotins). Note that the length of the portion of the linker between the base and the N_3 group, the segment which remains attached to the extended DNA after TCEP cleavage, varies between the two purine (A and G) and the two pyrimidine bases (C and U) bases, enabling clear discrimination of their sizes by MALDI-TOF mass spectrometry. 69

Fig. 2.4. MALDI-TOF mass spectra of the DNA extension products, their subsequent cleavage products in solution, and released DNA products from streptavidin coated magnetic beads. (A) primers extended with ddATP-N_3-biotin (1) (8889 m/z); (B) their cleavage products (2) (8416 m/z); (C) released products from solid phase (3) (8416 m/z); (D) primers extended with ddGTP-N_3-biotin (4) (9018 m/z); (E) their cleavage products (5) (8545 m/z); (F) released products from solid phase (6) (8545 m/z) (G) primers extended with ddCTP-N_3-biotin (7) (8866 m/z); (H) their cleavage products (8) (8393 m/z); (I) released products from solid phase (9) (8866 m/z); (J) primers extended with ddUTP-N_3-biotin (10) (8980 m/z); (K) their cleavage products (11) (8507 m/z); (L) released products from solid phase (12) (8507 m/z).....72

Fig. 2.5. Polymerase extension reaction using ddATP-N_3-biotin as a substrate and TCEP cleavage of DNA fragments containing ddA-N_3-biotin on streptavidin coated beads. The DNA polymerase Thermo Sequenase incorporates ddATP-N_3-biotin to generate the extension product (1). Cleavage by TCEP of

the DNA extension products captured on streptavidin-coated beads produces released products (3), while the biotin moiety remains on the solid surface of the beads.....73

Fig. 2.6. Staudinger reaction with TCEP to cleave the azido-based linker and release the DNA extension products from the streptavidin coated surface.....74

Fig. 2.7. Comparison of cleavable biotinylated dideoxynucleotide (ddATP-N_3-biotin) and uncleavable biotinylated dideoxynucleotide (biotin-11-ddATP). (A) Chemical structures of ddATP-N_3-biotin and biotin-11-ddATP); (B) the purification process for each nucleotide. Note: after cleavage, the biotin-11-ddATP incorporated DNA template requires an extra ethanol precipitation step; (C) mass spectrum of final product for starting template: a, 0.5 pmol, b, 1.0 pmol, c, 2.5 pmol. 76

Fig. 2.8. Structures of 5- or 7-propargylamino-ddNTPs.....78

Fig. 2.9. Synthesis and structures of Biotin-N_3-linker attached ddNTPs 81

Fig. 3.1. The mass spectra of mock A, C, G, and T sequencing reactions containing mixtures of synthetic oligonucleotides. (A) Individual spectra; (B) The spectra are overlaid and displayed on the same mass scale. [5] 90

Fig. 3.2. Solid-phase-capture (SPC) sequencing. (A) A SPC-sequencing scheme to isolate pure DNA fragments for MS analysis; (B) A DNA sequencing mass spectrum generated after extension with biotinylated terminators.[11].....92

Fig. 3.3. Scheme for purification of DNA-sequencing fragments for MALDI-TOF MS analysis. DNA-sequencing fragments are isolated from the sequencing solution containing excess primers, falsely stopped fragments and salts by streptavidin-coated magnetic beads. Then the sequencing fragments are cleaved from the beads with TCEP for MALDI-TOF MS analysis, leaving the biotin moiety still bound to the surface.....95

Fig. 3.4. Mass-sequencing spectrum generated using ddNTP-N_3-biotins on a synthetic template. The nested insets show increasing magnifications of the lower intensity region. 97

Fig. 3.5. Gel electrophoresis of PCR amplification within RHOD gene. The expected fragment size is 150 bp. 1. Low molecular weight marker; 2, PCR reaction products (+); 3, Negative control (no DNA template, -).....98

Fig. 3.6. Mass-sequencing spectrum generated using ddNTP-N_3-biotins on a PCR product. The nested insets show increasing magnifications of the lower intensity region. 99

Fig. 3.7. Sanger sequencing for RHOD gene fragment, the result of which is consistent with that obtained by the MS-based sequencing.....99

Fig. 4.1. The human mtDNA map, showing the location of selected pathogenic mutations within the 16,569-base pair genome. Genes are designated by the abbreviations outside the ring, and mitochondrial syndrome abbreviations and key mutation sites are displayed inside the ring (e.g., MERRF mutation at position 8344). 105

Fig. 4.2. The SPC-SBE approach for multiplex genotyping by MALDI-TOF MS using cleavable biotinylated dideoxynucleotides. DNA fragments that contain target SNP positions are generated by uniplex or multiplex PCR, and serve as templates for single base extension reactions. A set of SBE primers that are adjacent to SNP sites are used to generate single base extension products which are then isolated from the reaction mixture containing unextended primers, salts and other contaminants by streptavidin-coated magnetic beads. SBE products are then cleaved from the beads with TCEP in preparation for MALDI-TOF MS analysis, leaving the biotin moiety still bound to the bead surface. 110

Fig. 4.3. Gel electrophoresis of PCR condition testing. A. PCR on mitochondrial DNA at different annealing temperatures and template concentrations: 1. Low Molecular Weight (LMW) ladder, 2, 55^{o}C (100ng), 3, 58^{o}C (100ng), 4, 53^{o}C (100ng), 5, 55^{o}C (10ng); B. PCR on mitochondrial DNA-negative sample: 1. LMW ladder, 2, 100ng, 3, 10ng, 4, positive control (mitochondrial DNA sample, +), 5, negative control (no template, -) 112

Fig. 4.4. PCR amplification of region 1 containing MERRF SNP at mitochondrial genome position 8344 on different samples: channel 1, 8344 wild type sample (8344WT), 2, anonymous healthy donor (WT-2), 3, 8344 mutant sample (8344MT), 4 patient sample (PT), 5, mitochondrial DNA-negative cell line (Rho^0), 6, negative control (-), 7, 3255 mutant sample (3255MT), 8, negative control. 112

Fig. 4.5. 2-plex PCR on different samples: channel 1, mitochondrial DNA-negative (rho^o) cell line; 2, anonymous healthy donor (WT-2), 3, 8344 wild type sample (8344WT), 4, 8344 mutant sample (8344MT); 5 patient sample (PT); 6, 3255 mutant sample (3255MT); 7, negative control (-). 113

Fig. 4.6. Mass spectrometric detection of nucleotide variation at mtDNA locus 8344 from (a) 8344 wild-type sample; (b) anonymous healthy subject; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample; (f) mitochondrial DNA-negative cells. Letters in red refer to the mutant type. 114

Fig. 4.7. Sanger sequencing results on different samples. (a) 8344 wild-type sample; (b) anonymous healthy subject; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample. 116

Fig. 4.8. Mass spectrometric results of 5-plex SNP genotyping for MERRF syndrome. (a) 8344 wild-type sample; (b) anonymous healthy donor; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample; (f) mitochondrial DNA-negative cell line. Red labels refer to the mutated forms. 118

Fig. 4.9. Real-time PCR for quantitative analysis of A8344G wild type and mutant type samples. A. The Ct value versus the concentration of DNA for both MT and WT samples; B. Direct comparison between MT and WT using linear regression. 120

Fig. 4.10. PCR result for different levels of artificial heteroplasmy (%). 120

Fig. 4.11. Quantitative analysis of position 8344 heteroplasmy in mitochondrial DNA based on various mixture ratios of purified PCR products (A) and original biological samples (B). 121

Fig. 4.12. Detection of low 8344 heteroplasmy levels for sensitivity testing. (A) Detection of 5% and 2.5% heteroplasmy, based on original sample DNA mixing, by MALDI-TOF MS; (B) Detection of 5% and 2% heteroplasmy, based on purified PCR products mixing, by MALDI-TOF MS (A) and (B) shows the reproducible detection of <5% heteroplasmy. (C) Sanger sequencing analysis for 5% and 95% mixtures; (D) PCR-RFLP analysis for wild-type, mutant type, 5% heteroplasmy and anonymous patient sample. LMW refers to low molecular weight… 124

Table 4.1 SBE sites and corresponding SBE primers to generate DNA extension products. 129

Fig. 5.1. Overall scheme of cleavable SPC-SBE approach for SNP genotyping. 138

Fig. 5.2. (A) Schematic of the microfluidic SNP genotyping device. C1 is the central chamber for the single base extension and cleavage reaction, C2 is the

streptavidin bead loading channel, and C3 is the C18 reverse-phase channel. P1 to P5 represent the outlet/inlet ports for sample loading and waste collection. The light blue rectangle indicates the PDMS layer. (B) A close-up view of the temperature sensor; (C) A close-up view of the central heater; (D) A close-up view of a filter; (E) Photograph of microdevice. (F) A light microscopic blow up view of the temperature sensor. 139

Fig. 5.3. Temperature sensor calibration. The formula shown is the linear regression fitting result, with y referring to resistance (Ohms) and x referring to temperature (^{o}C).....142

Fig. 5.4. Results of on-chip single base extension reaction. (A) Scheme of SBE reaction with ddTTP; (B) MALDI-TOF mass spectrum of the SBE extension products.....143

Fig. 5.5. On-chip solid phase capture and cleavage result. (A) Scheme of the whole process including SBE, SPC and cleavage; (B) MALDI-TOF mass spectra of the products.....144

Fig. 5.6. (A) The microchannel packed with C18 coated particles. (B) Mass spectrum after on-chip desalting of 0.5 pmol DNA/TCEP solution.....146

Fig. 5.7. A simplified device process flow referring to cross section a-a' in Fig. 5.2. (a) Metal Cr/Au deposition on glass slide. (b) Lithography for metal patterning and hole drilling in the slide. (c) Thin PDMS layer bonding. (e) PDMS channel fabrication. (f) PDMS channel demolding. (f) Bonding and packaging.....147

Fig. 5.8. (A) Microfluidic experimental test setup with closed-loop feedback temperature control. (B) Bottom view of microdevice; (C) Valving system set-up. 150

Fig. 6.1. Schematic illustration of rationales for cleavable fluorescent nucleotide reversible terminators (CF-NRTs). (A) Stereo diagram of the polymerase active site for incorporating an incoming ddCTP;[4] The ring in red refers to the base (cytosine), and the ring in blue refers to the sugar. The 5 postion of the base and the 3' postion of the sugar are indicated by arrows. (B) General structure of our design of CF-NRTs. 158

Fig. 6.2. Molecular structures of 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP, 3'-*O*-N_3-dCTP and 3'-*O*- N_3-dTTP.....160

Fig. 6.3. Molecular structures of 3'-*O*-N_3-dATP-N_3-ROX, 3'-O-N_3-dUTP-N_3-R6G, 3'-O-N_3-dCTP-N_3-Bodipy-FL-510 and 3'-*O*-N_3-dGTP-N_3-Cy5.....161

Fig. 6.4. Scheme of sequencing by synthesis using CF-NRTs (3'-*O*-N_3-dNTP-N_3-fluorophores) and NRTs (3'-*O*-N_3-dNTPs) 162

Fig. 6.5. Molecular structures of cleavable fluorescent dideoxynucleotide terminators: ddCTP-N_3-Bodipy-FL-510 , ddUTP-N_3-R6G , ddATP-N_3-ROX and ddGTP-N_3-Cy5 164

Fig. 6.6. A hybrid SBS scheme for 4-color sequencing on a chip using the nucleotide reversible terminators (3'-*O*-N_3-dNTPs) and cleavable fluorescent dideoxynucleotide terminators (ddNTP-N_3-fluorophores). 165

Fig. 6.7. Approaches to overcome short read-length. (A) Paired-end sequencing; (B) Library preparation for mate-pair sequencing 167

Fig. 6.8. General scheme of primer walking strategy for extending the read-length. 171

Fig. 6.9. The effects of primer concentration: one base extension with fluorescent reversible terminators (ddNTP-N_3-fluorophores) after hybridization of primers at different concentrations. (A) 37mer primer (B) 52mer primer.....173

Fig. 6.10. Comparison of CL and HD slides: one base extension after primer hybridization. 173

Fig. 6.11. 4-color hybrid SBS data on a linear template (Template1) using ddNTP-N_3-fluorophores and 3'-*O*-N_3-dNTPs. 174

Fig. 6.12. MALDI-TOF MS spectra of single base extension products using self-priming SP26T template and dTTP with DNA polymerase (A) Therminator II, (B) Pfu DNA polymerase, (C) Tgo DNA polymerase and (D) Thermo Sequenase.....178

Fig. 6.13. MALDI-TOF MS spectra of self-priming SP26T template extended with dATP, dGTP and dTTP using (A) Thermo Sequenase (B) Thermo Sequenase and PyroE solution.....180

Fig. 6.14. MALDI-TOF MS spectrum of Primer5163 extended with dTTP, dGTP and dCTP using Thermosequenase and PyroE solution 180

Fig. 6.15. MALDI-TOF MS spectrum of Primer5164 extended with dTTP, dATP, dCTP and 3'-*O*-N_3-dGTP using 9^oN polymerase 183

Fig. 6.16. Four-color DNA sequencing by hybrid SBS after primer "walking". (A) A scheme for primer walking followed by hybrid SBS sequencing. The primer hybridized to the linear templates was extended by polymerase using three dNTPs (dCTP, dATP, dTTP) and 3'-*O*-N_3-dGTP, and extension stopped after the first G. After a capping step to achieve synchronization, the 3'-OH of the extended primer was regenerated by removing the 3'-*O*-N_3 group. Hybrid SBS sequencing started after 3 walking cycles. (B) Four-color sequencing data obtained by hybrid SBS sequencing after primer "walking" 25 bases. 185

Fig. 6.17. Implementation of primer walking strategy for extending the read-length. (A) Scheme of SBS integrated primer walking. Not shown are capping in between each sequencing incorporation cycle. (B) 4-color SBS sequencing data obtained by combining 1^{st} and 2^{nd} round SBS using CF-NRTs. The first 30 bases were sequenced during the 1^{st} round of SBS. 23 bases were sequenced in the second round, which generated the overall read-length of 53 bases.....187

Table 6.1 Sequence information for the DNA templates on slide.....188

Table 6. 2 Volumes of solution A and B in each SBS cycle during hybrid SBS.191

Table 6.3 Volumes of solution A and B in each SBS cycle during hybrid SBS after three walking cycles 195

Fig. 7.1. Scheme of SBS integrated emulsion PCR-beads-on-chip for ultra-massively parallel sequencing.....202

Fig. 7.2. Chemical structures of 3'-*O*-N_3-dCTP-N_3-Bodipy-FL-510, 3'-*O*-N_3-dCTP, 3'-*O*-N_3-dUTP-N_3-R6G, 3'-*O*-N_3-dTTP, 3'-*O*-N_3-dATP-N_3-ROX, 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP-N_3-Cy5 and 3'-*O*-N_3-dGTP 203

Fig. 7.3. Fluorescence imaging data on Chemagen PVA beads: a, one-step EDC coupling; c, two-step EDC coupling with sulfo-NHS; e, PEG coated first and then coupled to DNA; b, d, f are the respective negative controls. 205

Fig. 7.4. Mechanism of EDC coupling 206

Fig. 7. 5. Pyrosequencing data.....207

Fig. 7.6. (A) Scheme of incorporation test with ddNTP-dyes and dNTPs; (B) Sequence data achieved by using ddNTP-dyes/dNTPs 208

Fig. 7.7. (A) Scheme of SBS using 3'-*O*-N_3-dNTPs and 3'-*O*-N_3-dNTP-N_3-fluorophores; (B).....209

Fig. 7.8. (A) Scheme of SBS on Streptavidin beads with an alternative strategy; (B) 4-color SBS data on Streptavidin beads.....210

Fig. 7.9. Sanger sequencing result for PCR products on beads. The associated bar graphs indicate the quality of the sequence: good sequence is indicated in blue. 211

Fig. 7.10. Emulsion formation containing 5 μm beads (A) and 8 μm beads (B).212

Table 7.1 Coupling experiments on different classes of beads.....214

Fig. 8.1. The emission and excitation spectra of QD 605 and Cy5 226

Fig. 8.2. Scheme of QD-based FRET binary oligonucleotide probes for DNA detection. 228

Fig. 8.3. The distribution of carboxyl-QD (1), DNA (2) and carboxyl-QD-DNA conjugates (3) after gel electrophoresis 229

Fig. 8.4. Fluorescence intensity of conjugation products generated with different ratios of carboxylic QD to DNA.....229

Fig. 8.5. The time-dependent spectral evolution of the FRET between carboxylic QD-DNA and Cy5-DNA with hybridization time, presented (A) as fluorescence spectra and (B) in line graph form. 231

Fig. 8.6. Comparison of FRET efficiency between Carboxylic QD based BPs and Streptavidin QD based BPs: the fluorescence spectra before and after hybridization with target DNA 3. (A) Carboxylic QD-Cy5 system; (B) Streptavidin QD-Cy5 system. 232

Fig. 8.7. Steady-state fluorescence spectra of carboxyl-QD and Cy5-DNA hybridization with different targets and without targets. Inset is an enlargement of just the target-containing hybridization reactions with the abscissa contracted for clarity. 235

Fig. 8.8. Time-resolved lifetime fluorescence spectra of carboxyl-QD-DNA and Cy5-DNA with different targets.....235

Table 8.1 Lifetime data for QD-DNA, Cy5-DNA hybridization with different targets 236

Table 8.2 Sequences of the probes and targets.....237

Fig. 9.1. Theoretical SBS read-length based on sequencing cycle efficiency.....247

Acknowledgements

The last five years I spent at Columbia University will be a lifetime asset for me during which I had the fortune to work with the most talented people. They truly helped me grow into a mature person.

First and foremost, I would like to express my sincere thanks to my mentor, Professor Jingyue Ju, for his guidance, supervision, and support throughout the entire course of my graduate study. It has been and will always be my privilege to learn from Prof. Ju, a motivated scholar and a passionate scientist.

I would also like to convey my gratitude to Dr. James Russo, who has always been patient and kind to me and has taught me a lot from biological knowledge to experimental design and to scientific writing. Advice and support from Dr. Zengmin Li and Dr. Shiv Kumar in the area of chemistry are greatly appreciated. The same appreciation is extended to Drs Xiaoxu Li, Shundi Shi, Sergey Kalachikov, Irina Morozova and Minchen Chien for their support across the chemistry and biology. My thanks go out to my wonderful colleagues as well: Dr. Yu Lin, Dr. Jia Guo and Mirkó Palla for their pleasant collaborations and friendship; Wenjing Guo, Dr. Jian Wu and Ning Xu, for their friendship and help.

I send my acknowledgement to Prof. Eric A. Schon, Dr. Ali B. Naini and Dr. Jiesheng Lu for providing me with mitochondrial samples and their insightful suggestions; to Prof. Qiao Lin and his group members, Dr. ThaiHuu Nguyen and Jing

Zhu for teaching me lab-on-chip technologies and helping me design and construct the microfluidic device, to Prof. Yiru Peng, for working with and guiding me through the imaging project, and to Prof. Nicholas Turro and Dr. Steffen Jockusch for giving me insightful suggestions and help.

I would also like to thank my defense committee members, Professors Qiao Lin, Ben O'Shaughnessy, Vanessa Ortiz, and Eric A. Schon for their time and invaluable advices.

In addition, I was so fortunate to have my special friends, who have accompanied me through the years at Columbia: Chao Sun, Shelby Yue Zhang, Zheyuan Chen, Tianxia Jia, Xueyu Pang and Jie Xu, to just name a few. Last but not least, I would like to thank three very special people who have always showed me love and care regardless of my ups and downs: my father Zhihe Qiu, my mother Qiuxiang Yan, and my fiancée Sai Zhang. I cannot imagine myself accomplishing anything without their selfless devotion and support. My special acknowledgement goes to my beloved grandmother, Minglan Gong, whose spirit has always been a source of support through my life.

Finally, I appreciate all the people who cared about me and helped me in any way. Thank you very much!

Dedicated to my beloved grandmother, my parents and Sai

Abbreviations and Symbols

3-HPA	3-hydroxypicolinic acid
9^oN polymerase	9^oN A485L/Y409 (exo-) DNA polymerase (Therminator^{TM} II DNA Polymerase)
A	adenine
ACN	acetonitrile
ATP	adenosine 5'-triphosphate
Bodipy	4,4-difluoro-5,7-dimethyl-4-bora-3α,4α,-diaza-s-indacene
BP	binary probe
bp	base pair
B/W buffer	binding and washing buffer
C18	octadecyl carbon chain
C	cytosine
CF-NRT	cleavable fluorescent nucleotide reversible terminator
Cy5	cyanine-5
Da	Dalton
dATP	2'-deoxyadenosine triphosphate
dCTP	2'-deoxycytidine triphosphate

ddNTP	dideoxynucleotide triphosphate
dGTP	2’-deoxyguanosine triphosphate
DMF	*N,N*-Dimehtylformamide
DMSO	dimethyl sulfoxide
DNA	deoxyribose nucleic acid
dNTP	deoxyribonucleoside 5’-triphosphate
dsDNA	double-stranded DNA
dTTP	2’-deoxythymidine triphosphate
dUTP	2’-deoxyribouridine-5’-triphosphate
EB	ethidium bromide
EDC	1-ethyl-3-(3-dimethylaminopropyl)carbodiimide
EDTA	ethylenediaminetetraacetic acid
FRET	fluorescence resonance energy transfer
FAM	5-caboxyfluorescein
G	guanine
Hae	*Haemophilus aegyptius bacteria*
HPLC	high performance liquid chromatography
LOC	lab-on-a-chip
MALDI-TOF MS	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MB	molecular beacon

MES	2-(N-morpholino)ethane sulfonic acid
MERRF	myoclonic epilepsy with ragged red fibers
MT	mutant-type
mtDNA	mitochondrial DNA
MSBE	multiple single base reaction
MW	molecular weight
nDNA	nuclear DNA
NGS	next generation sequencing
NHS	*N*-hydroxy succinimidyl
NRT	nucleotide reversible terminator
nt	nucleotide
OPC	oligonucleotide purification cartridge
Parylene	poly(p-xylyene) polymers
PCR	polymerase chain reaction
PDMS	polydimethylsiloxane
PEG	polyethylene glycol
PID	proportional-integral-derivative
PPi	pyrophosphate
QD	quantum dot
RFLP	restriction fragment length polymorphism
RHOD	ras homolog gene family D

RNA	ribonucleic acid
ROX	6-carboxy-X-rhodamine
R6G	6-carboxyrhodamine 6G, hydrochloride
S/B	signal-to-background ratio
SBE	single base extension
SBS	sequencing by synthesis
SDS	sodium dodecyl sulfate
SMS	single molecule sequencing
SMRT	single-molecule real-time
SNP	single nucleotide polymorphism
SP	self-priming
SPC	solid phase capture
SPSC	sodium phosphate sodium chloride
ssDNA	single-stranded DNA
T	thymine
Taq	*Thermus aquaticus*
TEAA	triethylammonium acetate
TCEP	tris(2-carboxyethyl)phosphine
TCR	temperature coefficient of resistance
UV	ultraviolet
WT	wild-type

λ_{abs}	maximum absorption wavelength (nm)
λ_{em}	maximum emission wavelength (nm)

Chapter 1 Introduction to Genomic Analysis Technologies -- DNA sequencing, Genotyping, Nucleic Acid Detection

For thousands of years, people have been seeking the answers to the mystery of life, what defines us inherently, what makes us different from others, and what causes malfunctions of our body, etc. Thousands of years of accumulated efforts by generations of pioneers and followers have revealed many of wonders of living organisms. From the discovery of DNA as the hereditary material to the central dogma, from the recognition of nucleotides as DNA building blocks to DNA's double helical structure, from the deciphering of the sequential information to the tracking of the molecular events inside the cells, advances in genomic studies have continued to revise and deepen our understanding of life through abundant new discoveries, and are leading us to an epoch of synthetic genomics and personalized medicine with the promise of extending life expectancy. Nevertheless, the advances in genomic studies are more and more relying on sophisticated tools, wherein DNA sequencing, genotyping, and detection of nucleic acid *in vivo* have played some of the most revolutionary roles. DNA sequencing, as the means to determine the sequential order of nucleotides of a genetic information carrying DNA strand, together with genotyping, the fine interrogation of genetic variations in the DNA leading to phenotypic differences and disease, have shed light on novel biomarker and drug discoveries as well as gene therapy. The probing of nucleic acids inside cells to track the transmission and transcription of genetic information, is leading to deeper

understanding of biological phenomena and functions. Progress in any of these genomic technologies will eventually bring us incredible advances in our knowledge of cells and tissues in health and disease.

1.1 Introduction to DNA sequencing technology

1.1.1 Background and significance

Deoxyribonucleic acid, or DNA, is a double helical polymer composed of deoxyribonucleotide monomers, each containing a base, a sugar and phosphate group[1]. These building blocks of the DNA molecule, normally called deoxyribonucleoside 5'-trisphosphates (dNTPs) are shown in Fig. 1.1, where N refers to the base. Specifically, the 3'-hydroxyl group of the sugar of one nucleotide is joined to the 5'-phosphate group of the sugar in the adjacent nucleotide by a phosphodiester bond, generating the polymer strand (DNA) with a phosphate group on the 5'-terminal sugar unit (5'-end) and a free hydroxyl group at the 3'-terminal sugar unit (3'end). Sugar and phosphate groups of DNA molecules perform structural supporting roles, whereas it is the bases of DNA molecules that carry genetic information. There are four different bases, adenine (A), guanine (G), thymine (T) and cytosine (C), with thymine replaced by uridine in ribonucleic acid (RNA). Adenine and guanine are purine derivatives, while thymine and cytosine are pyrimidine derivatives. In the DNA double helix structure, adenine is always paired with thymine, and guanine with cytosine, which is often referred to "Watson and Crick base-paring" or "complementary base-paring". The order of these bases (A, G, C

and T) determines specific genetic information.

The DNA molecule undergoes two biological reactions: replication for transmission of the genetic information to daughter cells during cell division, and transcription for guiding the synthesis of RNA that is then translated into amino acid sequence to complete the flow of genetic information. Both of these processes play tremendous roles in cell biology. The deciphering of the complete sequence of the genomic DNA is of great significance for our inherent understanding of living species.

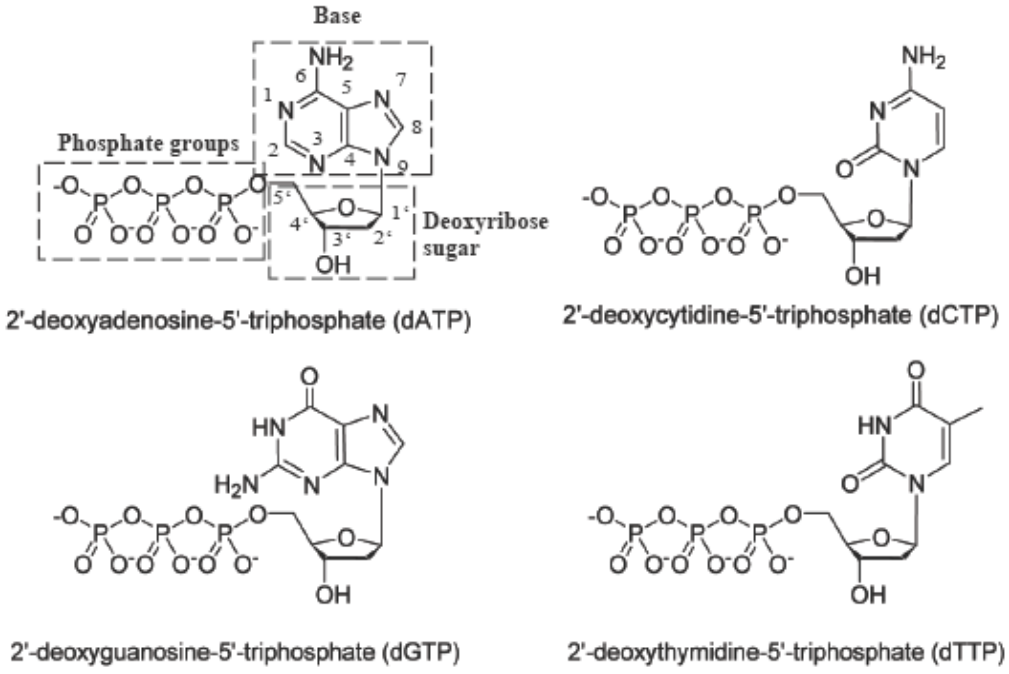


Fig. 1.1 Chemical structures of 2'-deoxyribonucleotides (dNTPs). Each nucleotide is composed of a base (adenine, guanine, cytosine or thymine), a sugar and a phosphate group.

Inspired by the DNA replication process, a template-directed event in which DNA polymerase accurately catalyzes the addition of complementary nucleotides to the 3' end of a growing DNA strand via the formation of a phosphodiester bond between the

terminal 3'-OH group and the 5'-phosphate group of the incoming nucleotide, Sanger first developed chain-termination sequencing methods.[2, 3] Since the first genome, bacteriophage phi x 174 DNA, was completely sequenced by Sanger,[4] DNA sequencing technologies have made a significant impact in the biomedical sciences, with its applications across such fields as comparative genomics and evolution, forensics, epidemiology and applied medicine for diagnostics and therapeutics.

With the completion of the Human Genome Project, the National Human Genome Research Institute (NHGRI) has called for new sequencing technologies to reduce the cost of the current Sanger based method by 100-fold in the near term ($100K Genome) and by a further 100-fold in the near future ($1,000 Genome).[5] This invigorated the whole sequencing industry, resulting in a variety of new technologies and an astonishing number of scientific advances in the field of genomics and beyond.

1.1.2 DNA sequencing technologies overview

As a well established DNA sequencing method, Sanger sequencing has its strength in long read-length for *de novo* sequencing but limitations in its throughput as well as accuracy. The growing demand for sequencing has driven the development of alternative sequencing technologies towards rapid and inexpensive DNA sequencing, aiming at the vision of the $1000 genome. Various current and developing sequencing technologies as well as current underdeveloped sequencing technologies will be reviewed here.

1.1.2.1 Conventional Sanger Sequencing

Since it was first presented in 1977, Sanger sequencing has been one of the most influential innovations in biological research.[3] The principle of Sanger sequencing, also known as the chain-termination sequencing method, is to stochastically terminate the primer extension with labeled, distinguishable dideoxynucleotide (ddNTP) terminators to generate different fragment lengths that can aid in interrogating the terminating base. As shown in Fig. 1.2, in current Sanger sequencing reactions, uniquely fluorescently labeled ddNTPs (ddNTP-fluorophores) are mixed with non-labeled, natural deoxynucleotides (dNTPs) to generate elongation fragments ending at all positions along the template. The sequencing biochemistry is actually a cycle sequencing reaction with template denaturation, primer annealing and primer extension, resulting in a linearly amplified mixture of end-labeled extension products, which can be separated by gel electrophoresis. The sequence is then obtained by reading the fluorescent signal from the terminating nucleotides in length order.

For decades, with the introduction of new improvements, such as capillary electrophoresis,[6] laser induced fluorescent excitation of energy transfer dyes,[7] a fully automated system[8] and others, Sanger sequencing has evolved into a well-established sequencing method with an average sequencing read-length of 800 bases[9] and has become the choice for numerous sequencing projects, including the first human genome project.

However, electrophoresis based Sanger sequencing has inherent limitations due to

the resolution of the product separation matrix, and it is also too expensive and time-consuming for current and envisioned projects. Though people have been trying to further improve the Sanger sequencing method by pursuing multiplexing and miniaturization, there appears to be a few remaining approaches for achieving significant reductions in the cost and massively parallel performance. The limitations of the Sanger sequencing technique and the increasing demand of genome sequencing projects have propelled the new generations of sequencing technologies.

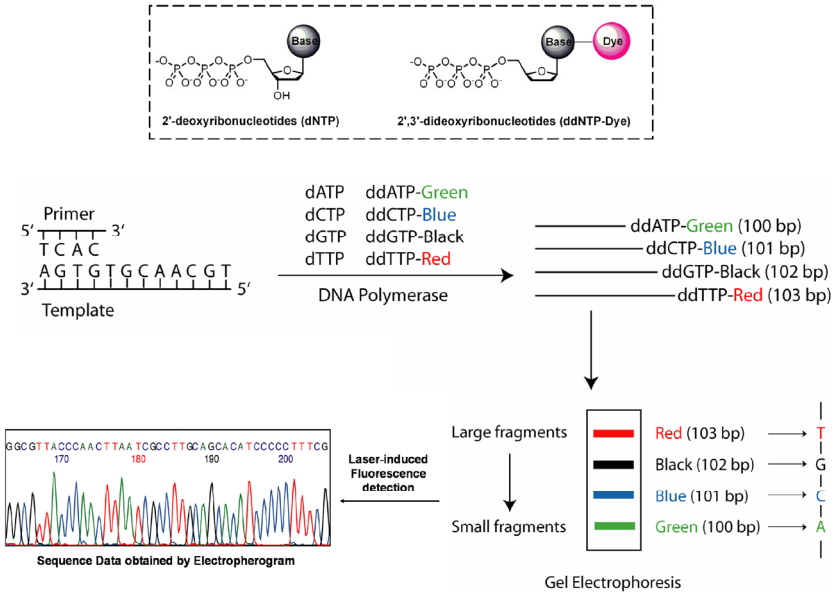


Fig. 1.2. Principle of Sanger sequencing.

1.1.2.2 Mass Spectrometry based sequencing

Since the first application of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) to oligonucleotide analysis in 1990[10], MALDI-TOF MS has become one of the most powerful tools for DNA analysis

due to its high speed, high accuracy and high resolution. It has been widely explored as an alternative method to electrophoresis based Sanger sequencing. The most straightforward approach is to replace the polyacrylamide gel separation of the Sanger sequencing products with mass analysis,[11] as shown in Fig. 1.3 A. The sequence is determined by the mass differences between the fragments produced by Sanger sequencing reactions. This approach was further improved by Ju *et al.* by introducing solid phase capturable and mass-tagged dideoxynucleotides,[12, 13] which will be described in detail in Chapter 3.

The other commonly used strategy is mass ladder sequencing by enzyme digestion, which was first described by Pieles *et al.*[14] Generally, as shown in Fig. 1.3 B, the oligonucleotides are cleaved using an enzyme that sequentially removes nucleotides from either the 5' or the 3' end, like the 5'-exonuclease calf spleen phosphodiesterase (CSP) or the 3'-exonuclease snake venom phosphodiesterase (SVP). The sequence is determined by the mass change after each cleavage. Bentzley *et al.* further extended this method to achieve a read-length of 24 bases.[15] It is also possible to identify bases with methyl modifications and sugar-modified nucleosides with different digestion strageties.[16] This exonuclease digestion method has proven to be helpful for the sequence determination of moderate length oligonucleotides.

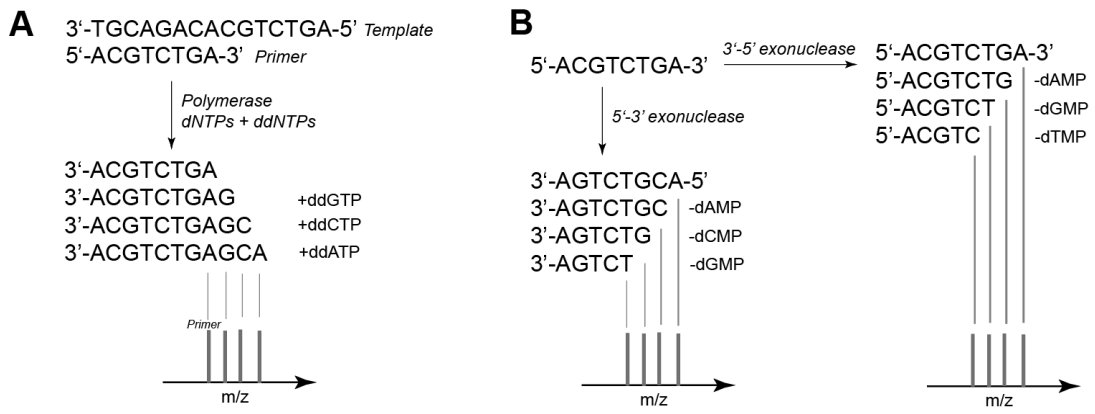


Fig. 1.3. Schematic process of mass spectrometry assisted sequencing. (A) Sanger sequencing reaction based MS sequencing; (B) Enzymatic digestion based MS sequencing.

As MALDI-TOF MS is well-suited to analyzing RNA, RNase digestion-based protocols have also been developed to investigate RNA sequences.[17] As an alternative strategy to sequence ladder generation, Hahner *et al.* developed an approach to analyze the sequence by converting DNA to RNA and generating fragments using sequence-specific RNase.[18] Gut *et al.* further modified it to a DNA/RNA chimera approach by replacing one of the four dNTPs with nucleoside triphosphate (NTP), allowing fragmentation by alkali backbone cleavage for the analysis.[19] Both of these methods require a comparison between the unknown sample and a reference sequence, limiting their application in sequencing.

Several other strategies for DNA sequencing with mass spectrometry were also reported, including gas-phase fragmentation that made use of fragmentation during mass spectrometric analysis to determine the sequence across the numerous fragments, and sequence determination based on mass ladders of the failure products generated in solid-phase oligonucleotide synthesis. However, technically, it is hard to control the

gas-phase fragmentation, and the sequencing of failure products is not applicable for natural oligonucleotides.[20]

Though different mass spectrometry based sequencing approaches have been developed, mass analysis of Sanger sequencing products remains the most straightforward and simplest, and has the potential for unambiguous high speed sequencing. However, the main challenges of mass spectrometry lie in limitation of the range of masses one can obtain on a given instrument, the stringent requirement for purity, and difficulties in achieving massively parallel detection.

1.1.2.3 Sequencing by hybridization

With the availability of oligonucleotide synthesis and more and more understanding of DNA strand hybridization and denaturation,[21] sequencing by hybridization (SBH) was first proposed by Drmanac and Crkvenjakov in 1987 as an alternative to conventional Sanger sequencing.[22, 23] Originally, there were two formats for SBH. In one approach, the unknown target DNA was immobilized to a solid substrate and hybridized with a set of labeled oligonucleotide probes. The sequence was determined by assembling information from the overlapped probes. Drmanac has reported the sequencing of *p53* exons 5-8 using this approach with 10,000 7-mer probes.[24] The other approach is to determine the sequence through the hybridization of labeled unknown DNA targets to substrate bound probes, which has been widely utilized by most current SBH technology, like Affymetrix (Santa Clara, CA, USA), Perlegen (Moutain View, CA, USA) and Roche NimbleGen

(Madison, WI, USA).[25] Here, a DNA microarray/microchip with thousands of oligonucleotide probes is constructed, each probe representing a piece of reference sequence in the genome of interest. Though the SBH based microarray technology is useful in DNA resequencing and large scale genome analysis,[26] it is more and more used in nucleotide variation detection rather than sequencing, which will be discussed in detail in the section on SNP genotyping. Besides, there are some limitations to this approach. Accuracy is always an issue due to the complexity of DNA hybridization which depends on many factors, including GC content, length, concentration and secondary structure of the target and probe sequences. False positives often occur between exact matches and those with a single base difference, and a region with repetitive sequence stretches also poses challenges in unambiguous determination. The use of longer probes offers the possibility of deciphering longer sequence stretches, but at prohibitive cost due to the exponentially increasing number of probe combinations. With these limitations, the prospects for *de novo* genome sequencing remain unclear.

1.1.2.4 Sequencing by synthesis

The heart of sequencing by synthesis (SBS) is to iteratively identify the incorporation of each nucleotide during the extension reaction by DNA polymerase. It has emerged as the core technology of second generation sequencing, and is also being widely explored for single molecule based third generation sequencing. Sequencing by synthesis was first revealed by Heyman in his description of pyrosequencing,[27] and

quickly developed into a variety of approaches, such as pyrosequencing,[28] sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators,[29] single molecule sequencing with virtual nucleotide terminators[30] and SMART sequencing,[31] each of which will be reviewed here. A related method, sequencing by ligation,[32] will also be described.

1.1.2.4.1 Pyrosequencing

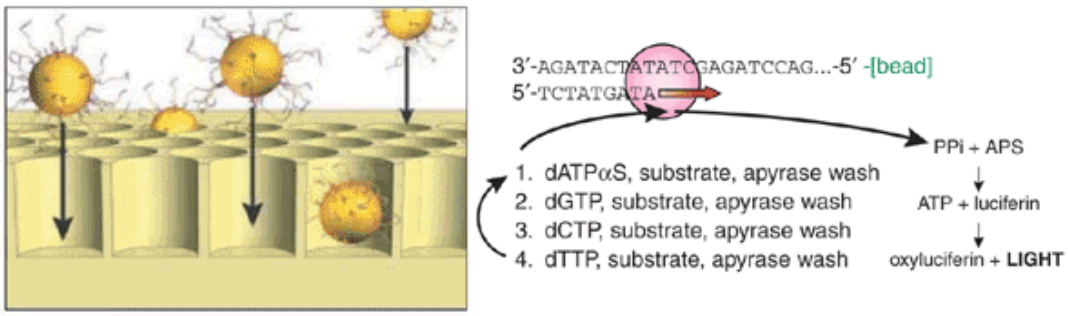


Fig. 1.4. Schematic view of pyrosequencing.[25]

The principle behind pyrosequencing is the detection of the pyrophosphate that is released during enzymatic DNA synthesis for nucleotide determination. It involves a cascade of enzymatic reactions.[27, 33] Briefly, it starts with a nucleotide polymerization reaction in which pyrophosphate (*PPi*) is released, one pyrophosphate per nucleotide. The *PPi* is subsequently converted to ATP by ATP sulfurylase, which drives the luciferin oxidization reaction by luciferase and emits a photon of light (luminescence). By continuous stepwise addition of four deoxynucleotide triphosphates and observing the light signal, the nucleotide that is complementary to the DNA template can be identified.

This pyrosequencing method was first employed by Hayman and further developed by Pål Nyrén and Mostafa Ronaghi. A series of strategies have been utilized to optimize pyrosequencing,[28] including the replacement of natural deoxyadenosine triphosphate (dATP) by deoxyadenosine α-thiotriphosphate (dATPαS) in the polymerase reaction to avoid a false positive signal from the added ATP analog,[34] the introduction of apyrase to degrade the excess nucleotides,[35] and the addition of single stranded DNA (ssDNA)-binding protein for longer template sequencing.[36] Pyrosequencing evolved rapidly and became the first commercialized next generation sequencing technology, launched by 454 Life Sciences.[37] In the Roche/454 pyrosequencing system, unlike conventional Sanger sequencing, the sample preparation avoids cloning and is instead achieved by emulsion PCR amplification of single DNA molecule on beads. As shown in Fig. 1.4, the larger amplicon carrying beads surrounded by smaller beads with luciferase and ATP sulfurylase are deposited onto a microfabricated array of picoliter-scale wells (picoTiterPlate). Other reagents are flowed in at each cycle. Only one of the four dNTPs is added in each cycle, generating light captured by a CCD camera if incorporation occurs. Apryase is flowed in to degrade excess nucleotides after each round. The current Roche/454 Titanium Platform has been used in whole-genome sequencing of different species with significantly improved throughput. However, there are some limitations. The major challenge is sequencing homopolymer regions, since the approach lacks accuracy in determining the number of nucleotides if continuous incorporation occurs. As the cycles accumulate, the efficiency of the immobilized enzyme drops. Synchronization

is a problem since the data is generated from a cluster of DNA molecules.

1.1.2.4.2 Sequencing by ligation

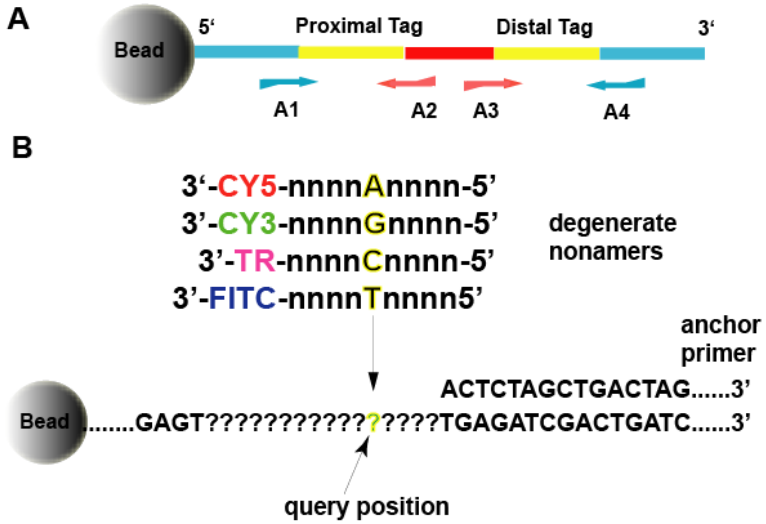


Fig. 1.5. Principle of sequencing by ligation. (A) DNA template molecule with two mate-pair tags of unique sequence which are flanked and separated by universal sequences complementary to amplification or sequencing primers. (B) Steps of sequencing by ligation.

Strategies for sequencing by ligation (SBL) were first developed in polony sequencing of the *Escherichia coli* genome by Church *et al.*[32] Generally, the sample preparation before SBL involves library construction of DNA molecules with two mate-pair (sequencing reads derived from opposite ends of each fragment) tags of unique sequence which are flanked and separated by universal sequences complementary to amplification or sequencing primers (Fig. 1.5 a), emulsion PCR for generating multiple copies of a single DNA template (colony) on 1 μm magnetic beads, and the immobilization of colony carrying beads on a solid support for performing the

sequencing cycles. As shown in Fig. 1.5 B, sequencing by ligation is divided into four steps. An anchor primer is first hybridized to a priming site (immediately 5' or 3' to one of the two tags, A1 to A 4, in Fig. 1.5 A, indicating the priming site for each anchor primer) within the single-stranded template. A ligation reaction is then performed with a pool of fully degenerate, fluorescently labeled nonanucleotides with one encoding base that requires perfect matching for hybridization and ligation. Therefore, the nucleotide at the query position can be identified according to the fluorophore signal from the nonamer that is exactly complementary to this position. The cycle is repeated after stripping away the anchor-primer/fluorescent-nonamer complex. Church demonstrated that a total of 26 bp was obtained through SBL, with 6 bases from the ligation junction in the 5' to 3' position and 7 bases from the ligation junction in the 3' to 5' direction for both mate-paired tags. The development of sequencing by ligation contributed to the commercial release of the Polonator and the ABI SOLiD, the latter of which introduced a slight modification to the original SBL method, the use of labeled degenerate octamers with two base encoding, 5 base extensions in each cycle with primer resetting (shift by one nucleotide) after multiple cycles.[38, 39] However, limitations exist, including difficulties in localization of beads due to random dispersion, a higher error rate of raw sequence data than in the Sanger method, incomplete dye removal, false read out due to proximity of beads as well as a mixture of templates on beads, and errors and difficulties during *in vitro* amplification steps.[40, 41]

1.1.2.4.3 Sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators

An alternative sequencing by synthesis approach, SBS with cleavable fluorescent nucleotide reversible terminators (CF-NRTs), also known as array cyclic sequencing, was shown to have advantages over pyrosequencing in sequencing homopolymer regions and allowing addition of all four nucleotides at once, and possess advantages over the sequencing by ligation approach in terms of sequencing read-length and lower fluorescent background. It soon became one of the most widely explored sequencing technologies and has been commercially available for several years.

Generally, SBS with CF-NRT is composed of three steps: incorporation, detection (imaging), and deprotection (cleavage). The key is the reversible nucleotide terminators, which ideally should exhibit efficient incorporation by DNA polymerases, efficient deprotection under mild conditions and labels with desired characteristics.[42] Nevertheless, it remained unsuccessful until Ju's publication on the design of CF-NRTs by tethering a unique fluorescent dye to the 5-position of the pyrimidines (T and C) and the 7-position of the purines (G and A) via a cleavable linker, and attaching a small cleavable chemical moiety to cap the 3'-OH group.[43] The rationale is that, though polymerase is sensitive to modification of the 3' position due to its proximity to the amino acid residues in the active site of the polymerase, some modified DNA polymerases are highly tolerant to nucleotides with extensive modifications with bulky groups at the 5-position of the pyrimidines (T and C) and the 7-position of the purines (G and A). Different sets of

cleavable nucleotide reversible terminators have been developed, some of which have been employed in current next-generation sequencing.[29, 44, 45] The whole process of SBS with CF-NRTs is illustrated in Fig. 1.6. Two sets of nucleotide reversible terminators are used, one for the incorporation step and one for capping. The first set consists of the four nucleotides, each linked to a different fluorescent dye, with their 3' position blocked with chemical group R (CF-NRTs), while the second set consists of non-fluorescently labeled nucleotide reversible terminators (NRTs), but with their 3' position blocked with the same R group. After the incorporation step, a template-dependent primer extension occurs with a CF-NRT, so the nucleotide can be identified according to the fluorescent signal. For synchronization, a "capping" step with NRTs is performed to extend un-reacted templates. Then the fluorescent dye and the blocking group are cleaved away, so the 3' hydroxyl group is regenerated, in readiness for the next sequencing cycle. As the cycles progress, the sequence of the template can be identified base by base continuously. Hundreds of millions of amplified templates immobilized on slides can be decoded simultaneously using CCD cameras or other imaging approaches.

Similar to other SBS approaches, the current limitation of this approach is the relatively short read-length, due to the DNA template cluster synchronization, the surface chemistry, and the efficiency of each step. These somewhat limit its application for *de novo* sequencing.

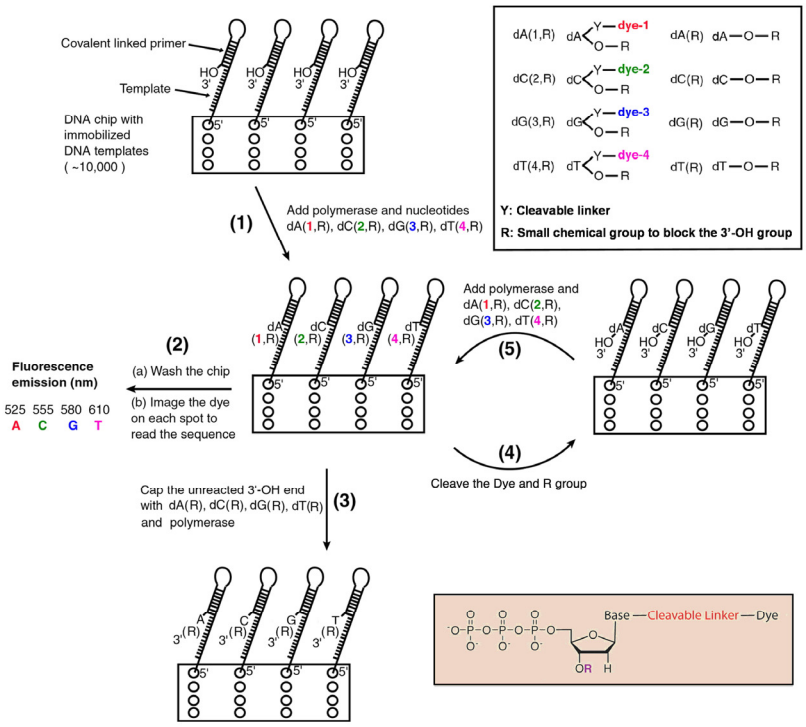


Fig. 1.6. Principle of sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators.

1.1.2.4.4 Single molecule sequencing-by-synthesis

The previously described sequencing-by-synthesis methods rely on PCR to grow clusters of a given DNA template on a solid surface (beads or directly on flow cells) that are then imaged in a phased approach. This not only requires tedious labor and a large amount of genomic DNA, but also brings in biases via PCR amplification and synchronization problems during sequencing.[41] Therefore, sequencing-by-synthesis without a prior amplification step, also known as single molecule sequencing, is currently being pursued. The introduction of single molecule sequencing (SMS), free of PCR amplification, will simplify sample preparation, avoid biases or errors from the library preparation or amplification step, allow asynchronous extension of single

molecules and flexibility in the kinetics of sequencing chemistry, enable extremely high densities of templates, make it possible to identify DNA modifications that are lost in the *in vitro* amplification process, and enable direct RNA sequencing.

Since the first demonstration of the feasibility of single-molecule sequencing in 2003,[46] it took five years to launch the first sequencer that was able to sequence individual molecules instead of molecular ensembles created by an amplification process, the HeliScope single-molecule sequencer by Helicos Biosciences.[30] The input genomic DNA is first fragmented, denatured, and a poly-A-tail added by polyadenylate polymerase, with the last adenine labeled with Cy3. The labeled poly-A-tail helps in localization of the DNA strand after it hybridizes to the poly-T oligonucleotides immobilized on the flow cell and also serve as primers during the incorporation process. The fluorescent label of adenine is released before the sequencing reactions start. The core of the Helicos single molecule sequencing method is their so called asynchronous virtual terminator chemistry. A virtual terminator is a type of nucleotide that is attached to a fluorescent dye via a disulfide bond linker (Cy5-12ss-dNTP analogs), while its 3' position is unblocked.[47] Virtual terminators can be incorporated by polymerase at a slower rate than natural nucleotides and are able to temporarily block incorporation of a second nucleotide on a homopolymer template if polymerase and excess nucleotides can be washed away instantly (termed virtual termination). Therefore, in each cycle, as shown Fig. 1.7, Cy5-labeled nucleotides are incorporated, asynchronous growth of individual DNA molecules is monitored by fluorescence imaging, and the fluorescent

dye is cleaved off for the next cycle of sequencing. Using this single-molecule method, Quake *et al.* have reported the sequencing of an individual human genome, with an average read length of 32 bp.[48]

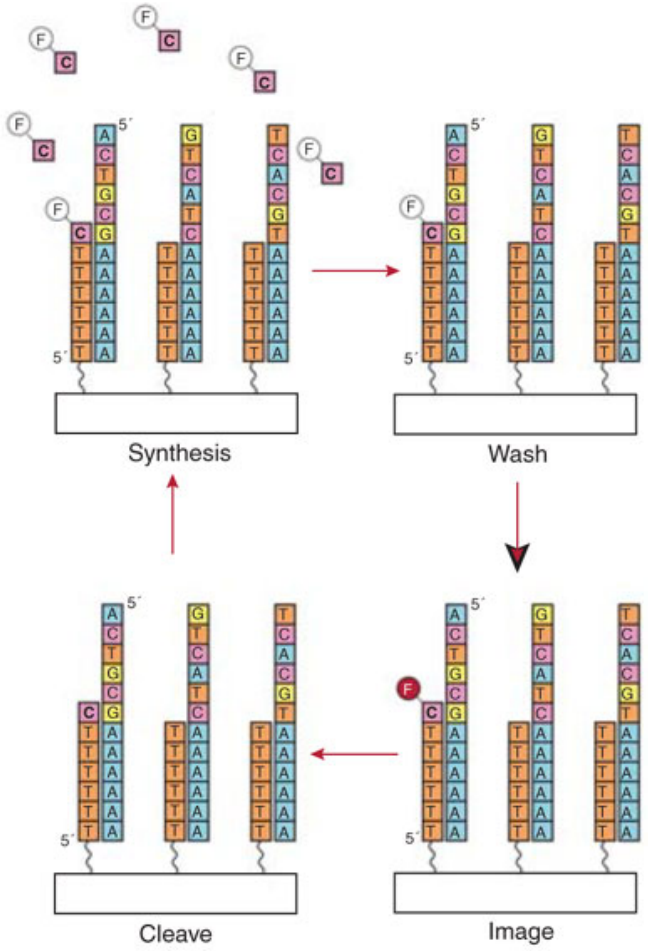


Fig. 1.7 Schematic of the single molecule sequencing-by-synthesis approach with virtual terminators.[49]

The introduction of Zero Mode Wavelength (ZMW) technology[50] has spawned a different approach, Single-Molecule Real Time Sequencing-by-synthesis (SMRT), developed by Pacific Biosciences.[31, 51, 52] As shown in Fig. 1.8, ZWM is a nanophotonic structure, with nanoscale wells of 70 nm in diameter and the light projects inward at the

opening, allowing an illumination detection volume of 20 zl (10^{-21} l) where a single DNA polymerase is immobilized. This enables detection of an individual fluorophore on an incorporating nucleotide against a dense background of labeled nucleotides, hence guaranteeing fast and processive enzyme synthesis as well as a significantly reduced signal-to-noise ratio.[53] Distinguishable fluorophores are tethered to the phosphate chain of the nucleotide through a phospholinker so that the natural strand of DNA will be regenerated after cleaving away the linker-dye, enabling continuous incorporation without steric hindrance. As shown in Fig. 1.8 B**,** the DNA sequence is determined by detecting the fluorescent pulse from the correctly base paired (cognate) nucleotide binding in the active site of the polymerase. The lifetime of the fluorescence is controlled by the rate of catalysis, and ends after the cleavage, the period of which is much longer than diffusion and non-cognate sampling, allowing correct identification of incorporation reactions. The period between each pulse reflects the DNA molecule's translocation and subsequent nucleotide binding. Based on the direct observation of a processive DNA incorporation reaction, SMRT has the potential for continuous observation of DNA synthesis over thousands of bases with a maximum read length in excess of 10,000 bp. Chin *et al.* have successfully employed this technology to sequence the whole genome of the pathogen responsible for the recent cholera epidemic in Haiti, the Haitian Vibrio cholera outbreak strain.[54] In addition, since modifications of the nucleotides could be identified by different polymerase kinetics, appearing as different arrival times and durations of the resulting fluorescence pulses through the incorporation, Flusberg *et al.*

were able to directly identify nucleotide methylation and hydroxymethylation, including mA, mC and hmC sites, without bisulfite conversion.[55] However, due to the very short intervals, dissociation of the complementary nucleotide before incorporation, and spectral misalignment of fluorescent dyes, the accuracy of SMRT is still very low. Furthermore, the throughput is limited by the number of ZMWs that can be read simultaneously. [9, 56]

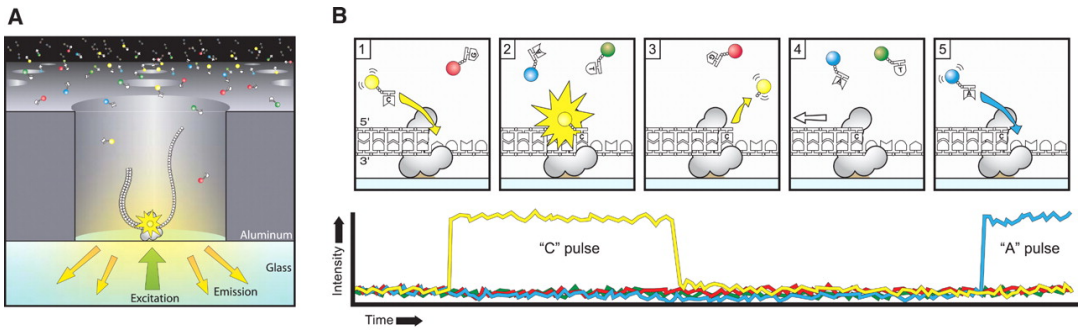


Fig. 1.8 Principle of single-molecule, real-time DNA (SMRT) sequencing.[31]

Other SBS based single molecule sequencing strategies are also being pursued. For example, Visigen Biotechnologies is trying to modify SMRT technology by using fluorescence resonance energy transfer (FRET) donor attached polymerase and different FRET acceptor modified nucleotides to identify the base by real time observance of FRET.[57] As an alternative detection method, Ion Torrent's semiconductor sequencer now sold by Life Technologies is trying to exploit semiconductor technology to create a high density array of micro-machined wells for the detection of released hydrogen ions during base incorporation, in an approach which is free of light, scanning and cameras.[9]

1.1.2.5 Sequencing by direct physical recognition of the DNA molecule

Besides sequencing by synthesis based single molecule sequencing, efforts have been undertaken to develop a sequencing strategy involving direct physical recognition of the DNA molecule, avoiding the attachment of a fluorophore to the nucleotides and the associated limitations of the polymerase enzyme. This, together with SBS-based SMS, has been explored as so-called next-next (third) generation sequencing.[56, 58]

1.1.2.5.1 Nanopore sequencing

Nanopore sequencing was proposed in 1988 and its feasibility was first demonstrated by Deamer *et al.* in 1996,[59, 60] which has stimulated nanopore research in the ensuing years.[61] As shown in Fig. 1.9 A, the principle underlying nanopore sequencing is that a single stranded DNA (ssDNA) molecule can be electrically driven through a nanoscale pore (1.5-5 nm) in strict linear form under an applied voltage, resulting in a change of electrical signals, such as ionic current blockages, transverse tunneling currents, or capacitance, for nucleotide discrimination. The two most common classes of nanopores are protein pores in lipid bilayers, such as α-hemolysin (αHL) pores and *Mycobacterium smegmatis* porin A (MspA), and synthetic solid nanopores formed in a variety of thin films, including Si-based materials, graphene and carbon nanotubes. The development of nanopore sequencing is a challenging process due to the control of DNA velocity and orientation during translocation, construction of nanopores with comparable channel lengths to a single nucleotide (~ 0.4 nm), discrimination of individual

nucleotides with subtle chemical and electrical differences, and the high background level. Nevertheless, various methods have been tested in order to employ nanopore sequencing.

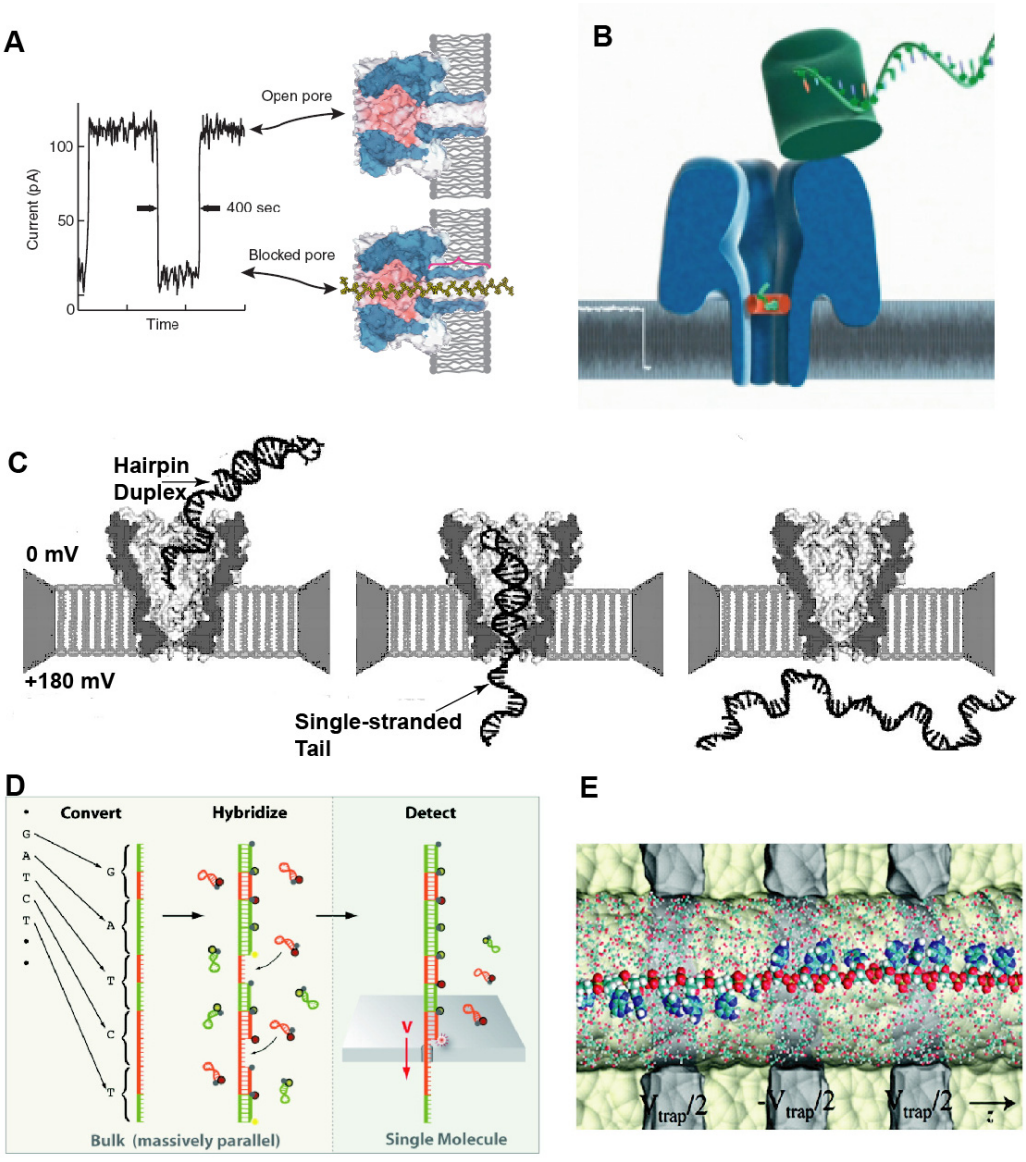


Fig. 1.9. Schematic of various Nanopore sequencing method. (A) Fundamental principle of nanopore sequencing; (B) Nanopore sequencing approach by Oxford Technology[65]; (C) duplexes halting translocation sequencing by MspA nanopore;[66] (D) Nanopore single-molecule optical detection approach;[69] (E) Schematic of the DNA transistor. The light blue regions with voltage labeled are the conductors, while the light green regions are insulators.[67]

Under the high applied potentials required for threading DNA molecules, freely moving DNA molecules pass through the wild-type αHL pore too quickly for bases to be identified. To solve this problem, Bayley and his colleagues used an engineered nanopore with a molecular adaptor to capture and identify the nucleotides, thereby slowing the movement of DNA.[62] They reported the identification of unlabeled 5'-monophosphate molecules with an average accuracy of 99.8% using an αHL nanopore with a covalently attached cyclodextrin molecular adapter[63] and also demonstrated three recognition sites in the transmembrane β-barrel of an engineered αHL pore for identification of all 4 bases in an immobilized single-stranded DNA molecule.[64] They further utilized the enzyme to reduce the speed of DNA translocation, and initiated a nanopore-exonuclease approach, being co-developed by Oxford Nanopore Technologies. As shown in Fig. 1.9 B, an exonuclease was attached onto the protein pore. As the DNA strand is digested by exonuclease, released deoxynucleoside monophosphates are captured and identified within the pore. The sequencing of independent readings reflects the order in which the bases are cleaved from the DNA.[64, 65] In addition, this group is also trying to develop so called "strand sequencing", via a combination of nanopores with processive enzymes, wherein individual nucleotides passing through the nanopore are identified and then incorporated by an enzyme covalently attached to the nanopore. Again, the enzyme plays a role as the speed controller for DNA strand translocation.[65]

As an alternative to the αHL nanopore, the MspA nanopore has emerged as an even more promising biological nanopore with its shorter blockade region, addressing axial

resolution limitations of the αHL nanopore and yielding better-resolved current signatures.[9, 66] By using engineered MspA nanopores, Derrington *et al.* developed a duplex halting translocation sequencing approach. In brief, as shown in Fig. 1.9 C, a region of double-stranded DNA (dsDNA) is introduced into a target single-stranded DNA (ssDNA). When the DNA strand is passing the nanopore, the double-stranded section will cause the translocation to halt while holding the single-stranded region of interest within the pore's constriction for nucleotide identification, and the double stranded region will then be denatured due to the voltage, allowing complete passage of the single strands. In this way, they were able to distinguish nucleotides through differences in conductivity, even in a homopolymer region.

For the control of DNA velocity and orientation, the IBM "DNA transistor" offers an alternative approach for precise control of the DNA movement through the nanopore with single nucleotide accuracy.[67, 68] As shown in Fig. 1.9 E, the DNA transistor is a silicon based, multilayer metal/dielectric nanopore device with voltage generated inside the nanopore. By utilizing the interaction of discrete DNA backbone charges with a modulated electric field inside the nanopore, it can trap and slowly release the individual DNA molecules. This enables optimal base orientation and sufficient sampling of a base.

The similar chemical structure of the four nucleotides leads to their similar electronic signatures, which poses challenges to distinguish them in this way. This might be solved by either reducing background or amplifying signals through base modification or combination with other detection methods, such as optical detection.[56] With the aim of

overcoming the challenges of high signal contrast required for single nucleotide differentiation, Soni and Meller developed a nanopore single-molecule optical detection approach,[69] wherein the nucleotides A, T, G and C are first converted to a pair of 10-mer oligonucleotides with 2-unit codes of 0 and 1, the combination of which represents each respective nucleotide (for example, A is 1 1, G is 1 0, C is 0 0, T is 0 1). The converted DNA strand is then hybridized with two different 10-mer oligonucleotide "molecular beacons" designed to complement either 1 or 0, producing "amplified" fluorescence signals specific for each nucleotide. The emitted fluorescence of the molecular beacons is detected when the beacons are sequentially stripped off the complementary converted DNA strand passing through a nanopore.[69] These investigators also demonstrated the parallel readout of signals from multiple nanopores, which could be potentially applied to massively parallel large nanopore arrays.[70]

1.1.2.5.2 Tunneling and transmission electron microscopy based single molecule DNA sequencing

The interest in transmission electron microscope (TEM) based DNA sequencing traces back to Feynman's vision of directly reading DNA molecules by TEM in 1959, which stimulates people nowadays to study and achieve this goal despite many technological hurdles. Halcyon Molecular[71] developed a pioneering approach that could chemically detect atoms associated with nucleotides in a DNA strand by using annular dark-field electron microscopy, together with a number of supporting technologies, such

as stretching DNA on a substrate. ZS Genetics is trying to label atoms within the nucleotides, whereby individual bases can be identified according to their size and intensity differences under a high-resolution electron microscope.[9]

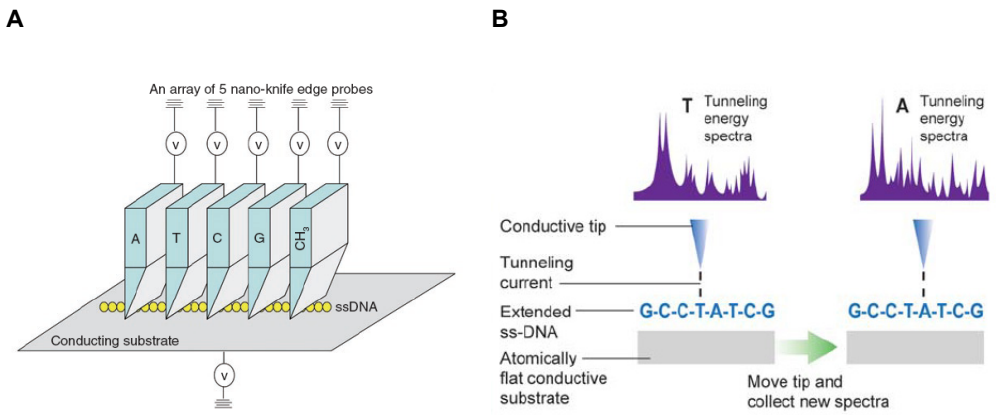


Fig. 1.10 (A) Reveo STM-based sequencing.[49] (B) DNA sequencing by direct inspection of DNA using electron microscopy [9]

Scanning tunneling microscopy (STM) could also be used to detect DNA bases according to the characteristic electronic differences among the four bases. Xu *et al.* has reported the electronic fingerprints of DNA bases deposited on a gold surface, such as bandgap and molecular energy levels.[72] Tanaka *et al.* later showed that guanine bases are brighter under STM by depositing the DNA molecule onto a copper surface with an oblique pulse-injection method.[73] With the assistance of single-walled carbon nanotubes, theoretically 100% base identification could be achieved.[74] Enhanced tunneling current was observed between a base specific modified STM tip and its complementary nucleotide. These indicate the possibilities of STM based DNA sequencing. Reveo is

now developing a technology called nano-knife-edge probes for reading sequence,[49] which utilizes the selective excitation of molecular vibrations by electron tunneling as its main principle. In their system, as shown in Fig. 1.10A, the DNA strand is immobilized and stretched in a 10 μm wide channel. Multiple nano-knife edge probes with 10 μm width which are tuned to recognize one specific nucleotide will pass over the DNA and identify the base according to the tunneling current generated when the probes touch the corresponding nucleotides.

These tunneling and transmission electron microscopy based single molecule DNA sequencing ideas appear straightforward, but the execution is going to be a long journey due to many factors, including preparing well-ordered and stretched ssDNA on the surface, the high cost of the microscopes, the construction of modified STM tips, and so on.

1.1.3 Conclusion

Since Sanger's introduction of the dideoxynucleotide chain terminator sequencing method, DNA sequencing has revolutionized the biological sciences and become one of the most powerful technologies in biology. Driven by the $1000 Genome project initiated by NHGRI, a variety of sequencing technologies have been explored, each with its own advantages and drawbacks. Among these new technologies, sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators (CF-NRTs) has shown the most promise for next generation sequencing, and has potential applications for single

molecule sequencing in the third generation platforms. Based on some of the ideas sprouted from our laboratory, part of this thesis will focus on the improvement of SBS with CF-NRTs, with the hope of pushing this particular platform towards the $1000 Genome goal. On the other hand, taking advantages of the high accuracy and high speed of mass spectrometry, we will continue to explore MS-based sequencing, aiming at its application in particular circumstances requiring high accuracy but short read-length.

1.2 Introduction to SNP genotyping technology

1.2.1 Background and Significance

The successful sequencing of the human genome has ushered in a new era of fine mapping of subtle genetic variations for elucidating the fundamental molecular bases of diseases, drug response and genotype changes. The most abundant type of these variations are single nucleotide polymorphisms (SNPs), with more than 10 million in public databases, occurring approximately once every 100 to 300 bases.[75] By definition, a SNP is a single nucleotide variation, including the substitution, insertion or deletion of individual bases, at a specific location in the genome that is found in more than 1% of the population.[76] These are usually distinguished by geneticists and epidemiologists from mutations, which may also be single nucleotide polymorphisms, but usually occur at lower frequencies in the population as a whole, and often have more significant impacts, either beneficial or more commonly deleterious. Typically, SNPs present in human nuclear genome as heterozygous (1:1 wild-type to mutant ratio) or homozygous, but in

the mitochondrial genome as heteroplasmy (mixture of wild-type and mutant forms at various ratios) or homoplasmy due to its high copy numbers. SNPs are not evenly distributed across the genome. In coding regions, they occur less frequently, but may cause alterations in protein structure and hence biological function, leading to the development of disease or a change in environmental response. They are important genetic markers for pharmocogenomics, such as SNPs (more properly mutations) in the breast cancer genes BRCA1 and BRCA 2. In the non-coding regions, SNPs occur much more frequently. Though not altering the encoded protein, they may affect gene regulation, and serve as important genetic or physical markers for comparative or evolutionary genomic studies, or even forensics investigations. The analysis of SNPs in the human genome will provide the key to understand genetic differences between individuals and disease states, and eventually realize personalized medicine by the prediction of genetically related disease risk and drug response genes. Therefore, tremendous efforts have been made to develop a variety of methods that allow the efficient and accurate genotyping of SNPs.

1.2.2 Overview of SNP genotyping technologies

Though direct DNA sequencing is the most straightforward way to discover and analyze SNPs, it is more suitable for genome-wide SNP studies and unknown SNP discoveries. When analyzing highly targeted SNPs, it becomes too expensive, complex and unnecessary. The following section will focus on current SNP genotyping

technologies, most of which are employed in the analysis of known SNPs.

1.2.2.1 SNP genotyping by enzymatic cleavage

The use of restriction fragment length differences generated by DNA sequence-specific restriction endonucleases to identify DNA polymorphisms for human linkage studies was first demonstrated by Botstein *et al*.[77] This, together with the introduction of PCR in the 1980s, has developed into a simple and convenient laboratory technique, polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP), for the detection of SNPs.[78] DNA restriction enzymes can recognize specific sequences in DNA and catalyze endonucleolytic cleavages, yielding fragments with defined lengths. The introduction of one or more individual base differences will lead to the loss of a cleavage site or formation of a new one, resulting in DNA fragments with different lengths, which can be detected by gel electrophoresis. As shown in Fig. 1.11 A, PCR primers are designed to allow amplification of a portion of the DNA template encompassing the polymorphic site. Then the DNA fragments amplified by PCR are digested with the restriction enzyme that will cleave template bearing different alleles differently, revealed as different fragment sizes by electrophoresis after staining with ethidium bromide (EB). To improve the sensitivity, radioactive nucleotides, like ^{32}P-labled nucleotides, are used in PCR reactions for radiographic detection.[79] This PCR-RFLP method is applicable not only for SNPs but also for insertion/deletion polymorphism detection, however it has limitations in throughput, number of SNPs

detected as well as the availability of appropriate restriction enzymes for the analysis.

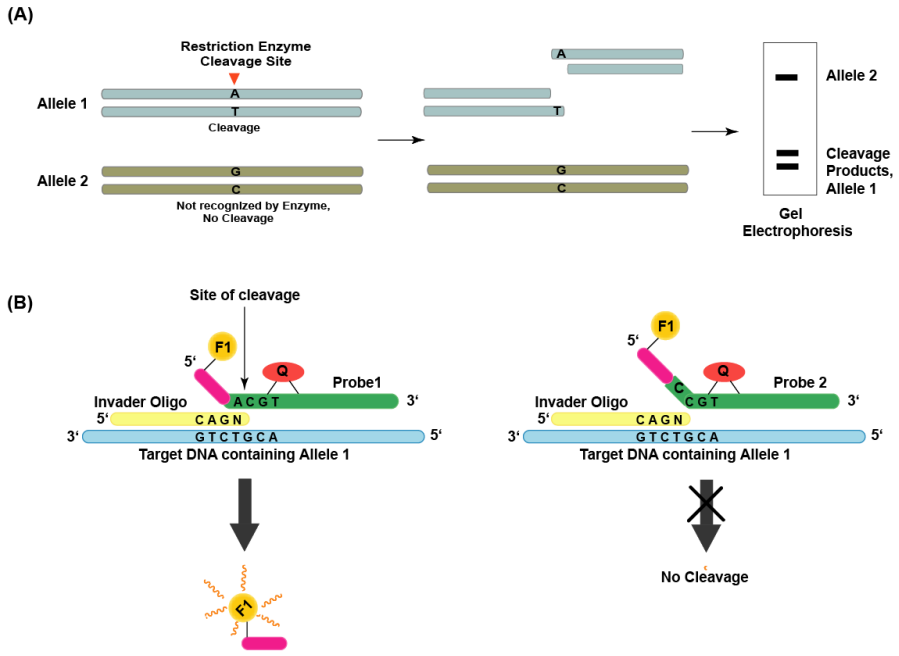


Fig. 1.11. Schematic diagram of SNP enzymatic cleavage assay. (A) PCR-RFLP assay; (B) Invasive cleavage assay.

Inspired by their discovery of thermostable flap endonucleases (FEN) in sequence mismatch sensitivity and structure-specific cleavage of DNA, Lyamichev and his colleagues developed the Invader® assay,[80] wherein FEN is used to cleave a three-dimensional complex formed by the hybridization of allele-specific overlapping oligonucleotides to target DNA with a polymorphic site, followed by the detection of cleavage products using several different approaches, such as fluorescent detection or mass spectrometric detection. An example of a fluorescence resonance energy transfer (FRET) detection based Invader assay is illustrated in Fig. 1.11 B. Two oligonucleotides are used here. One oligonucleotide, containing a fluorophore at its 5' end and an internal

quencher molecule, is the allele-specific probe, which consists of the complementary base to the SNP allele and additional sequence 5' of the SNP site. The other oligonucleotide strand is called the Invader® oligo, which is complementary to the target sequence immediately 3' of the polymorphic site, with a non-complementary base at the polymorphic site. When the two oligonucleotides hybridize to the target DNA, a three-dimensional invader structure over the SNP site is formed and can be recognized by the FEN enzyme. The enzyme cleaves the allele specific probe 3' of the base if it is complementary to the polymorphic site (3' of the overlapping Invader structure), thereby the cleavage reaction will separate the fluorophore from the quencher and generate a measurable fluorescent signal. If the allele specific probe does not match the SNP allele, then no overlapping Invader structure is formed, and the probe is not cleaved. This gives a highly specific signal.[81, 82] Hall *et al.* further improved this strategy by combining two invasive cleavage reactions into a single homogeneous assay, which is called the Serial Invasive Signal Amplification Reaction (SISAR).[83] However, this strategy also has limitations in throughput and multiplexing, since two allele-specific probes with both fluorophore and quencher molecules attached are required for the detection of one polymorphic site, making it cost-prohibitive for constructing probes for more SNP sites.

1.2.2.2 SNP genotyping by allele specific hybridization

As mentioned in the first section, allele specific hybridization has been mostly explored for SNP genotyping. Equivalent to sequencing by hybridization, in principle,

single base distinction can be achieved based on the thermal stability difference of double-stranded DNA between perfectly matched and mismatched target-probe pairs. Generally, the effectiveness in allele differentiation depends on many factors, such as length, sequence and GC content of the probes, secondary structures of the target, location of SNP in the probe, and hybridization conditions. And the selection of analyzable SNPs is highly dependent on the local SNP sequence, which, to some extent, limits the universal applicability of this method.

Allele specific hybridization has been developed in different ways. High density DNA microarrays, like the Genome-Wide Human SNP Array by Affymetrix, are a typical hybridization based approach for genome-wide or large scale SNP screening. Briefly, taking the Affymetrix GeneChip as an example,[84] a large set of nucleic acid probe sequences, matching one of the two SNP alleles, is immobilized in defined locations on the solid substrate, enabling acquisition of large amounts of genetic variation information in a single hybridization step. The DNA or RNA target is fragmented and labeled with a fluorophore, and then applied to the DNA array for hybridization with the surface-bound probes under controlled conditions, as shown in Fig. 1.12 A. The array is then imaged with a fluorescence-based reader to locate the target sequence. With the development of microfluidic and automation systems, DNA microarrays have enabled high throughput SNP detection; however, they are not suitable for fine mapping of SNPs due to their insufficient level of accuracy.

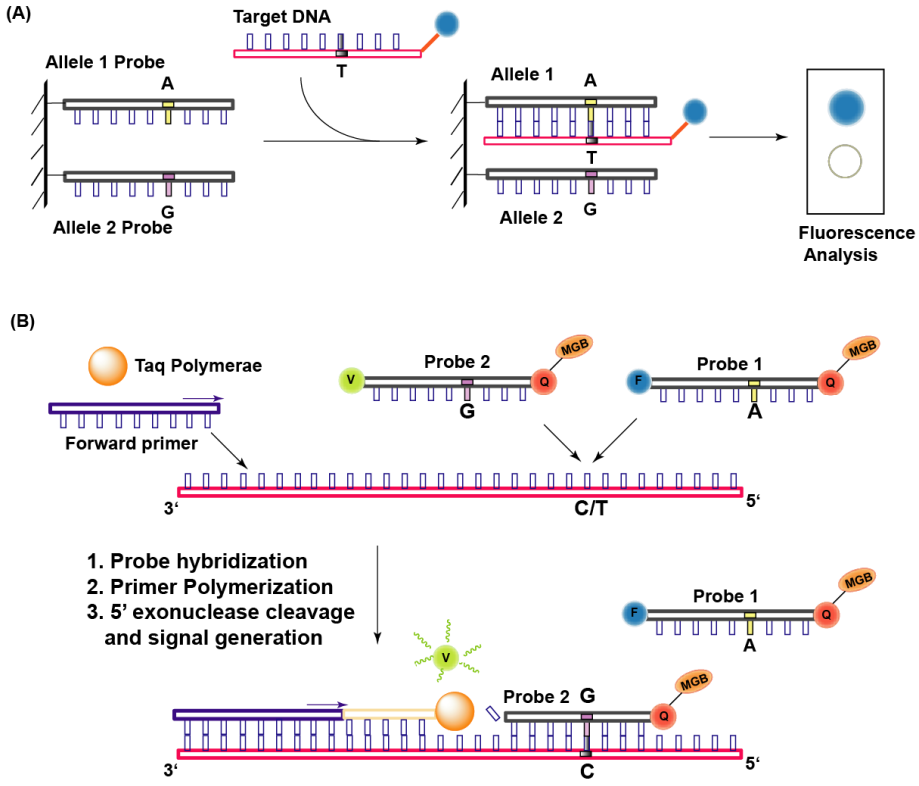


Fig. 1.12 Approaches of SNP genotyping by allele specific hybridization. (A) DNA microarray hybridization; (B) TaqMan assay.

Another example is the TaqMan SNP assay,[85] an oligonucleotide hybridization based assay capable of detecting the accumulation of PCR products in real time. As shown in Fig. 1.12 B, forward and reverse primers are designed to amplify the sequence of interest encompassing the SNP site, and two dye-labeled probes are used for allele-specific detection. Each allele specific probe is labeled with a different reporter dye at its 5' end: a VIC® dye for the Allele 1 probe and the 6FAM™ dye for the Allele 2 probe, while their 3' ends are labeled with a non-fluorescent quencher (NFQ). This allows the detection of the fluorescent signal only if the fluorophore dye and NFQ are separated. A minor groove binder (MGB) is also introduced to the allele specific probes at the 3' end, in order to

improve the stability and specificity of the hybridization and enhance the ability of the assay to handle difficult sequences. As the DNA primer extends to the 5' end of the allele specific probes, the 5' to 3' exonuclease activity of the Taq polymerase enzyme will degrade the nucleotide from the 5' end, hence release the 5'-bound reporter dyes from the probe, which will light up for allele identification. It is a relatively easy and convenient assay, yet it is currently used mostly for singleplex reactions: one tube for one SNP. Though it might be multiplexed to 4 or 5 SNPs per reaction, it is hard to achieve detection of more SNPs at a time due to the limitation of using fluorophores. The introduction of miniaturized reactions has enabled over 3000 samples to be tested on a chip, however the main drawback for the TaqMan SNP assay remains in its incapability of achieving multiplexing of SNPs.[86]

1.2.2.3 SNP genotyping by allele specific ligation

Combining the hybridization of oligonucleotides to the sequence of interest and the specificity of the ligase enzyme to distinguish mismatched nucleotides in the DNA double helix, a ligation strategy for SNP detection was firstly developed by Landegren *et al.*[87] In principle, only when two oligonucleotides hybridize to a single-stranded template DNA with perfect matching (the nucleotides at the junction are correctly base-paired with the target strand) and adjacent to each other (the 3' end of one oligonucleotide lies immediately adjacent to the 5' end of the other), DNA ligase will join them into a single oligonucleotide through the formation of a phosphodiester bond between the upstream

and downstream oligonucleotides. Three oligonucleotide probes are used, two of which are allele specific and bind to the SNP site. The third probe binds to the template adjacent to the SNP site, allowing ligation reaction if the allele specific probe perfectly matches the SNP site. The ligation products are then detected by various tools based on the labeling method for the oligonucleotides, which has been explored and gradually improved in different ways. In Landegren's original idea, one of the oligonucleotides was biotinylated and the other labeled with ^{32}P, so the biotinylated oligonucleotides would bind to a streptavidin coated solid support, and the radioactive oligonucleotides that have become ligated to the biotinylated oligonucleotides and remained on the support after washing were detected by autoradiography. Xue *et al.* introduced a silver enhancement chip-based platform for SNP screening, with one DNA probe immobilized on a glass substrate, and the other oligonucleotide probe attached to a gold nanoparticle (NP). The ligation allows the NP probe staying attached to the substrate, enabling detection by the naked eye or a flatbed scanner.[88] Notwithstanding these developments, the most common approach nowadays is fluorescent labeling. Taking fluorescence labeled oligonucleotides as an example (Fig. 1.13 A), two allele specific probes were tagged with different fluorescent dyes, with each representing one allele. The third probe could be annealed downstream to the polymorphic site and is biotinylated for separation of ligated products from the reaction mixture. After the ligation reaction and solid phase capture, the allele specific probe that is complementary to the polymorphic site will stay attached to the solid surface through ligation with the biotinylated probe and generate a fluorescent

signal for detection, while the one that does not match the polymorphic site will be washed away.

By combining ligation for allele discrimination and rolling circle amplification (RCA) for signal enhancement, Qi *et al.* proposed a method called ligation-rolling circle amplification (L-RCA).[89] As shown in Fig. 1.13 B, short padlock probes coupled with generic primers are used, the circulation of which by T4 ligase or thermostable ligase can be used to discriminate alleles in target DNA sequences. The circulation product is amplified by DNA polymerase and visualized by UV illumination after staining. Pickering *et al.* further introduced energy transfer primers for rolling circle amplification, avoiding the gel electrophoresis.[90] In their method, the circularized probes are detected by amplification using two primers: the first primer hybridizes to its complementary region on the probe backbone, and is extended by polymerase to generate a single-stranded concatamer of the original probe; while the second primer, which contains a 5'-hairpin loop end labeled with a fluorophore and quencher, binds to each tandem repeat of the original probe, resulting in strand displacement and branching of the RCA products, as well as a strong fluorescent signal due to the separation of the fluorophore and quencher when the hairpin stem is displaced by DNA polymerase during extension.

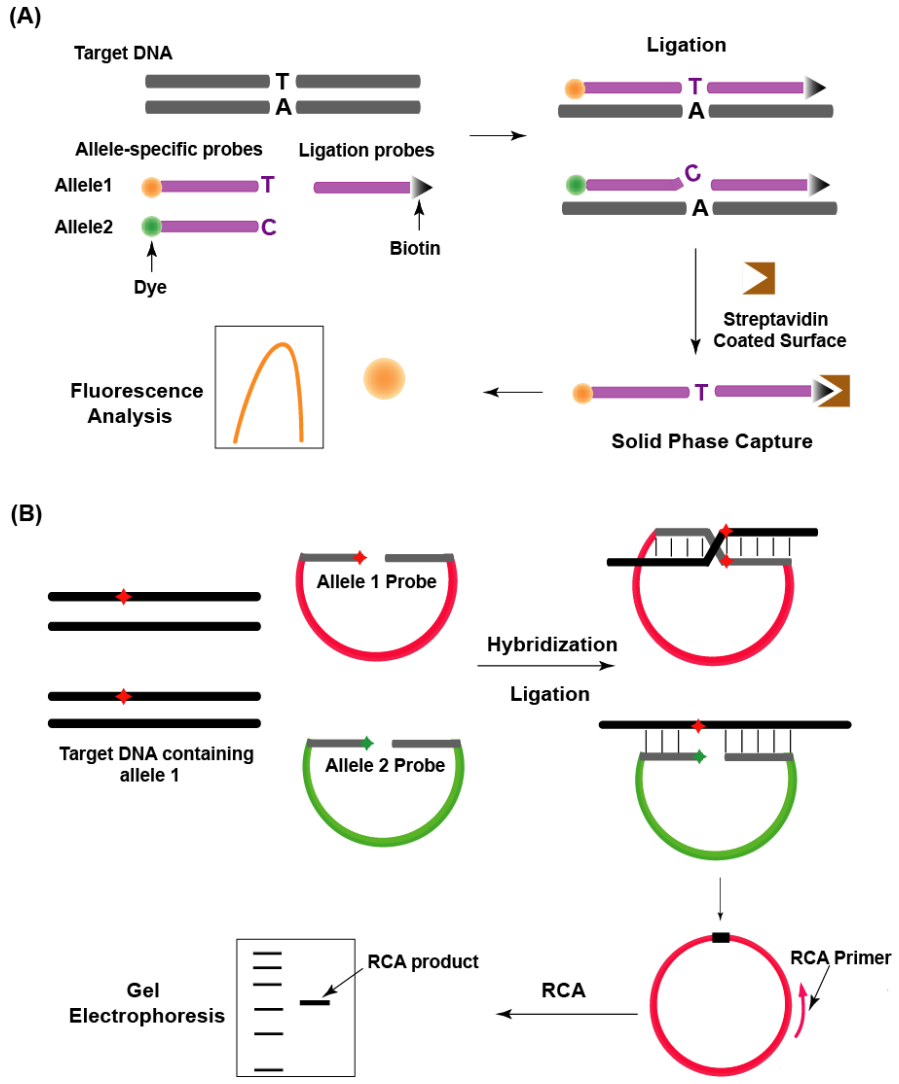


Fig. 1.13 Methods for SNP genotyping by ligation. (A) Fluorescence based ligation assay; (B) Ligation-rolling circle amplification assay.

1.2.2.4 SNP genotyping by allele specific primer extension

Allele specific primer extension has been widely explored for the development of effective and accurate genotyping tools. Generally, it involves allele-specific incorporation of nucleotides in primer extension with a DNA template, utilizing enzyme specificity to achieve allele discrimination. In principle, a primer is designed to anneal to the target DNA, with its 3'end immediately adjacent to the polymorphic site, and then

extended with nucleotides by polymerase enzyme. The identification of the base at the SNP site is based on the specific detection method, whether fluorescent or mass spectrometric detection. Compared to previous methods, it gives a more specific signal, because of the high fidelity of DNA polymerase in incorporating complementary nucleotides. It is also suitable for multiplex SNP detection, with multiple primers of different length designed for each target SNP site, and the primer selection and assay design is simple and flexible.[91]

The approaches that are based on fluorescence detection involve the utilization of fluorescently labeled nucleotides. As shown in Fig. 1.14 A, after single base extension reaction, the primer is end labeled with fluorescent dyes by incorporating fluorescently tagged nucleotides, which could then be visualized to determine the base at the polymorphic site according to the reporter dye. The fluorescence based allelic extension approach has been implemented in different formats, including homogeneous detection and solid phase detection. For example, Illumina's Infinium assay uses single-base extensions on a DNA array with a labeled base to call the SNP,[86] and the SNaPshot® approach developed by Applied Biosystems uses capillary array electrophoresis to detect fluorescently labeled extension products using fluorescently labeled ddNTPs.[91] In contrast to the direct detection of a fluorescence signal, the mini-sequencing assay developed by Perkin Elmer, fluorescence polarization-template-directed dye incorporation (FP-TDI), monitors the fluorescence polarity of the fluorescent nucleotide terminators. In their system, after PCR amplification of the SNP containing sequence, a

third primer (single base extension primer) is annealed immediately adjacent to the polymorphic site, and extended with allele specific fluorescent nucleotide terminators by polymerase. The mode of detection is to monitor the change in fluorescence polarization, according to the different fluorescence polarization properties of the incorporated and unincorporated nucleotides: under polarized fluorescent light, the incorporated nucleotide would cause the light to become depolarized, while the dyes incorporated into the primer will keep most of the light polarized. This avoids the extra step of clean-up of the excess nucleotides.[86]

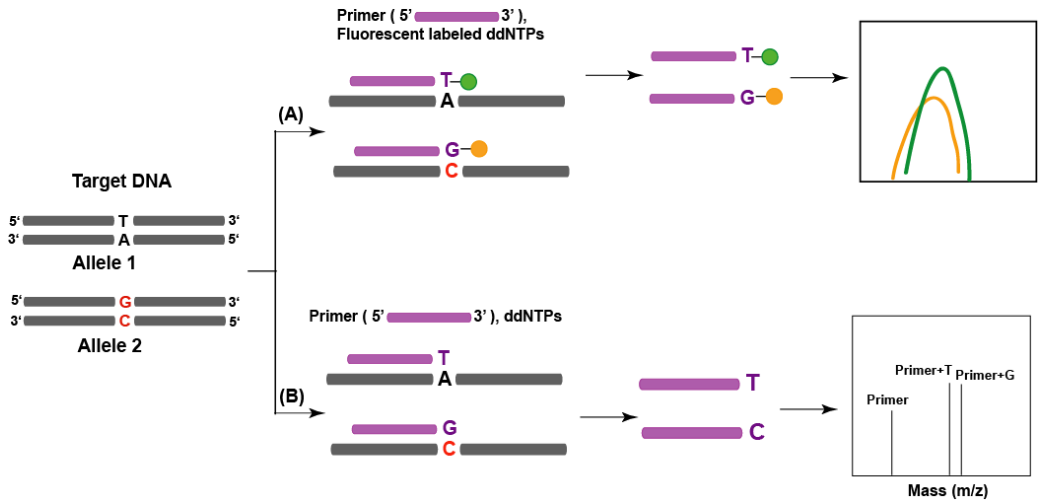


Fig. 1.14 Allele specific primer extension. (A) Fluorescence based detection; (B) Mass spectrometry based detection.

The combination of high density bead array, allele-specific primer extension, adapter ligation and amplification assay protocols propelled the development of the GoldenGate genotyping assay by Illumina. In this system, two allele specific oligonucleotides (ASO) (P1 and P2) and one locus specific oligonucleotide (LSO) (P3) that will anneal 1-20

bases downstream of the SNP site are used, all of which contain universal PCR primer sites. The LSO also contains a unique code sequence complementary to a particular bead type. In their method, genomic DNA is first biotinylated and attached to the streptavidin beads. In a single base extension, the allele specific oligonucleotide (P1 or P2) that perfectly matches the target sequence at the SNP site is extended by polymerase up to the site of the locus-specific oligonucleotides (P3) and ligated to P3, providing the template for PCR reactions. In subsequent PCR reactions, two universal primers complementary to the priming sites in P1 and P1 are labeled with Cy3 and Cy5 respectively, which serve as the allele-specific fluorescence reporters. The final PCR product will bind to the bead array through the built-in encoding sequence in P3 for accurate genotype calling.[92, 93]

The other big category of allele specific extension approaches couple with mass spectrometry, taking advantage of the high efficiency and accuracy of mass spectrometric detection. As shown in Fig. 1.14 B, SNP-specific primers are extended with dideoxynucleotides using PCR products as the template to yield extension products of different molecular weights, which then are analyzed by mass spectrometry to identify the incorporated nucleotides. The unmodified dideoxynucleotides and primers were first used to run single base extension in the PinPoint assay, the result of which also demonstrated the potential of MALDI-TOF based methods for multiplex SNP detection.[94, 95] However, the peaks of excess primers and the small molecular weight difference between each nucleotide reduced the accuracy of the detection. Modifications to the extension primers or nucleotides were gradually developed to overcome these

problems and commercial platforms were developed.[96-98] In the GOOD assay, primers with 3' phosphorothioate modifications and α-S-ddNTPs were used to generate charge-tagged DNA fragments. This eliminates the need for sample purification before MS analysis and improves the resolution by using shorter extension fragments.[96] Nevertheless, the minuscule differences between ddNTPs remained the limitation for base discrimination. Mass-tagged ddNTPs were introduced as the alternative solution.[99] In combination with the solid phase capture (SPC) technique, the Ju group further introduced a modified SPC-SBE approach for isolation of extension products using molecular affinity between streptavidin and biotin, wherein the biotinylated nucleotides were used in single base extension, and the replacement of regular biotin-11-ddUTP with biotin-16-ddUTP (longer linkage between biotin and nucleotide) was shown to improve the resolution between C and T.[100] However, the sample purification by solid phase capture before the MS analysis as well as the mass-tagged strategy still needs further improvement to realize high throughput, multiplex genotyping of high accuracy and sensitivity. We will discuss our modifications to accomplish this goal in Chapter 4.

1.2.3 Conclusion

Single nucleotide polymorphisms are the most common genetic variations, and the development of an accurate, precise, time-efficient and cost-effective SNP genotyping method is of particular interest to genetics, medicine and other areas of biology. Different types of SNP genotyping methods were reviewed, each with its advantages and

limitations with respect to specific applications. For a particular project, one should choose the type of assay based on the number of SNPs and number of samples to be genotyped. The mass spectrometry based single base extension approach outperforms many of the others in terms of its speed, accuracy, and potential for multiplexing, which is particularly applicable for relatively small scale projects with a high accuracy and sensitivity requirement. Part of my thesis focuses on the development of a reversible solid-phase-capture single-base-extension approach for SNP genotyping and its application in heteroplasmy detection of mitochondrial DNA as the validation of its accuracy and sensitivity.

1.3 Introduction to the detection of nucleic acids by oligonucleotide probes

1.3.1 Background and significance

DNA sequencing and genotyping technologies reveal the genetic information inherent in nucleic acids but are not able to tell us how these biomolecules are involved in processes within the cell. Visualization of DNA or RNA in cells and tissues can suggest possible roles of these molecules in physiological and pathological states, via the information about their location, kinetics, etc. For example, the real-time detection of specific mRNAs can provide information on their localization and expression level, hence suggest how they are involved in post-transcriptional processes in living cells.[101, 102] Nevertheless, direct observation of cellular DNA, especially RNA, remains

particularly difficult, mainly due to the non-specific interaction of probes with other cellular components, the high background signal inside the cell, and particularly for mRNAs, their abundance and short half-lives, and the accessibility of specific sites in mRNA molecules with highly ordered structures. Therefore, development of novel, sensitive and selective sensors for the detection of DNA and RNA is of particular interest in biology. In the following section, we will only discuss methods for visualizing nucleic acids inside cells, not for their quantitation.

1.3.2 Overview of oligonucleotide probes for detection of nucleic acids

The selective detection of specific DNA and RNA sequences can be achieved by using oligonucleotide-based antisense hybridization probes.[103] In general, the sequences of these oligonucleotide probes are complementary to the target sequences, and the attachment of a reporter group to these probes enables their visualization by a spectroscopic technique, such as fluorescence spectroscopy.[104] Two of the typical categories of these oligonucleotide probes are molecular beacons (MBs) and binary probes (BPs), as will be described in the following section.

1.3.2.1 Molecular beacons

The MB concept was introduced in 1996 by Tyagi and Kramer[105] for nucleic acid hybridization assays and is among the most promising technologies under development for quantitative nucleic acid detection *in vitro* and *in vivo*. As shown in Fig. 1.15 A, a MB is a fluorophore and quencher labeled (or dual fluorophore-labeled) DNA hairpin

comprised of a probe region (loop) complementary to a target sequence and a self-complementary region of five to six nucleotides at the opposite ends (stem). In the absence of the target, the complementary parts of the probe hybridize and form a hairpin conformation, thereby bringing the fluorophore (donor) and quencher (acceptor) into close proximity. This interaction results in strong fluorescence quenching through the deactivation of the fluorophore excited state through förster resonance energy transfer (FRET). In the presence of the target, the hairpin structure opens up and the loop region hybridizes to the target DNA sequence, resulting in the separation of the fluorophore and quencher accompanied with brightening of the fluorophore.

An ideal MB should not exhibit any fluorescence emission in the "closed" hairpin conformation but a strong fluorescence emission in the "open" conformation when hybridizing with the target. The efficiency of target detection of a MB is measured by its signal-to-background ratio (S/B), which is the ratio of the fluorescent signal in the presence of the target to the fluorescent signal before the addition of the target.[106] Molecular Beacons have enabled increased assay detection by directly measuring the fluorescence changes rapidly after DNA probe hybridization and have been exploited in various biological applications. However, the opening of the hairpin and separation of the fluorophore and the quencher evoked by hybridization of this oligonucleotide to the target probe region is not easily adapted for *in vivo* studies because of interference from unbound probes, non-specific hybridization and autofluorescence from the biomolecules in the sample. These result in unacceptably high background signals and false positive

signals.[105, 107, 108] Furthermore, in the probe-quencher configuration, the quenching of the fluorophore in the hairpin state makes it difficult to track the injection of MBs inside the cells.

With dual fluorophore-labeled molecular beacons, a signal due to FRET from acceptor to donor will be observed in the hairpin state since the two fluorophores are thereby brought close to each other, and the hairpin conformation will evoke a strong donor emission. While this makes it possible to visualize the injection of MBs, non-specific opening of the MBs remains a significant challenge.[104]

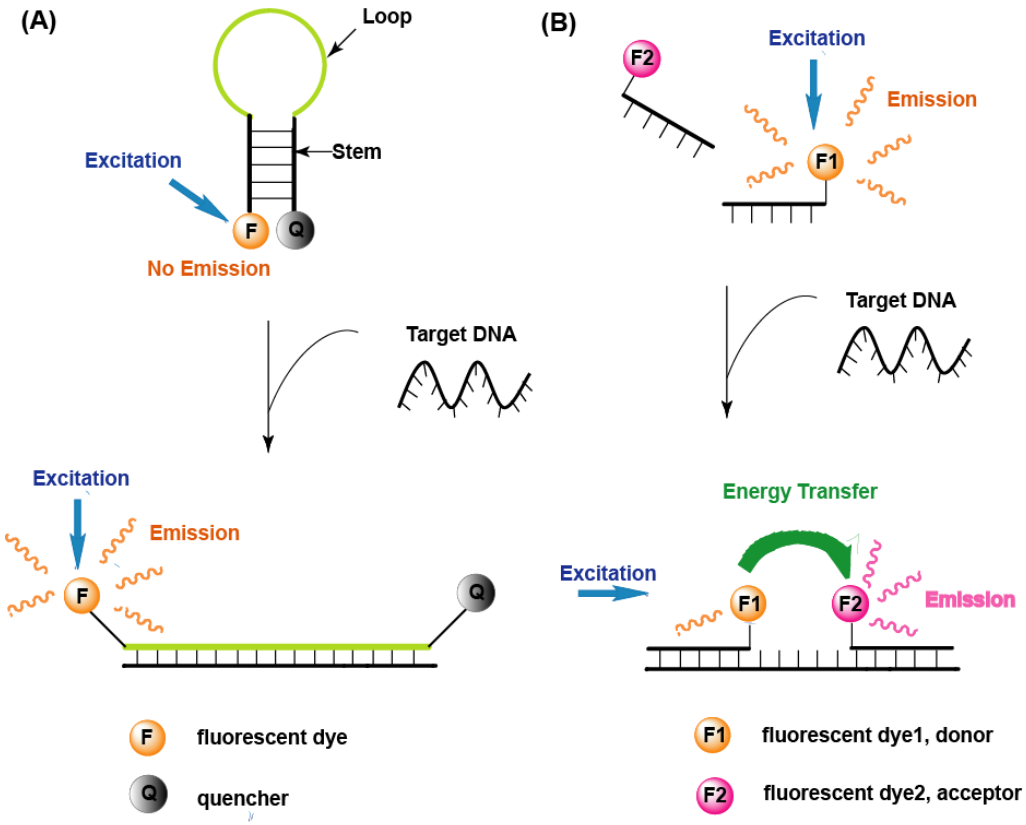


Fig. 1.15. Principle of oligonucleotide probes for detection of nucleic acid. (A) molecular beacons; (B) binary probes.

1.3.2.2 Binary oligonucleotide probes

An alternative strategy to the use of molecular beacons is the binary oligonucleotide probes, which have been shown to display high specificity within living cells since the requirement for efficient FRET relies on the strict spatial requirements of the fluorescent donor and acceptor.[109] As shown in Fig.1.15 B, the BP system is composed of two single-stranded DNA (ssDNA) molecules which are complementary to adjacent regions of a common target sequence where a donor and acceptor are attached at opposite ends. When the pair of BPs selectively hybridizes to their targets, the two oligonucleotide strands are drawn into close proximity and a distinctive FRET signal is elicited. The resulting FRET signal is different from that of the non-hybridized probes, and nonspecific binding to cellular components will be unlikely to promote FRET because of the strict spatial requirements. Therefore, higher specificity is shown to be an advantage of the BP over the MB approach.

In the ideal 2-color binary probe system, only the fluorescence from the donor fluorophore is observed in the absence of the target, while in the presence of the target, only the fluorescence of the acceptor fluorophore is observed because of FRET from the donor. Hence, the BP system undergoes a color switch when changing from random distribution in the solution state to the hybridization state of binding to the target. Similar to molecular beacons, the efficiency of BPs could also be measured by S/B ratio, which can be generally described as the ratio of the fluorescence signal of the acceptor to that of the donor in the presence of target divided by the ratio of the fluorescence intensity of the

acceptor to that of the donor in the absence of the target (S/B = ($F_{Acceptor+Target}$* F_{Donor})/($F_{Donor+Target}$*$F_{Acceptor}$)).[110] The S/B ratio relies on the efficiency of FRET, which depends on various parameters, including the distance between the donor and acceptor fluorophores and the spectral overlap between the donor emission and the acceptor absorption spectra.[111] Theoretically, it requires that the emission band of the donor should overlap as much as possible with the excitation band of the acceptor, but the excitation band of the donor should be far from that of the acceptor to reduce direct excitation of the acceptor leading to a background signal. However, no available dye pairs fulfill the above criteria perfectly and sometimes it is necessary to compromise energy transfer efficiency in order to reduce direct excitation. The other challenge of binary probes is their slow kinetics of hybridization, because both probes need to hybridize to the target, resulting in a larger entropy loss compared with molecular beacons. Although people have been trying to construct different binary probes to address these problems, such as connected binary probes[112] and pyrene excimer probes,[104] and there are various biological applications of BPs so far, a fundamental improvement in the technology is still needed to allow tracking of typical mRNAs in living cells.

1.3.3 Conclusion

Due to the great potential of *in vivo* and *vitro* detection of nucleic acids, oligonucleotide-based probes, molecular beacons and binary probes, have engendered an intensive research field with novel and creative approaches. However, the effective use of

these probes requires overcoming challenges such as non-specific opening of MBs, injection and distribution visualization of MBs within the cell, a low S/B ratio in BPs due to direct excitation of the acceptor fluorophore and/or overlap of the fluorescence spectra of the dyes, low kinetics of BPs, cell autofluorescence, and others. Modification of the classical architecture of MBs or BPs has allowed for the enhancement of detection capabilities. However, many improvements are necessary to advance their application. Because of the high specificity of 2 dye-binary probes and the ease of tracking them *in vivo*, part of my thesis focuses on the development of novel binary probes with quantum dot-based donor molecules and organic dye-based acceptors to address some of the challenges of traditional binary probes.

References

1. Watson JD, Crick FHC. A structure for deoxyribose nucleic acid. *Nature*, **1953**, *171*, 737-738.

2. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **1975**, *94*, 441-448.

3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science of the United States of the America*, **1977**, *74*, 5463-5467.

4. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, FIDDES JC, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriphage phi x174 DNA. *Nature,* **1977**, *265*, 687-695.

5. http://www.grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html

6. Huang XC, Quesada MA, Mathies RA. DNA sequencing using capillary array clectrophoresis, *Anal Chemistry*, **1992**, *64 (18)*, 2149-2154.

7. Ju J, Ruan C, Fuller CW, Glazer AN, Mathies RA. Fluorescence energy transfer dye-labeled

primers for DNA sequencing and analysis. *Proceedings of the National Academy of Sciences of the United States of America,* **1995**, *92 (10)*, 4347-4351.

8. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequencing analysis, *Nature*, **1986**, *321*, 674-679.

9. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Human Molecular Genetics*, **2010**, *19*, R227-R240.

10. Spengler B, Yang Y, Cotter J, Kan L-S. Molecular weight determination of underivatized oligodeoxyribonucleotides by positive-ion matrix-assisted ultraviolet laser-desorption mass spectrometry. *Rapid Communications in Mass Spectrometry*, **1990**, *4*, 99-102.

11. Murray KK. DNA sequencing by mass spectrometry. *Journal of Mass Spectrometry*, **1996**, *13*, 1203-1215.

12. Edwards JR, Itagaki Y, Ju J. DNA sequencing using biotinylated dideoxynucleotides and mass spectrometry. *Nucleic Acids Research*, **2001**, *29*, e014.

13. Edwards JR, Ruparel H, Ju J. Mass-spectrometry DNA sequencing. Mutation Research, **2005**, *573*, 3-12.

14. Pieles U, Zürcher W, Schär M, Moser HE. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res*, **1993**, *21*, 3191-3196.

15. Bentzley CM, Johnston MV, Larsen B, Gutteridge S. Oligonucleotide sequence and composition determined by matrix-assisted laser desorption/ionization. *Analytical Chemistry*, **1996**, *68*, 2141-2146.

16. Smirnov IP, Roskey MT, Juhasz P, Takach EJ, Martin SA, Haff LA. Sequencing oligonucleotides by exonuclease digestion and delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Analytical biochemistry*, **1996**, *238,* 19-25.

17. Kirperkar F, Douthwaite S, Roepstorff P. Mapping posttranscriptional modifications in 5S ribosomal RNA by MALDI mass spectrometry. *RNA,* **2000,** *6*, 296-306.

18. Hahner S, Lüdemann H-C, Kirpekar F, Nordhoff E, Roepstorff P, Galla H-J, Hillenkamp F. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) of endonuclease digests of RNA. *Nucleic Acids Research*, **1997**, *25* (10), 1957-1964.

19. Mauger F, Bauer K, Calloway CD, Semhoun J, Nishinoto T, Myers TW, Gelfand DH, Gut IG. DNA sequencing by MALDI-TOF MS using alkali cleavage of RNA/DNA chimeras. *Nucleic Acid Research*, **2007**, *34(3)*, e18.

20. Castleberry CM, Chou C-W, Limbach PA. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of oligonucleotides. *Current Protocols in Nucleic Acid Chemistry*, **2008**, 10.1.1-10.1.21.

21. Drmanac R, Drmanac S, Chui G, *et al*. Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Advances in Biochemical Engineering/Biotechnology*, **2002**, *77*, 75-101.

22. Drmanac R, Crkvenjakov R. Method of sequencing of genomes by hybridization with oligonucleotide probes. **1987**, Yugoslav patent application 570/87.

23. Drmanc R, Crkvenjakov R. Method of sequencing of genomes by hybridization with oligonucleotide probes. **1989**, US patent 5,202,231.

24. Drmanac S, Kita D, Labat I, Hauser B, Schmidt C, Burczak JD, Drmanac R. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nature Biotechnology*, **1998**, *16*, 54-59.

25. Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology*, **2008**, *26 (10)*, 1135-1145.

26. Gresham D, Dunham MJ, Botstein D. Comparing whole genomes using DNA microarrays, *Nature Reviews Genetics*, **2008**, *9*, 291-302.

27. Hyman ED. A new method of sequencing DNA. *Analytical Biochemistry*, **1988**, *174*, 423-436.

28. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **2001**, *11*, 3-11.

29. Ju J, Li Z, Edwards J, Itagaki Y, **2003**, *US Patent* 6,664,079

30. Harris TD, Buzby P, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, DiMeo J, Efcavitch JW, *et al.* Single-molecule DNA sequencing of a viral genome. *Science*, **2008**, *320*, 106-109

31. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science*, **2009**, 323, 133-138.

32. Shendure J, Porreca GJ, Reppas NB, Lin X, Mccutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **2005**, *309*, 1728-1732.

33. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity, *Analytical Biochemistry*, **1987**, *167*, 235-248.

34. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry,* **1996**, *242*, 84-89.

35. Ronaghi R, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. *Science*, **1998**, 281, 363-365.

36. Ronaghi M. Improved performance of pyrosequencing using single-stranded DNA-binding protein. *Analytical Biochemistry*, **2000**, *286*, 282-288.

37. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **2005**, *437*, 376-480.

38. Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, **2009**, *85*, 142-154.

39. Pattersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. *Genomics*, **2009**, *93*, 105-111.

40. Kircher M, Kelso J. High-throughput DNA sequencing – concepts and limitations. *Bioessays*, **2010**, *32*, 524-536.

41. Metzker ML. Sequencing technologies – the next generation. *Nature Reviews Genetics*, **2010**, *11*, 31-46.

42. Metzker ML. Emerging technologies in DNA sequencing. *Genome Research*, **2005**, *15*, 1767-1776

43. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Science of the United States of America*, **2006**, *103*, 19635-19640.

44. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers

DJ, Barnes CL, Bignell HR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **2008**, *456*, 53-59.

45. Guo J, Yu L, Turro NJ, Ju J. An integrated system for DNA sequencing by synthesis using novel nucleotide analogues. *Accounts of Chemical Research*, **2010**, *43*, 551-563.

46. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequencing information can be obtained from single DNA molecules. *Proceedings of the National Academy of Science of the United States of America*, **2003**, *100*, 3960-3964.

47. Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, *et al*. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods*, **2009**, *6*, 593-595.

48. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, **2009**, *27*, 847-852.

49. Blow N. DNA sequencing: generation next-next. *Nature Methods*, **2008**, *5*, 267-274.

50. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **2003**, *299*, 682-686.

51. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructure. *Proceedings of the National Academy of Science of the United States of America,* **2008**, 105, 1176-1181.

52. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology*, **2010**, *472*, 431-455.

53. McCarthy A. Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. *Chemistry & Biology*, **2010**, *17*, 675-676

54. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, et al. The origin of the Haitian Cholera Outbreak Strain. *The New England Journal of Medcine*, **2011**, *364*, 233-242.

55. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, **2010**, *7 (6)*, 461-465.

56. Xu M, Fujita D, Hanagata N. Perspectives and Challenges of emerging single-molecule DNA sequencing technologies. *Small*, **2009**, *5 (23)*, 2638-264.

57. www.visigenbio.com

58. McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano lett*ers, **2000**, *10*, 2237-2244.

59. Kasianowica JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the Academy of Science of the United States of the America,* **1996**, *93*, 13770-13773.

60. Branton D, Deamer DW, Marziali Andre, Bayley H, Benner SA, Butler T, Ventra MD, Garaj S, Hibbs A, Huang X, et al. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, **2008**, *26 (10)*, 1146-1153.

61. Rhee M, Burns MA. Nanopore sequencing technology: nanopore preparations. *Trends in Biotechnology*, **2007**, *25*, 174-181

62. Astier Y, Braha O, Bayley H. Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonulceoside 5'-monophosphates by using an engineered protein nanoopore equipped with a molecular adapter. *Journal of the American Chemical Society,* **2006**, *128*, 1705-1710.

63. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, **2009**, *4*, 265-270.

64. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United of States*, **2009**, *106(19)*, 7702-7707.

65. http://www.nanoporetech.com/sections/index/80

66. Derrington IM, Butler TZ, Collins MD, Manrao E, pavlenok M, Niederweis M, Gundlach. Nanopore DNA sequencing with MspA. *Proceedings of the Academy of Science of the United States of the America*, **2010**, *107*, 16060-16065.

67. Harrer S, Ahmed S, Afzali-Ardakani A, et al. Electrochemical Characterization of thin film electrodes toward developing a DNA transistor. *Langmuir,* **2010**, *26 (24)*, 19191-19198.

68. Polonsky S, Rossnagel S, Stolovitzky G. Nanopore in metal-dielectric sandwich for DNA position control. *Applied Physics Letters*, **2007**, *91*, 153103.

69. Soni GV, Meller A. Progress toward ultrafast DNA sequencing using solid-state nanopores. *Clinical Chemistry*, **2007**, *53 (11)*, 1996-2001.

70. McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano letters*, **2010**, *10*, 2237-2244.

71. www.Halcyonmolecular.com

72. Xu M, Endres RG, Arakawa Y. The electronic properties of DNA bases. *Small*, **2007**, *3*, 1539-1543.

73. Tanaka H, Kawai T. Partial sequencing of a single DNA molecule with a scanning tunneling microscope. *Nature Nanotechnology*, **2009**, *4*, 518-522.

74. Meng S, Maragakis P, Papaloukas C, Kariras E. DNA nucleoside interaction and identification with carbon nanotubes. *Nano letter*, **2007**, *7*, 45-50.

75. http://www.ncbi.nlm.nih.gov/SNP

76. Brookes AJ. The essence of SNPs. *Gene*, **1999**, *234*, 177-186.

77. Botstein D, White RL, Skolnick M, David RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *The American Journal of Human Genetics*, **1980**, *32*, 314-331.

78. Ota M, Fukushima H, Kulski JK, Inoko H. Single nucleotide polymorphism detection by polymerase chain reaction-restriction fragment length polymorphism. *Nature Protocols*, **2007**, *2 (11)*, 2857-2864.

79. Naini A, Shanske S. Detection of mutations in mtDNA. *Methods in Cell Biology*, **2007**, *80*, 437-463.

80. Lyamichev V, Mast AL, Hall JG, Prudent JR, Kaiser MW, Takova T, Kwiatkowski RW, Sander TJ, de Arruda M, Arco DA, Neri BP, Brow MA. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nature Biotechnology*, **1999**, *17*, 292-296.

81. Olivier M, The invader® assay for SNP genotyping. *Mutation Research*, **2005**, *573 (1-2)*, 103-110.

82. Che Y, Shortreed MR, Olivier M, Smith LM. Parallel single nucleotide polymorphism genotyping by surface invasive cleavage with universal detection. *Analytical Chemistry*, **2005**, *77*, 2400-2405.

83. Hall JG, Eis PS, Law SM, Reynaldo LP, Prudent JR, Marshall DJ, Allawi HT, Mast AL, Dahlberg JE, Kwiatkowski RW, de Arruda M, Neri BP, Lyamichev VI. Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction. *Proceedings of the National Academy of Science of the Unites of America,* **2000**, *97*, 8272-8277.

84. McGall GH, Christians FC. High-density GeneChip oligonucleotide probe arrays. *Advances in Biochemical Engineering/Biotechnology*, **2002**, *77*, 21-42.

85. Schleinitz D, DiStefano JK, Kovacs P. Targeted SNP genotyping using the TaqMan® Assay. *Disease Gene Identification, Methods in Molecular Biology*, **2011**, *700*, 77-87.

86. Perkel J. SNP genotyping: six technologies that keyed a revolution. *Nature Methods*, **2008**, *5*, 447-453.

87. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. *Science*, **1988**, *241*, 1077-1080.

88. Xue X, Xu W, Wang F, Liu X. Multiplex single-nucleotide polymorphism typing by nanoparticle-coupled DNA-templated reactions. *Journal of the American Chemical Society*, **2009**, *131*, 11668-11669.

89. Qi X, Bakht S, Devos KM, Gale MD, Osbourn A. L-RCA (ligation-rolling circle amplification): a general method for genotyping of single nucleotide polymorphism (SNPs). *Nucleic Acids Research*, **2001**, *29*, e116.

90. Pickering J, Bamford A, Godbole V, Briggs J, Scozzafava G, Roe P, Wheeler C, Ghouze F, Cuss S. Integration of DNA ligation and rolling circle amplification for the homogeneous, end-point detection of single nucleotide polymorphisms. *Nucleic Acids Research*, **2002**, *30* (12), e60.

91. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering*, **2007**, *9*, 289-320.

92. Shen R, Fan J, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A. High-throughput SNP genotyping on universal bead arrays. *Mutation Research*, **2005**, *573*, 70-82.

93. Fan J, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Garcia EW, Lebruska LL, Laurent M, Shen R, Barker D. Illumina universal bead arrays. *Methods in Enzymology*, **2006**, *410*, 57-73

94. Ross P, Hall L, Smirnov I, Haff L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature Biotechnology*, **1998**, *16*, 1347-1351.

95. Haff LA, Sminov IP. Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Research*, **1997**, *7*, 378-388.

96. Sauer S, Gelfand DH, Boussicault F, Bauer K, Reichert F, Gut IG. Facile method for automated genotyping of single nucleotide polymorhpisms by mass spectrometry. *Nucleic Acids Research*, **2002**, *30*, e22.

97. http://www.sequenom.com/

98. Jurinke C, van den Boom D, Cantor CR, Köster H. Automated genotyping using the DNA MassArray technology. *Methods in Molecular Biology*, **2002**, *187*, 179-192

99. Fei Z, Ono T, Smith LM. MALDI-TOF mass spectrometric typing of single nucleotide polymorphisms with mass-tagged ddNTPs. *Nucleic Acids Research*, **1998**, *26*, 2827-2828.

100. Kim S, Ruparel HD, Gilliam TC, Ju J. Digital genotyping using molecular affinity and mass spectrometry. *Nature Review Genetics*, **2003**, *4*, 1001-1008.

101. Sotelo-Silveira J R, Calliari A, Kun A, Koenig E, Sotelo JR. RNA trafficking in axons, *Traffic*, **2006**, *7*, 508-515.

102. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Development*, **2006**, *20*, 515-524.

103. Lamond AI, Sproat BS. Antisense oligonucleotide made of 2'-*O*-alkyl-RNA: their properties and applications in RNA biochemistry. *FEBS letters*, **1993**, *325*, 123-127.

104. Martí AA, Jockusch S, Stevens N, Ju J, Turro NJ. Fluorescent hybridization probes for sensitive and selective DNA and RNA detection. *Accounts of Chemical Research*, **2007**, *40*, 402-409

105. Tyagi S, Kramer FR. Molecular beads: probes that fluoresce upon hybridization. *Nature Biotechnology*, **1996**, *14*, 303-308.

106. Tsourkas A, Bao G. Shedding light on health and disease using molecular beacons. *Briefings in*

Functional Genomics and Proteomics, **2003**, *4*, 372-384.

107. Tyagi S, Bratu BP, Kramer FR, Multicolor molecular beacons for allele discrimination, *Nature Biotechnology.,* **1998**, 16, 49-53.

108. Tan W, Fang X, Li J, Liu X. Molecular beacons: a novel DNA probe for nucleic acid and protein studies, *Chemistry--A European Journal,* **2000**, *6*, 1107-1111.

109. Kolpashchikol DM, Binary probes for nucleic acid analysis, *Chemical Reviews*, **2010**, *110***,** 4709-4723.

110. Turro NJ. Modern Molecular Photochemistry; University Science Books: Sausalito, CA, **1991**.

111. Lakowicz JR. *Modern Molecular Photochemistry*; University Science Books: Sausalito, CA, **1991**.

112. Yang CJ, Martinez K, Lin H, Tan W. Hybrid molecular probe for nucleic acid analysis in biological samples. *Journal of the American Chemistry Society*, **2006**, *128*, 9986-9987.

Part I Mass Spectrometric DNA Sequencing and SNP Genotyping with Cleavable Biotinylated Dideoxynucleotides

MALDI-TOF mass spectrometry offers an attractive option for DNA analysis due to its high accuracy, sensitivity and speed. In this section, as a significant improvement on non-cleavable biotinylated dideoxynucleotides, we described the design and synthesis of cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins) and their particular applications in solid phase capturable mass spectrometric sequencing and single base extension-based mass spectrometric detection of single nucleotide polymorphisms (SNPs) that are associated with mitochondrial disease. We also explore the use of a SNP genotyping microfluidic lab-on-a-chip device with the potential for high throughput, miniaturization and automation.

Chapter 2 Design and Synthesis of Cleavable Biotinylated Dideoxynucleotides for DNA Sequencing and Genotyping by MALDI-TOF Mass Spectrometry

2.1 Introduction

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) was developed by Tanaka and Kara in 1988,[1, 2] with the initial application in proteomics studies. It was not until the introduction of a 3-hydroxypicolinic acid (3-HPA) matrix by Becker's group[3] and the addition of ammonium ions to the matrix by Pieles *et al.*[4] that nucleic acid analysis by MALDI-TOF MS started to gain momentum. The full potential for DNA analysis was demonstrated in 1995[5] and for RNA in 1998,[6] from which time MALDI-TOF MS has been widely used for rapid and accurate analysis of nucleic acids. In brief, as shown in Fig. 2.1, for MALDI-TOF MS analysis of DNA, the analytes, single-stranded nucleic acid molecules, need to be generated and deposited with the matrix molecules (typically ultraviolet (UV) or infrared (IR) light-absorbing small organic molecules, such as 3-HPA) on a flat sample plate and co-crystallized for the analysis. The analyte/matrix molecules are then irradiated by a laser, inducing energy desorption and ionization, after which the charged ions are accelerated under a constant electric voltage and pass through the flight tube to a detector at the opposite end. Molecules are distinguished by their time of flight, which is proportional to their individual masses. Hence, the masses of the charged ions are

determined from their time of flight to the detector, represented as m/z, mass per charge ratio.[7]

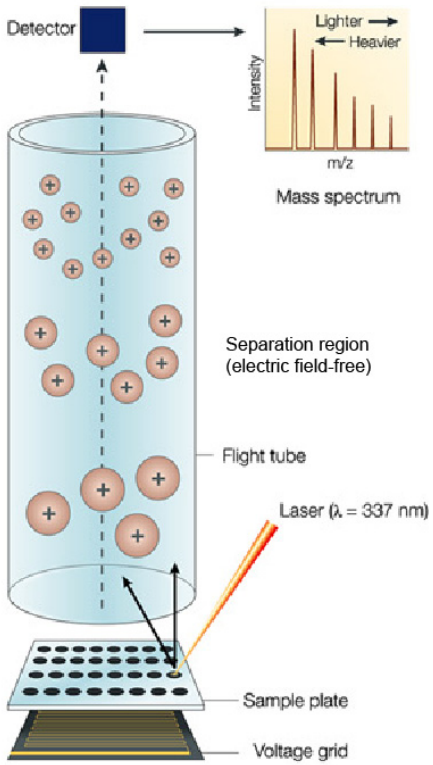


Fig. 2.1. Principle of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS).[7]

MALDI-TOF MS based DNA analysis has the advantages of accuracy, speed and potential for automation. Compared with gel mobility and hybridization-based fluorescent labeling assays, the absolute mass value represents an intrinsic property of a molecule, and therefore more informative and accurate, free from secondary structure effects that typically plague gel electrophoresis and the false positive signals resulting from mismatch hybridization. The mass data acquisition can be accomplished in less than 1 min, and the fragment "separation" process can be fully automated. Mass spectrometry

(MS) has been applied to a number of different areas of DNA analysis, such as sequencing, mutation and polymorphism analysis, and virus or bacteria detection,[8] among which MS-based Sanger sequencing and single base extension (SBE)-based single nucleotide polymorphism (SNP) genotyping have gained the most attention. Nevertheless, a major challenge of MALDI-TOF MS analysis is the stringent purity required for the DNA fragments introduced into the mass detector, which demands DNA fragments free of alkali and alkaline earth salts as well as other contaminants.[9]

One of the approaches for purifying DNA samples is to utilize the strong, specific and stable interaction of biotin and streptavidin coupled with solid-phase capture.[10] Fu *et al.* reported the use of a 5' biotinylated primer for sample purification by streptavidin coated magnetic beads in their MS sequencing. The Ju lab introduced solid phase capturable dideoxynucleotides, biotinylated ddNTPs, in both their sequencing and genotyping strategy to further eliminate falsely stopped DNA (i.e., preliminarily terminated) fragments and excess primers.[7, 9] This significantly improved the performance of MS analysis. Nonetheless, harsh conditions are required to cleave the biotin-streptavidin bond, such as treatment with formamide at a high temperature. This complicates downstream procedures, since the isolated products need to be ethanol precipitated prior to the desalting step for MS analysis. This is a tedious, time consuming process and is prone to sample loss. Though some groups have tried to develop mild approaches for breaking the bond between biotin and streptavidin,[11] the presence of the biotin moiety in purified DNA fragments introduces its own complications in higher

resolution analysis since biotin contains a sulfur atom which exists as four major stable isotopes. All of these problems could be eliminated by taking advantage of biotinylation analogs with cleavable linker arms.[12] As early as 1985, Shimkus *et al.* reported the use of a cleavable disulfide bond to link the biotin moiety to the nucleotide for the recovery of protein-DNA complexes.[13] Olejnik *et al.* attached biotin to peptides and oligonucleotides through a photocleavable linker for isolation and subsequent purification.[14] Bai *et al.* first introduced a nucleotide analog containing a photocleavable biotin that can be incorporated into the DNA strand by polymerase, which significantly facilitated the utilization of the cleavable linkers.[12] In their study, they attached the biotin moiety to the 5-position of 2'-deoxyribouridine 5'-triphosphate (dUTP) via a photocleavable 2-nitrobenzyl linker, and demonstrated it to be a good substrate for Thermo Sequenase in single and multiple primer extension reactions.

On the other hand, the decreasing resolving capacity of the mass spectrometer for larger DNA fragments requires appropriate mass differences to achieve high resolution and accuracy. This could be addressed by having different length linkers between the biotin and the nucleotide. For example, in their DNA sequencing analysis, Edward *et al.* replaced the biotin-11-ddUTP with biotin-16-ddUTP, to overcome the small mass difference between biotin-11-ddUTP and biotin-11-ddCTP (1 Da). The introduction of the longer linker arm shifted the mass difference to >80 Da, allowing better resolution in sequence determination.

This part of the thesis describes the design, synthesis and evaluation of a complete

set of further modified mass tagged, chemically cleavable dideoxynucleotides for application in DNA sequencing and SNP genotyping.

2.2 Experimental Rationale and Overview

The chemically cleavable biotinylated dideoxynucleotides are designed in such a way that the biotin moiety is covalently attached to the 5-position of the pyrimidines and the 7-position of the purines through chemically cleavable linkers of different lengths. As shown in Fig. 2.2, these chemically modified dideoxynucleotides are incorporated into the DNA strand by the polymerase which thereby generates DNA fragments with a cleavable biotin moiety at the 3' end. The biotinylated DNA fragments are captured by streptavidin interaction, while other contaminants from the polymerase reaction are washed away. Then the fragments are released from the streptavidin-coated surface by cleaving the linker under mild chemical conditions, leaving the biotin attached to the surface.

Here, a chemically cleavable biotinylated dideoxynucleotide set, ddNTP-N_3-biotins (ddATP-N_3-biotin, ddGTP-N_3-biotin ddCTP-N_3-biotin, and ddUTP-N_3-biotin) was generated for the DNA polymerase extension reaction and DNA sequencing as well as genotyping by MS applications. The nucleotide analogs were designed to have a biotin moiety attached to the 5 position of pyrimidines (C and U) or the 7 position of purines (A and G) via an azido based linker, and the nucleotide precursors with two different length carbon linker arms were used to increase the mass differences among these nucleotide

analogs. Previously, we reported that chemically cleavable fluorescent dideoxynucleotides with azido based linkers could be used successfully for DNA sequencing by synthesis, and the fluorophores were shown to be completely removable under very mild cleavage conditions, using an aqueous tris(2-carboxyethyl) phosphine (TCEP) solution.[15] Hence, azido linkers were chosen to attach the biotin to the nucleotides in the current design, since the linker can be cleaved by TCEP very efficiently and TCEP is compatible with the downstream desalting step.

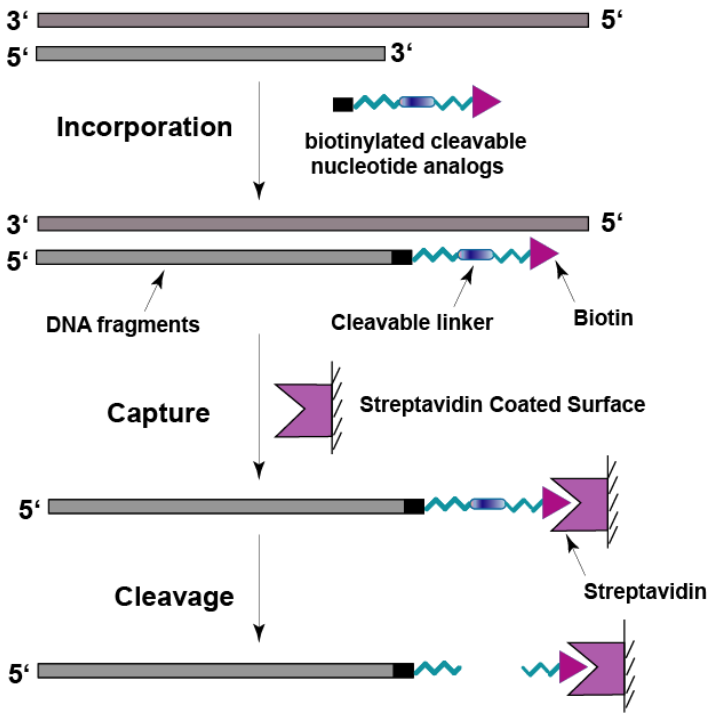


Fig. 2.2. Scheme of single base extension, solid phase capture and cleavage using chemically cleavable dideoxynucleotides

It is demonstrated here that ddNTP-N_3-biotins were able to be incorporated into DNA strands during the polymerase extension reaction, and that DNA strands with

biotinylated dideoxynucleotides at their 3'-end can be efficiently captured by streptavidin coated magnetic beads and then released from the beads with TCEP. Such nucleotide analogs that carry a biotin and a chemically cleavable linker will allow the isolation and purification of DNA fragments under mild conditions for MS-based sequencing and multiplex SNP detection, as well as facilitate DNA-protein complex purification in a non-denaturing fashion.

2.3 Results and Discussion

2.3.1 Design and synthesis of cleavable biotinylated dideoxynucleotides

Synthetic work for the full set of cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins) was accomplished by Dr. Shiv Kumar in our organic chemistry laboratory. Introduction of a chemically cleavable biotin moiety into the nucleotides enabled highly effective isolation and purification of the DNA products, which could then be applied to MALDI-TOF MS for DNA sequencing and genotyping analysis. It has been shown previously that when the 5 position of the pyrimidines (C/U) and the 7 position of the purines (A/G) were modified with bulky fluorescent dyes through azido-based linkers, DNA polymerase could still incorporate the modified nucleotides into the growing DNA strand, and the azido-based linker could be efficiently cleaved by the Staudinger reaction using an aqueous TCEP solution.[15] Therefore, azido-based linkers were chosen for this study. Here, biotin-azido linkers were pre-synthesized. To enlarge the mass difference between the modified nucleotides, propargylamino-ddGTP and

propargylamino-ddUTP, available commercially, were modified with an extra linker module to produce a longer linker arm before connection to the shorter biotin-azido linkers previously prepared for all four nucleotides. This generated a set of cleavable biotinylated dideoxynucleotides, ddNTP-N_3-biotins, shown in Fig. 2.3, which have a biotin moiety attached to the 5 position of the pyrimidines (C/U) and the 7 position of the purines (A/G) via chemically cleavable azido-based linkers which are longer in the U and G than in the A and C analogues.

With increasing DNA fragment size, mass spectrum peak widths increase, resulting in a decreasing resolution of the mass peaks with larger DNA fragments. In addition, very high accuracy is required to demonstrate the existence of single base polymorphisms (SNPs). Hence, for the unambiguous determination of sequence in the higher mass range as well as heterozygote detection, it is important to design the modified nucleotides with clearly distinguishable mass tags.[9, 16]

In this study, dideoxynucleotide precursors with two different carbon linker arms were chosen to increase the mass difference between different nucleotides before and after cleavage. Thus, the smallest mass difference between two modified dideoxynucleotides is 23 Da (A and C), whereas it is only 9 Da for standard terminators (A and T), and 16 Da for previously used biotinylated nucleotides (A and G). Also, the mass difference between A and G has been shifted from 16 Da to 129 Da, and the mass difference between C and G shifted from 39 Da to 152 Da. These mass-tagged linker halves still remain after cleavage since they are positioned before the azido group in the

full linker. The adjusted masses of these tagged nucleotides clearly provide better resolution and enhanced accuracy within the separable range.

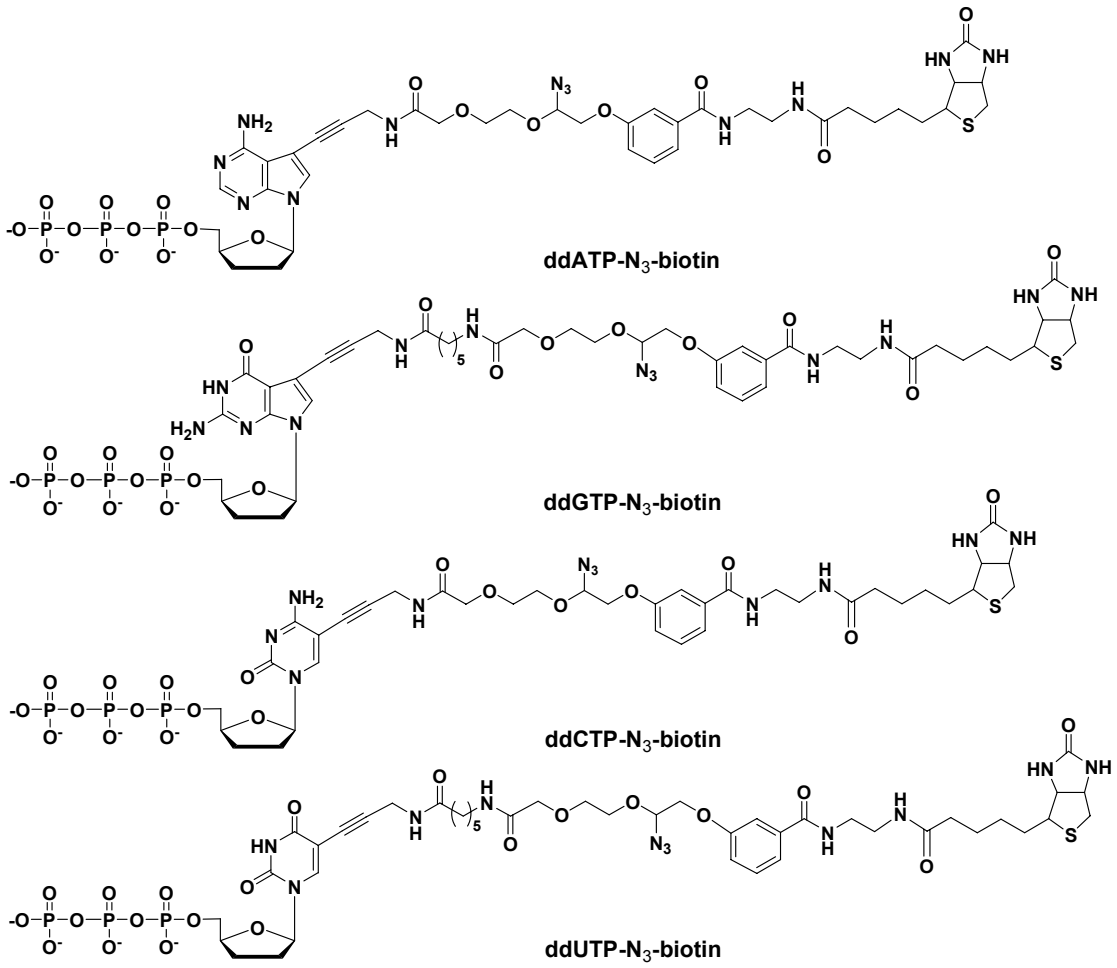


Fig. 2.3. Structures of cleavable dideoxynucleotides (ddNTP-N_3-biotins). Note that the length of the portion of the linker between the base and the N_3 group, the segment which remains attached to the extended DNA after TCEP cleavage, varies between the two purine (A and G) and the two pyrimidine bases (C and U) bases, enabling clear discrimination of their sizes by MALDI-TOF mass spectrometry.

2.3.2 Polymerase extension using ddNTP-N_3-biotins, solid phase capture and cleavage

In developing the MS genomic analysis assay, it was essential that the biotinylated nucleotides could be efficiently incorporated into the DNA strand during the polymerase reaction, the DNA strand bearing biotinylated nucleotides could be effectively captured on the streptavidin coated solid phase, and then efficiently released after cleavage while leaving the biotin behind on the solid phase. To verify this, single base extension reactions with four corresponding self-priming DNA templates were carried out in solution. As shown in Fig. 2.4 A, D, G, and J, 100% incorporation was confirmed with MALDI-TOF mass spectrometry by observing the total disappearance of the primer peak (7966 m/z) and the emergence of the extension product for each dideoxynucleotide (8889 for ddATP-N_3-biotin, 9018 for ddGTP-N_3-biotin, 8866 for ddCTP-N_3-biotin, and 8980 for ddUTP-N_3-biotin, m/z). Incubation of the extension products in TCEP solution led to the cleavage of the N_3-based linker tethering the biotin to the dideoxynucleotides. As shown in Fig. 2.4 B, E, H, and K, the mass peaks for the extension products have completely disappeared, whereas single peaks corresponding to the cleavage products appear at 8416, 8545, 8393, and 8507 (m/z), respectively, which indicates 100% cleavage efficiency. To evaluate these modified nucleotides for the solid phase purification procedure, after performing the same single base extension reactions, the extension products were captured on streptavidin coated magnetic beads, released by TCEP, and subsequently analyzed by MALDI-TOF MS. The results are shown in Fig. 2.4 C, F, I and

L, where single peaks corresponding to released extension products again appear at 8416, 8545, 8393, and 8507 (m/z), respectively, the expected masses of the cleavage products. The overall process of single base extension, solid phase capture and cleavage for each nucleotide is shown in the upper part of Fig. 2.4. Given ddA-N_3-biotin as an example, the detailed process is illustrated in Fig. 2.5. After single base extension, DNA extension product (1) (Fig. 2.5 A) was first captured by streptavidin-coated magnetic beads. After cleavage by TCEP, the DNA extension products captured on streptavidin-coated magnetic beads generated released product (3) (Fig. 2.5 C), while the biotin moiety stayed on the surface of the beads. The portion that remained on the nucleotides served as an enhanced mass tag to increase the mass differences between each of the modified nucleotides. The detailed mechanism of TCEP cleavage is shown in Fig. 2.6. In the Staudinger reaction, TCEP reduces the azido group into an amino group, which generates an active intermediate that is hydrolyzed efficiently in aqueous solution. This leads to the breakage of the linker.

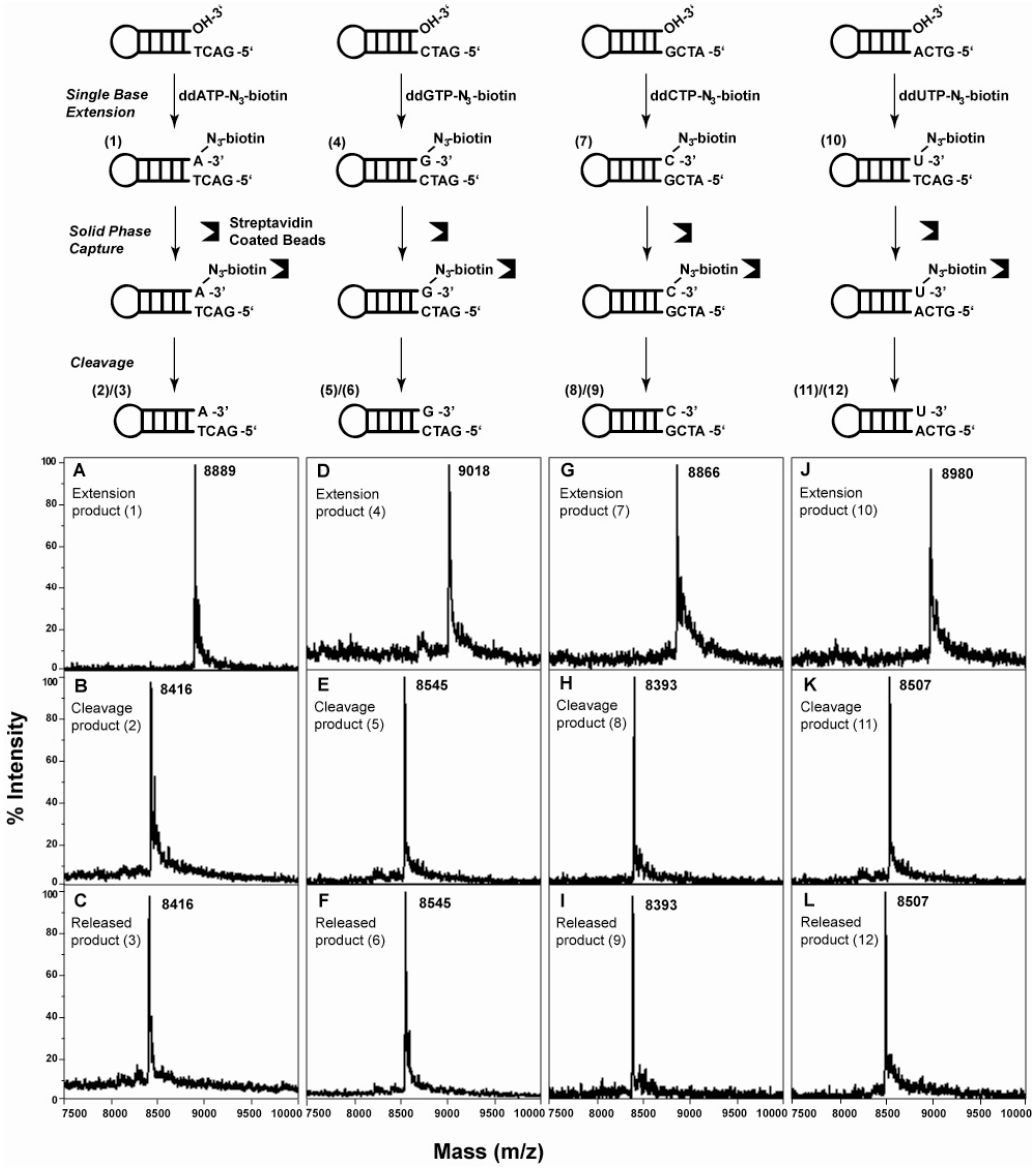


Fig. 2.4. MALDI-TOF mass spectra of the DNA extension products, their subsequent cleavage products in solution, and released DNA products from streptavidin coated magnetic beads. (A) primers extended with ddATP-N_3-biotin (1) (8889 m/z); (B) their cleavage products (2) (8416 m/z); (C) released products from solid phase (3) (8416 m/z); (D) primers extended with ddGTP-N_3-biotin (4) (9018 m/z); (E) their cleavage products (5) (8545 m/z); (F) released products from solid phase (6) (8545 m/z) (G) primers extended with ddCTP-N_3-biotin (7) (8866 m/z); (H) their cleavage products (8) (8393 m/z); (I) released products from solid phase (9) (8866 m/z); (J) primers extended with ddUTP-N_3-biotin (10) (8980 m/z); (K) their cleavage products (11) (8507 m/z); (L) released products from solid phase (12) (8507 m/z).

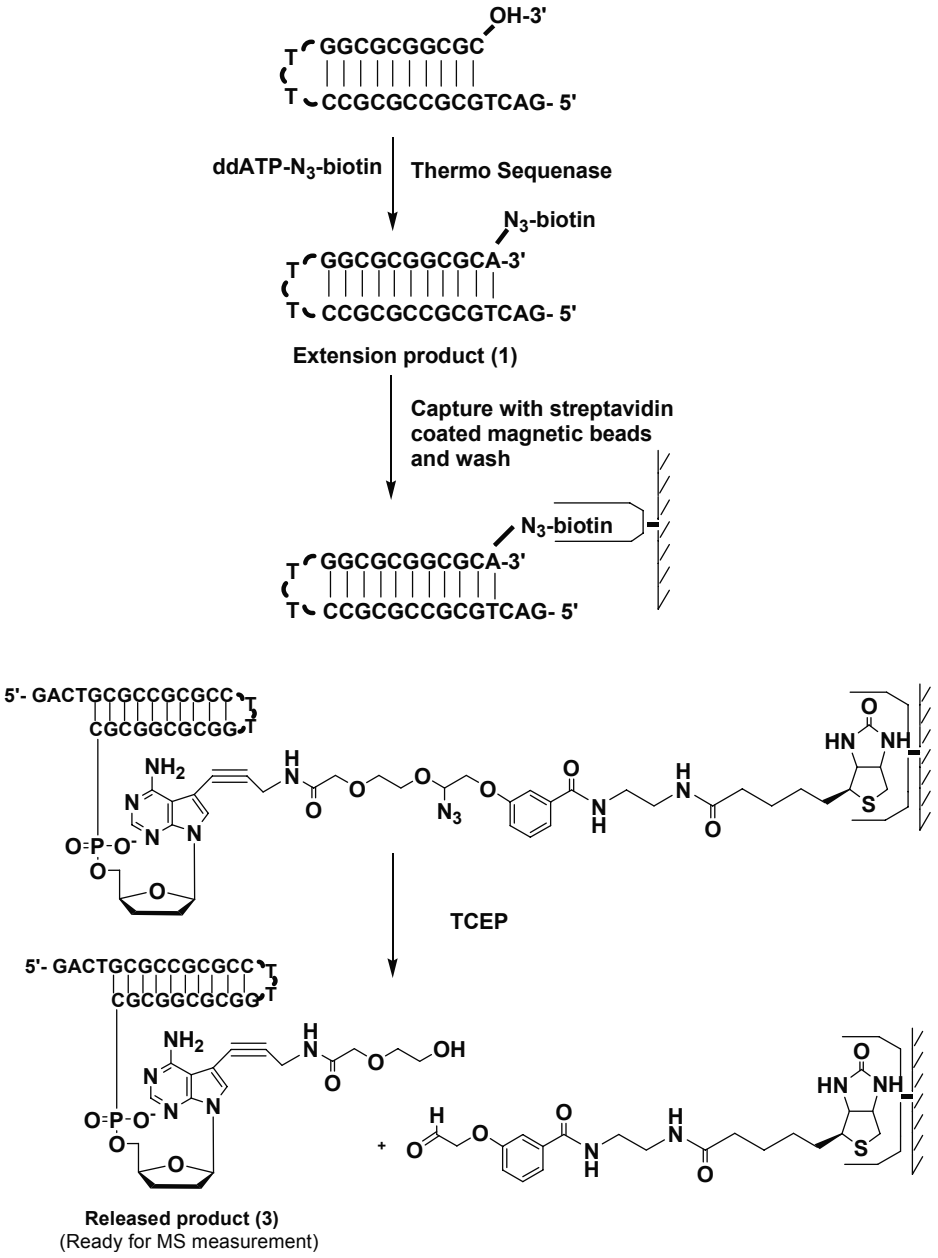


Fig. 2.5. Polymerase extension reaction using ddATP-N_3-biotin as a substrate and TCEP cleavage of DNA fragments containing ddA-N_3-biotin on streptavidin coated beads. The DNA polymerase Thermo Sequenase incorporates ddATP-N_3-biotin to generate the extension product (1). Cleavage by TCEP of the DNA extension products captured on streptavidin-coated beads produces released products (3), while the biotin moiety remains on the solid surface of the beads.

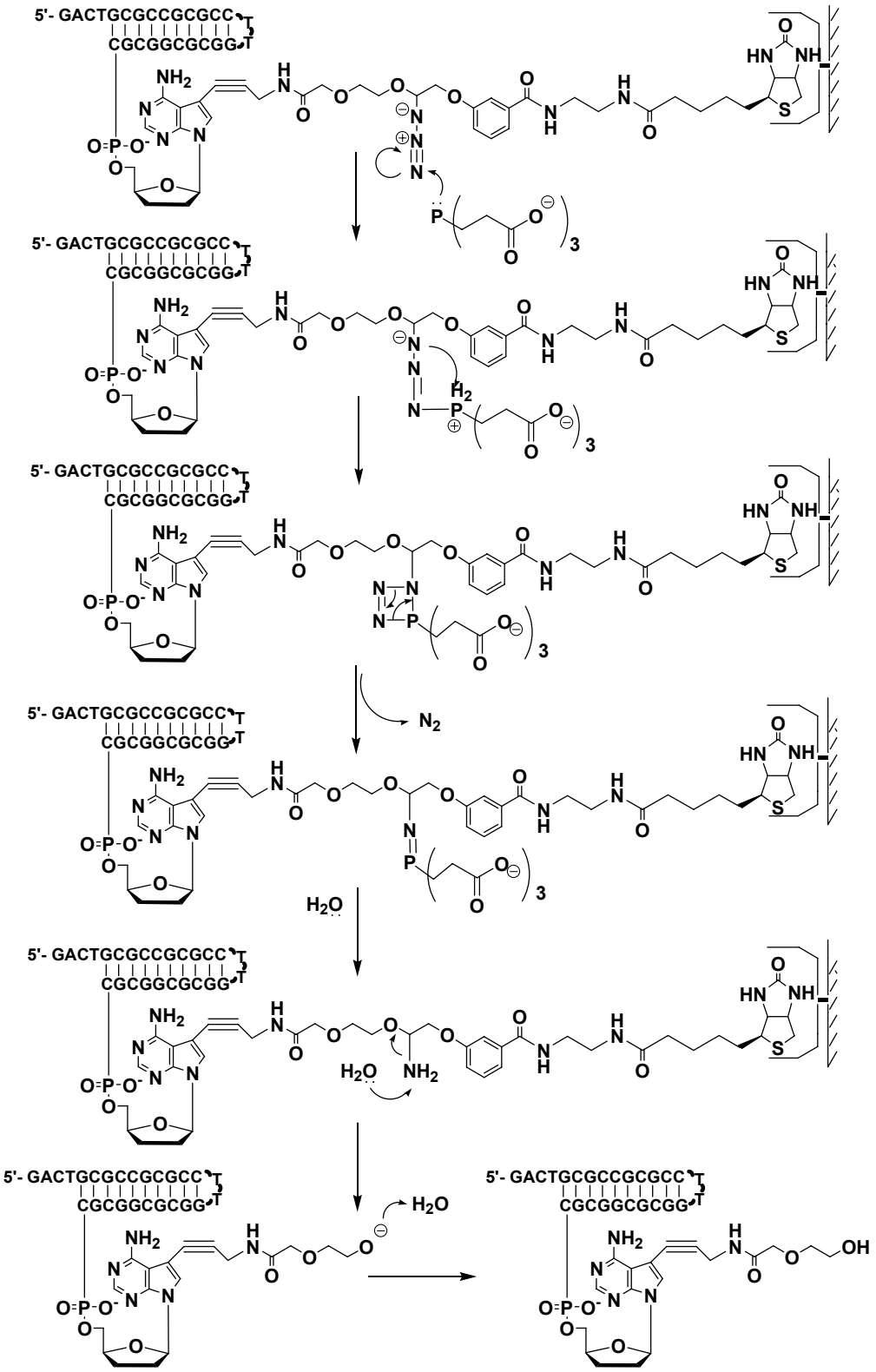


Fig. 2.6. Staudinger reaction with TCEP to cleave the azido-based linker and release the DNA extension products from the streptavidin coated surface.

2.3.3 Comparison with non-cleavable biotinylated dideoxynucleotides

The introduction of a chemically cleavable linker to the dideoxynucleotides was designed to facilitate the downstream process after solid phase capture. In order to verify the advantages of cleavable biotinylated dideoxynucleotides over noncleavable ones, taking ddATP-N_3-biotin and biotin-11-ddATP as an example, comparison studies were performed. Very small amounts of starting material, 0.5 pmol, 1.0 pmol and 2.5 pmol of self-priming template 26dA, were chosen to carry out the reaction to better compare the resulting sensitivity using each nucleotide. As shown in Fig. 2.7, following the same single base reaction and solid phase capture procedures, the ddATP-N_3-biotin terminated DNA strand was treated with TCEP, the product of which was taken directly to the desalting step since TCEP is compatible with the ZipTip procedure. In contrast, the biotin-11-ddATP extended DNA strand had to go through high temperature (95^{o}C) treatment in the presence of formamide, which required ethanol precipitation before the desalting step. Due to the small amount of starting material, an overnight ethanol precipitation was required to recover the maximum amount of sample. Hence, the use of cleavable nucleotides simplified the overall procedure and saved time. The mass spectrometry results further demonstrated that using cleavable biotinylated nucleotides (ddATP-N_3-biotin) leads to better sensitivity. As shown in Fig. 2.7 C, regardless of the amount of starting material, the signal generated from ddATP-N_3-biotin was more than 3 times higher than that from biotin-11-ddATP. The significant improvement of the detection signal was mainly due to higher cleavage efficiency and less sample loss.

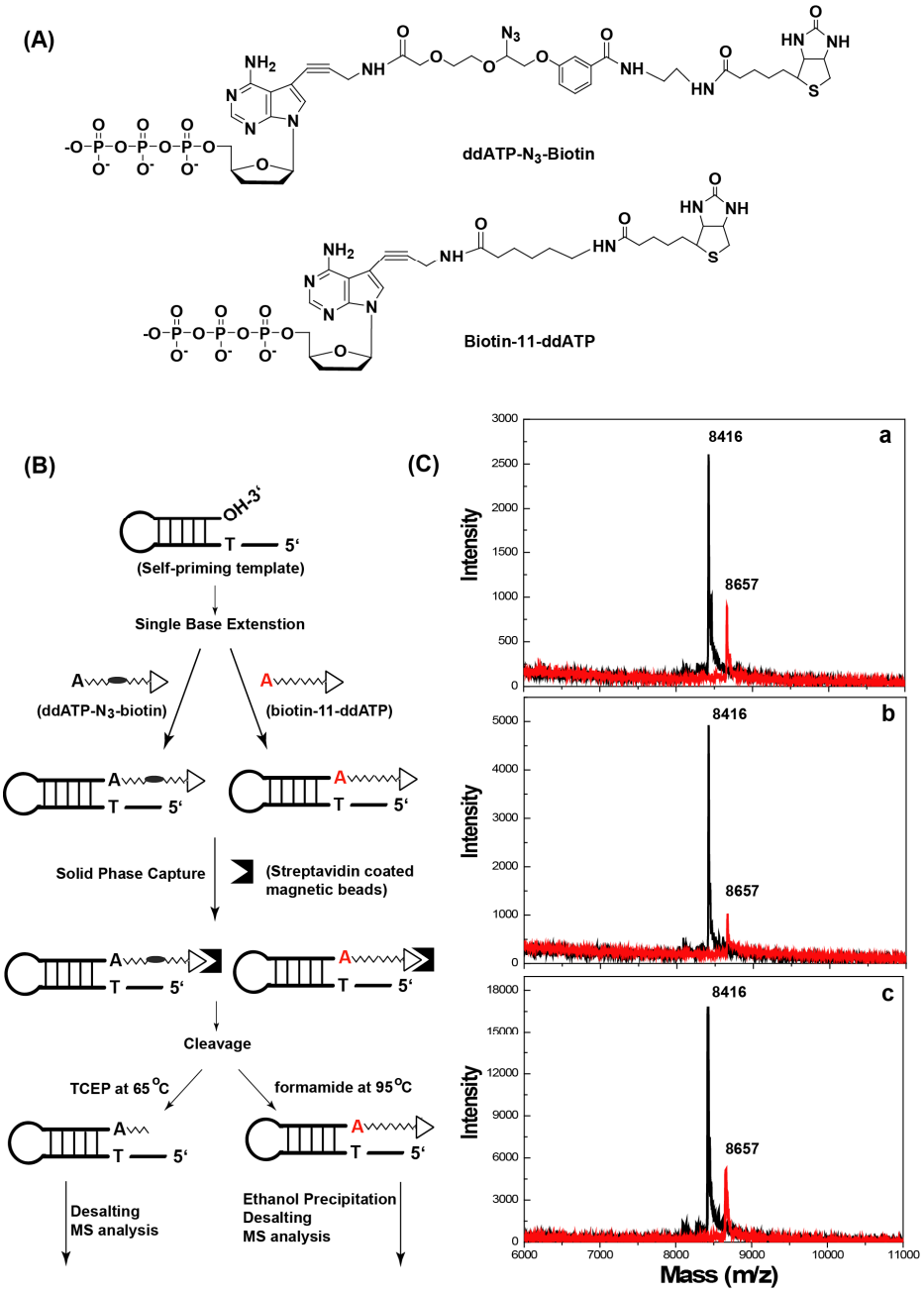


Fig. 2.7. Comparison of cleavable biotinylated dideoxynucleotide (ddATP-N_3-biotin) and non-cleavable biotinylated dideoxynucleotide (biotin-11-ddATP). (A) Chemical structures of ddATP-N_3-biotin and biotin-11-ddATP; (B) the purification process for each nucleotide. Note: after cleavage, the biotin-11-ddATP incorporated DNA template requires an extra ethanol precipitation step; (C) mass spectrum of final product for starting template: a, 0.5 pmol, b, 1.0 pmol, c, 2.5 pmol.

Furthermore, proper adjustment of the pH of the TCEP with ammonium hydroxide also avoids the production and retention of alkaline salts typically remaining after ethanol precipitation, which simplifies the overall desalting procedure. Additionally, though not demonstrated here, the released products from the solid phase do not contain the biotin moiety. This excludes the interference from the four isotopes of sulfur in the biotin molecule, and should result in a higher resolution and accuracy of the mass spectrum.

2.4 Materials and Methods

General information. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. ^{1}H-NMR spectra were recorded on a Bruker DPX-400 (400 MHz) spectrometer. Electrospray ionization (ESI) mass spectra were recorded on a Bruker Daltonics Esquire 6000 mass spectrometer. Mass measurement of DNA was performed on a Voyager DE™ MALDI-TOF mass spectrometer (Applied Biosystems by Life Technologies, San Diego, CA). HPLC was performed on a Waters system (Milford, MA) consisting of a Rheodyne 7725i injector, 600 controller and a 996 photodiode array detector. Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). Thermo Sequenase was from GE Healthcare (Piscataway, NJ). Streptavidin coated magnetic beads (Dynabeads® MyOne™ Streptavidin C1) were obtained from Life Technologies (San Diego, CA).

2.4.1 Synthesis of ddNTP-N3-biotins

The propargylamino-dideoxynucleotides (1-4) were either purchased from Perkin

Elmer Life and Analytical Sciences or prepared following the procedure described by Hobbs and Cocuzza.[17] The longer linker arm dideoxy nucleotides (5-6) were prepared according to Duthie *et al.*[18] and purified on HPLC (Fig. 2.8). Azido linker, (2-{2-[3-(2-amino-ethylcarbamoyl)-phenoxy]-1-azido-ethoxy}-ethoxyl)-acetic acid (7) was prepared according to Milton *et al.* [19]

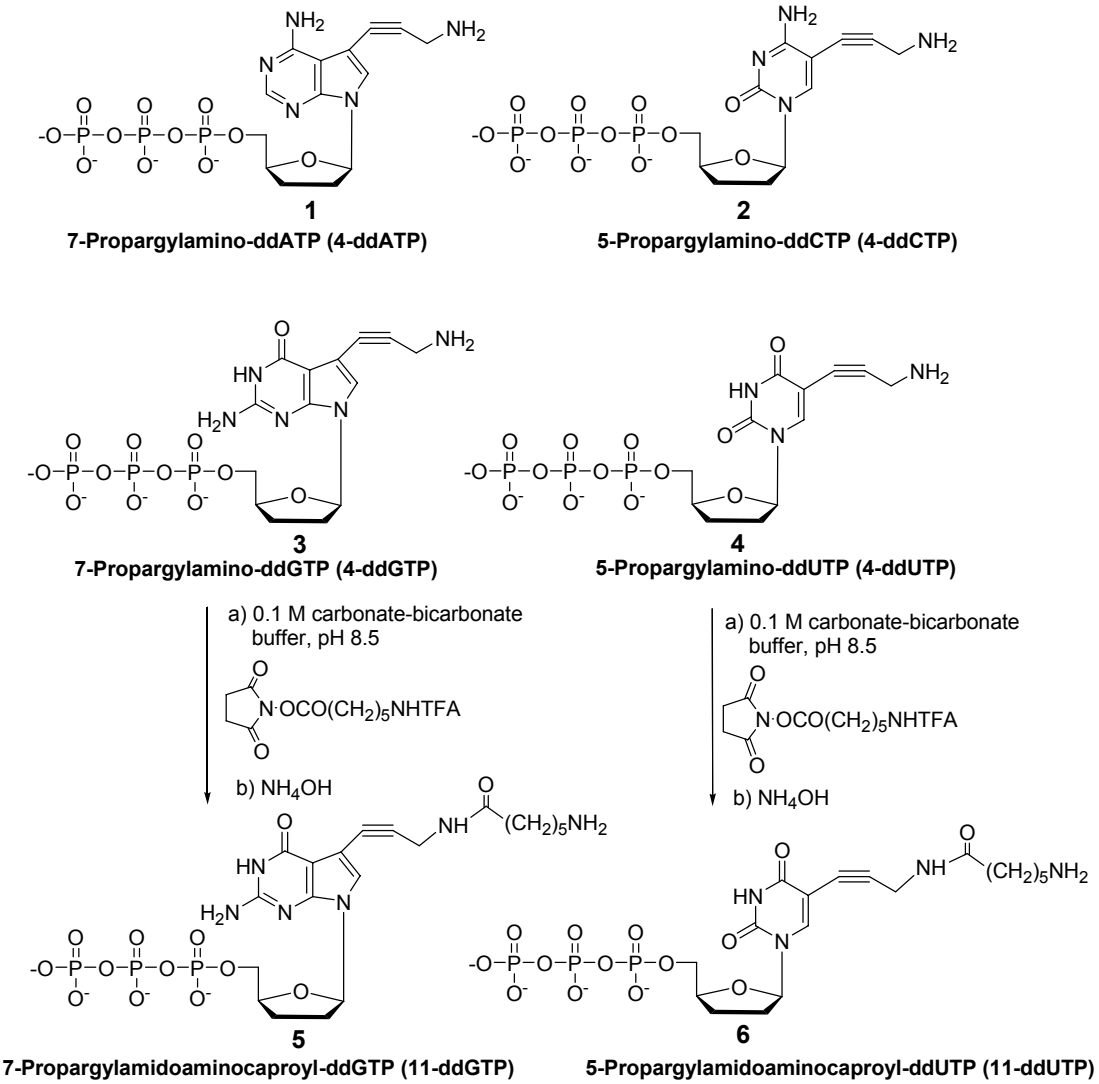


Fig. 2.8. Structures of 5- or 7-propargylamino-ddNTPs.

Synthesis of biotin-N_3-Linker Acid (8). Azido linker (7, 74 mg, 0.2 mmol) was

dissolved in anhydrous DMF (5 ml) and 1.5 ml of 1 M $NaHCO_3$ aqueous solution. A solution of Biotin-NHS ester (75 mg, 0.22 mmol) in 4 ml of anhydrous DMF was added slowly to the stirred reaction mixture and was stirred overnight at room temperature. The reaction mixture was concentrated *in vacuo* and purified on a silica gel chromatography column using 25% methanol in methylene chloride to 40% methanol in methylene chloride. The appropriate fractions were combined and concentrated to give a white solid (72 mg, 60%). ^{1}H NMR (400 MHz, D_2O): δ 7.48-7.39 (m, 3H), 7.16 (d, 1H), 5.05 (t, 1H), 4.48 (t, 1H), 4.29-4.20 (m, 3H), 4.00-3.90 (m, 4H), 3.78 (m, 2H), 3.53-3.45 (2d, 4H), 2.91-2.87 (dd, 1H), 2.69 (d, 1H), 2.36 (t, 2H), 1.66-1.37 (m, 6H); FAB`-MS *m/z*: calcd for $C_{25}H_{36}N_7O_8S$ ($M+H^+$), 594.65; found 594.75

General method for the synthesis of ddNTP-N3-biotins (10-13). Biotin-N_3-Linker acid (8, 4mg, 6.75 μmol) was co-evaporated with anhydrous DMF under reduced pressure and re-dissolved in anhydrous DMF (0.8 ml). A solution of O-(N-Succinimidyl)-1,1,3,3-tetramethyluronium tetrafluoroborate (TSTU, 20 μmol) in anhydrous DMF (0.4 ml) was added to the stirred solution under an argon atmosphere and the reaction mixture was stirred at room temperature for 15 min. The appropriate amino-ddNTP (1, 2, 5 or 6, 10 μmol) in 0.1 M $NaHCO_3$-Na_2CO_3 buffer (pH 8.7, 300 μl) was added to the activated ester and the reaction mixture stirred at room temperature overnight (Fig. 2.9). The mixture was purified by reverse-phase HPLC on a 150 X 4.6 C18 column (Supelco, PA) using mobile phase: A, 8.6 mM triethylamine/100 mM

hexafluoroisopropyl alcohol in water (pH 8.1); B, methanol. Elution was performed with 100% A isocratic for 10 min followed by a linear gradient of 0-50% B for 20 min and then 50% isocratic for another 20 min. The isolated pure adducts were characterized by the ESI-MS analysis and single base extension followed by MALDI-TOF MS analysis.

ddATP-N_3-biotin (10): HPLC retention time 30.7 min; TOF MS ES+ m/z: anal. Calculated for $C_{39}H_{52}N_{12}O_{18}P_3S$ ($M\text{-}H^-$) 1101.89; found, 1101.2

ddCTP-N_3-biotin (11): HPLC retention time 30.1 min; TOF MS ES+ m/z: anal. Calculated for $C_{37}H_{51}N_{11}O_{19}P_3S$ ($M\text{-}H^-$) 1078.85; found, 1078.2

ddGTP-N_3-biotin (12): HPLC retention time 30.8 min; TOF MS ES+ m/z: anal. Calculated for $C_{45}H_{63}N_{13}O_{20}P_3S$ ($M\text{-}H^-$) 1231.04; found, 1230.3

ddUTP-N_3-biotin (13): HPLC retention time 30.68 min; TOF MS ES+ m/z: anal. Calculated for $C_{43}H_{61}N_{11}O_{21}P_3S$ ($M\text{-}H^-$) 1192.99; found, 1192.2

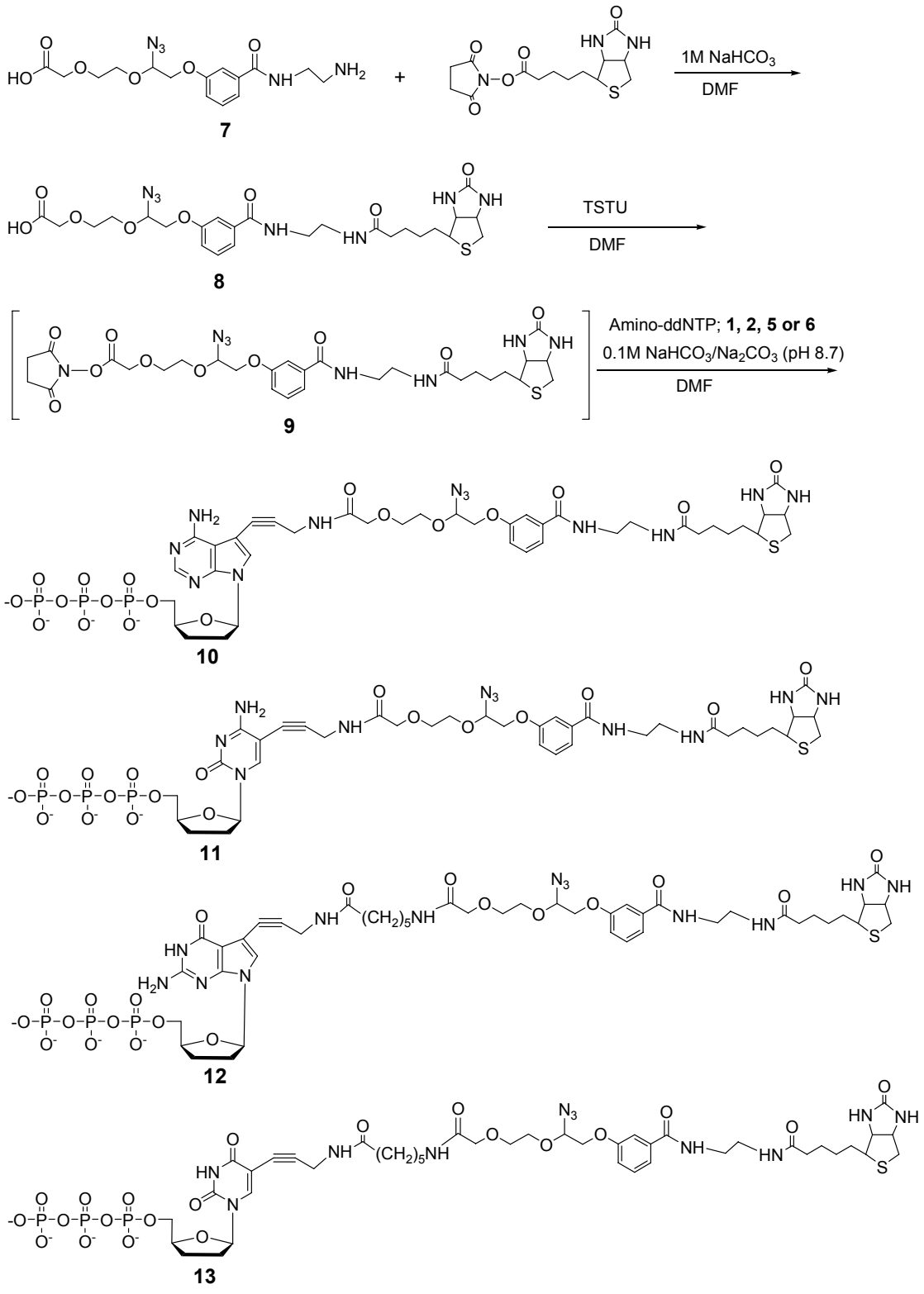


Fig. 2.9. Synthesis and structures of biotin-N_3-linker attached ddNTPs

2.4.2 Polymerase extension using ddNTP-N_3-biotins, solid phase capture and cleavage

The four cleavable biotinylated dideoxynucleotides, ddNTP-N_3-biotins (ddATP-N_3-biotin, ddGTP-N_3-biotin, ddCTP-N_3-biotin, and ddUTP-N_3-biotin) were first characterized by performing four separate single base extension reactions, each with a different self-priming DNA template allowing the four dideoxynucleotide analogs to be incorporated. The following four self-priming DNA templates (26-mer hairpin DNA with a 4-base 5'-overhang) were used for the extension: 5'-GACTGCGCCGCGCCTTGGCG-CGGCGC-3' for ddATP-N_3-biotin, 5'-GATCGCGCCGCGCCTTGGCGCGGCGC-3' for ddGTP-N_3-biotin, 5'-ATCGGCGCCGCGCCTTGGCGCGGCGC-3' for ddCTP-N_3-biotin, and 5'-GTCAGCGCCGCGCCTTGGCGCGGCGC-3' for ddUTP-N_3-biotin. Each of the extension reactions consisted of 40 pmol self-priming DNA template, 60 pmol of the corresponding ddNTP-N_3-biotin, 1X Thermo Sequenase reaction buffer and 2 U of Thermo Sequenase in a total volume of 20 μl. The reaction mixture was incubated at 65^{o}C for 15 min. For the incorporation test, the extension products were desalted by using ZipTips and analyzed by MALDI-TOF mass spectrometry. The matrix solution was made by dissolving 35 mg 3-hydroxypicolinic acid with 6 mg ammonium citrate in 800 μl 50% acetonitrile. For the cleavage, each extension product was mixed with 20 μl tris(2-carboxyethyl) phosphine (TCEP) solution (100 mM, pH 9.0 adjusted with ammonium hydroxide) and incubated at 65^{o}C to yield DNA cleavage products which were characterized by MALDI-TOF MS. To evaluate solid phase capture efficiency and

DNA recovery from the solid phase, the same single base extension reactions were performed, and 20 μl of each of the single base extension products was incubated with 20 μl streptavidin coated magnetic beads which had been prewashed with 1X binding and washing (B/W) buffer (5 mM Tris-HCl, 0.5 mM EDTA, 1 M NaCl, pH 7.5) three times, resuspended in 20 μl 2X B/W buffer, and allowed to incubate for 1 h at room temperature. The streptavidin coated magnetic beads bearing extended self-priming DNA templates were washed three times with 1X B/W buffer and three times with deionized water, then suspended in 20 μl TCEP solution and incubated at 65^{o}C for 25 min. This process removed the biotin moiety from the dideoxynucleotides and hence released extended self-priming templates from the magnetic beads. The supernatant was collected and desalted through ZipTip for MALDI-TOF MS characterization.

2.4.3 Comparison of ddATP-N_3-biotin and biotin-11-ddATP

A self-priming DNA template 5'-GACTGCGCCGCGCCTTGGCGCGGCGC-3' which can be used to incorporate modified ddATP was chosen for the single base extension reaction. 0.5 pmol, 1.0 pmol, and 2.5 pmol of template was used for ddATP-N_3-biotin and biotin-11-ddATP (Perkin Elmer) respectively. For the parallel comparison study, besides DNA template, each extension reaction also contained ddATP-N_3-biotin or biotin-11-ddATP, the amount of which was 1.5 times that of the DNA template, 2 U of Thermo Sequenase and 1X Thermo Sequenase reaction buffer in a total volume of 20 μl. The reaction mixture was incubated at 65^{o}C for 15 min. For

ddATP-N_3-biotin extended products, the same procedures as described in the last section were followed with mass spectrometric analysis directly after cleavage and ZipTip desalting. However, for biotin-11-ddATP extended products, a different cleavage treatment as well as downstream process was carried out. Briefly, after the solid phase capture and bead washing, the streptavidin coated magnetic beads bearing extended DNA templates were suspended in 10 μl of 95% formamide at 94^{o}C for 10 min. The released products were precipitated overnight with 100% ethanol at 4^{o}C and centrifuged at 14,000 rpm for 40 min, followed by a 70% ethanol wash with centrifugation at 14,000 rpm for 20 min. Then the products were analyzed by mass spectrometry after ZipTip desalting.

2.5 Conclusion

A set of mass tagged, chemically cleavable biotinylated dideoxynucleotide analogues, ddNTP-N_3-biotins, were synthesized and evaluated for their application for rapid and efficient recovery of DNA fragments captured on a solid surface under mild conditions. These dideoxynucleotide analogues were shown to be good substrates for the DNA polymerase, allowing accurate and fast incorporation by Thermo Sequenase. Biotin terminated DNA extension products generated from ddNTP-N_3-biotin incorporation could be captured on a streptavidin-coated solid surface, and subsequently released by using TCEP rather than harsh reagents. The cleavable linkers in the nucleotide analogues not only facilitated the DNA fragment recovery after solid phase capture, but also acted as mass tags to enlarge the mass differences among the nucleotides for better resolution.

It was also demonstrated that these chemically cleavable biotinylated dideoxynucleotides had substantial advantages over the conventional non-cleavable biotinylated dideoxynucleotides in terms of downstream workflow, cleavage efficiency and sample recovery. This set of biotinylated dideoxynucleotide analogues will be a valuable tool for DNA sequencing and genotyping by MALDI-TOF MS, as well as other genetic analysis. Examples of their utility for sequencing and genotyping are presented in the next two Chapters, with our work to develop a lab-on-chip microfluidic device for increasing the throughput of these protocols in Chapter 5.

References

1. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. Protein and polymer analysis up to m/z 100,000 by laser desorption time-of-flight mass spectrometry. *Rapid Communication Mass Spectrometry*, **1988**, *2*, 151-153.

2. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10, 000 daltons. *Analytical Chemistry*, **1988**, *60*, 2299-2301.

3. Wu KJ, Steding A, Becker CH. Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. *Rapid Communication Mass Spectrometry*, **1993**, *7 (2)*, 142-146.

4. Pieles U, Zürcher W, Schär M, Moser HE. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Research*, **1993**, *21*, 3191-3196.

5. Tang K, Fu D, Kötter S, Cotter RJ, Cantor CR, Köster H. Matrix-assisted laser desporption/ionization mass spectrometry of immobilized duplex DNA probes. *Nucleic Acids Res*, **1995**, *23*, 3126-3131.

6. Tolson DA, Nicholson NH. Sequencing RNA by a combination of exonuclease digestion and uridine specific chemical cleavage using MALDI-TOF. *Nucleic Acids, Research*, **1998**, *26*, 446-451.

7. Kim S, Ruparel HD, Gilliam C, Ju J. Digital genotyping using molecular affinity and mass spectrometry. *Nature Reviews Genetics*, **2003**, *4*, 1001-1008.

8. Fu D-J, Tang K, Braun A, Reuter D, Darnhofer-Demar B, Little DP, O'Donnell MJ, Cantor CR, Köster H. Sequencing exons 5 to 8 of the *p53* gene by MALDI-TOF mass spectrometry. *Nature Biotechnology*, **1998**, *16*, 381-384.

9. Edwards JR, Itagaki Y, Ju J. DNA sequencing using biotinylated dideoxynucleotides and mass spectrometry. *Nucleic Acids Research*, **2001**, *29 (21)*, e104.

10. Laitinen OH, Hytönen VP, Nordlund HR, Kulomaa MS. Genetically engineered avidins and streptavidins. *Cellular and Molecular Life Science,* **2006,** *63*, 2992-3017.

11. Holmberg A, Blomstergren A, Nord O, Lukacs M, Lundeberg J, Uhlen M. The biotin-streptavidin interaction can be reversible broken using water at elevated temperatures. *Electrophoresis*, **2005**, *26*, 501-510.

12. Bai X, Kim S, Li Z, Turro NJ, Ju J. Design and synthesis of a photocleavable biotinylated nucleotide for DNA analysis by mass spectrometry. *Nucleic Acids Research*, **2004**, *32*, 535-541.

13. Shimkus M, Levy J, Herman T. A chemically cleavable biotinylated nucleotide: usefulness in the recovery of protein-DNA complexes from avidin affinity columns. *Proceedings of the National Academy of Science of the United States of America,* **1985**, *82*, 2593-2597.

14. Olejnik J, Sonar S, Krzymanska-Olejnik E, Rothschild K. Photocleavable biotin derivatives: a versatile approach for the isolation of biomolecules. *Proceedings of the National Academy of Science of the United States of America,* **1995**, *92*, 7590-7594.

15. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Marma MS, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescen dideoxynucleotides. *Proceedings of the National Academy of Science of the United States of America,* **2008,** *105*, 9145-9150.

16. Fei Z, Ono T, Smith LM. MALDI-TOF mass spectrometric typing of single nucleotide polymorphisms with mass-tagged ddNTPs. *Nucleic Acids Research*, **1998**, *26*, 2827-2828.

17. Hobbs FW, Cocuzza AJ. Alkynylamino-nucleotides. **1991**. US Patent 5047519

18. Duthie RS, Kalve IM, Samols SB, Hamilton S, Livshin I, Khot M, Nampalli S, Kumar S Fuller CW. Novel cyanine dye-labeled dideoxynucleoside triphosphates for DNA sequencing.

Biocongugate Chemistry, **2002**, *13*, 699-706

19. Milton J, Ruediger S, Liu X. Labeled Nucleotides.**2006**, US Patent Application 2006/0160081 A1

Chapter 3 DNA Sequencing by MALDI-TOF Mass Spectrometry using Cleavable Biotinylated Dideoxynucleotides

3. 1 Introduction

With the completion of the first human genome project as example and reference, resequencing of target genes at high throughput, high fidelity and high sensitivity will greatly contribute to the understanding of the molecular basis of disease and the development of new therapeutics. While next generation sequencing approaches have become the methods of choice for very large scale projects such as resequencing of the whole genome or every exon, or for deep transcriptome sequencing, in many cases one is only interested in sequencing a limited area (one or a few genes or coding regions) and obtaining a rapid result with high accuracy. In these cases, next generation sequencing is unwarranted and too expensive, leaving only the traditional electrophoresis-based Sanger sequencing approach most conveniently performed with fluorescent dideoxynucleotides. Though the latter is well established, producing generally high quality and long sequence reads, it has some limitations. In addition to its being time consuming and carrying an overall high cost per read, in certain regions in which SNPs fall within or just beyond a homopolymer run, sequence quality becomes problematic. Moreover, despite substantial improvements in recent years, due to the resolution of the separation matrices and gel-based artifacts caused by secondary structures in GC rich

sequence, difficulties exist in identifying the first few bases after the priming site. One option that can overcome some of these limitations is to take advantage of mass spectrometry to distinguish the four nucleotides incorporated during the polymerase reaction.

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) has been explored widely for DNA sequencing, such as mass analysis of Sanger sequencing products,[1] mass laddering sequencing by enzymatic digestion,[2] RNA/DNA sequencing by base specific enzyme cleavage,[3] and mass analysis of oligonucleotide synthesis failure products,[4] among which direct mass analysis of Sanger sequencing products is the most straightforward approach, simply involving the replacement of gel electrophoresis determination of the size of DNA fragments with the exact molecular weight determination of the fragments. It also has the potential for sequencing longer DNA templates whereas the methods described above are only applicable for short oligonucleotides.

The idea of combining the Sanger cycle sequencing reaction with MS analysis originated with Fitzgerald *et al.* in 1993.[5] They first demonstrated the use of MALDI MS to analyze mock Sanger sequencing reaction mixtures of synthetic oligodeoxy-nucleotides ranging from 17 to 41 bases. In theory, the four nested sets of DNA fragments that are generated in Sanger sequencing reactions terminated with A, C, G or T dideoxynucleotides would each be subjected to MALDI analysis, and the four mass spectra would be overlaid to obtain the full sequence information (Fig. 3.1). However,

resolution and sensitivity were hurdles for further application of the method at that time. With the introduction of the delayed ion extraction technique for MALDI MS, Roskey *et al.* were able to sequence a synthetic ssDNA template by performing four separate

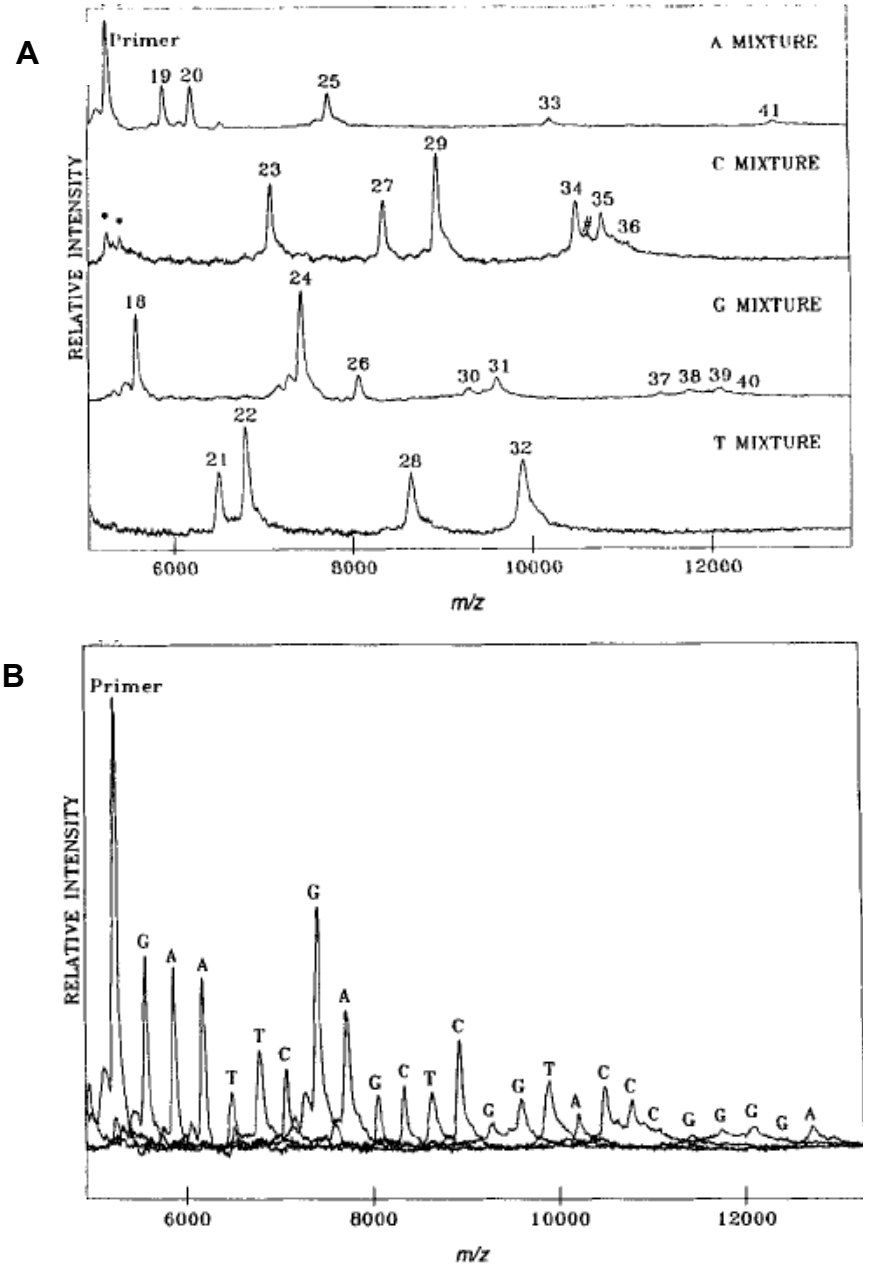


Fig. 3.1. The mass spectra of mock A, C, G, and T sequencing reactions containing mixtures of synthetic oligonucleotides. (A) Individual spectra; (B) The spectra are overlaid and displayed on the same mass scale.[5]

dideoxy termination reactions, one for each dideoxynucleotide terminator analog, to generate four mass spectra for obtaining full sequence information as described in Fitzgerald's method.[6] Mouradian *et al.* further reported MALDI analysis for sequencing an M13 bacteriophage single stranded template that was used in actual Sanger sequencing.[7] Furthermore, the potential of sequencing DNA templates of 130 bases and more by generating larger quantities of long DNA ladders was reported by Taranenko and colleagues.[8] One of the major challenges lies in the stringent purity requirement for MALDI-TOF MS analysis, which requires the analyte be free of alkaline salts and other contaminants. People have used the biotin- streptavidin interaction for purification of sequencing products. For example, Köster *et al.* succeeded in sequencing synthetic ssDNA using streptavidin-coated magnetic beads and biotinylated sequencing primers for purification of DNA ladders, removing deoxynucleoside and dideoxynucleoside triphosphates, enzyme and buffer salts before the MS analysis.[9] However, peaks corresponding to primer dimers were still observed, and the premature or false stops which are generated in Sanger-sequencing reactions when a DNA fragment terminates after incorporating a deoxynucleotide rather than a dideoxynucleotide were frequently observed. These extra peaks prevented the accurate interpretation of the sequence. Furthermore, the approach of four separate reactions, one for each dideoxynucleotide terminator, is cumbersome. And the small mass differences between the nucleotides also interfere with the accurate identification of nucleotides. Aiming at a single tube sequencing reaction with efficient elimination of salts and false stops for clean MS

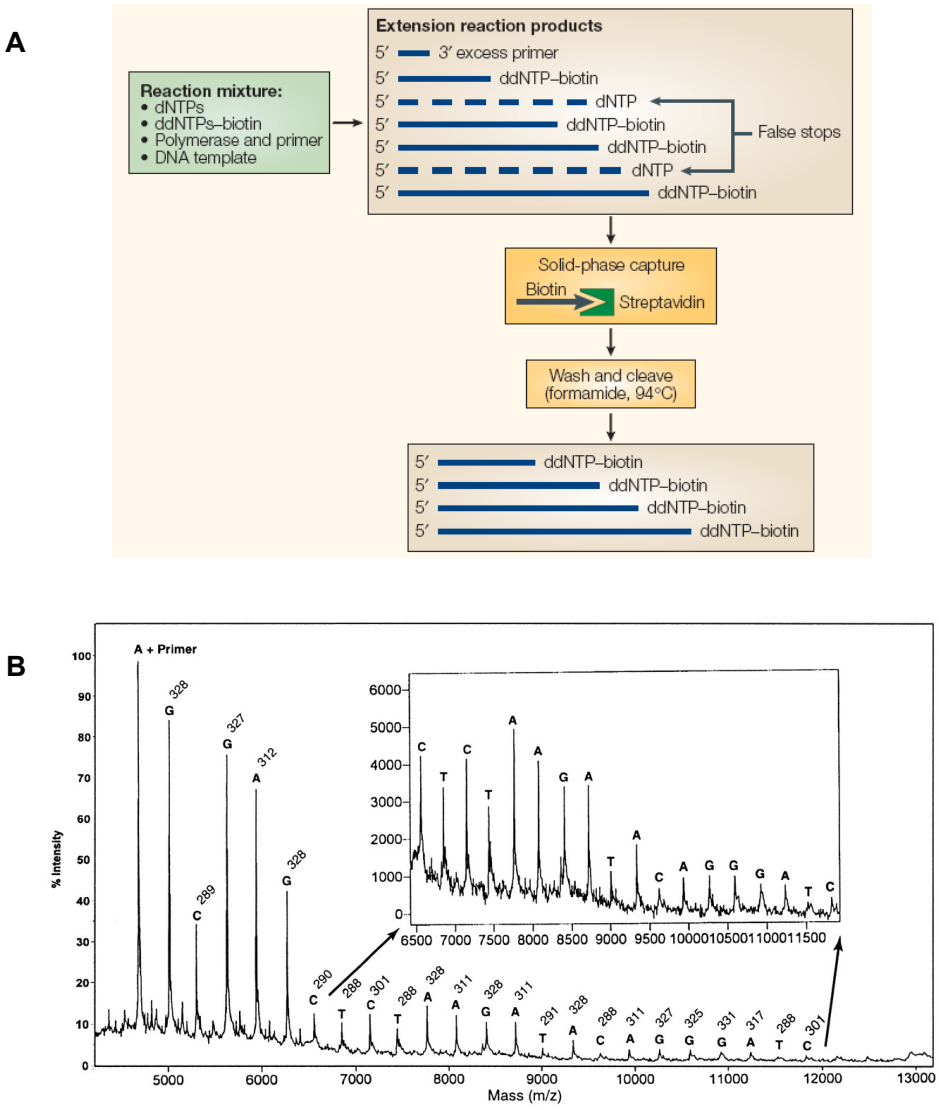


Fig. 3.2. Solid-phase-capture (SPC) sequencing. (A) A SPC-sequencing scheme to isolate pure DNA fragments for MS analysis; (B) A DNA sequencing mass spectrum generated after extension with biotinylated terminators.[11]

spectra, our group developed a solid phase capture (SPC) sequencing strategy by using biotinylated dideoxynucleotides.[10] As shown in Fig. 3.2, four deoxynucleotides (dNTPs) and four biotinylated dideoxynucleotides (ddNTP-biotins) generated by attaching biotin to the5-position of the pyrimidines and 7-position of the purines through a linker arm

were used in a Sanger sequencing reaction. Then the reaction products were subjected to streptavidin-coated magnetic beads for capturing biotin terminated fragments, whereas the false termination fragments, excess primers and other contaminants could be washed away. Finally, the Sanger sequencing fragments could be cleaved from the streptavidin-coated surface by formamide treatment at high temperature, followed by ethanol precipitation and desalting before the MS analysis. By using this approach, our group was able to unambiguously identify 24 consecutive bases.

Nevertheless, despite the improvements over the previous methods, the solid-phase capture sequencing still encountered problems in the release of fragments for MS analysis. The formamide treatment for breaking the streptavidin-biotin interaction requires the inclusion of an ethanol precipitation step, resulting in significant sample loss. Since the generation of sufficient termination products especially in the high mass range is crucial to achieve longer read-length with high resolution, it is very important to develop an alternative cleavage method which could reduce the sample loss before the MS analysis. In addition, the use of nucleotides with larger mass differences is also very important for unambiguous identification of nucleotides. Therefore, in this chapter, the newly synthesized, mass-tagged, chemically cleavable biotinylated dideoxynucleotides, described in Chapter 2, were utilized in MS-based solid phase capture sequencing by synthesis, with the hope of achieving higher efficiency and longer sequencing read-length. Our results demonstrate that read-lengths of over 30 bp are achievable on both synthetic and biological DNA templates.

3.2 Experimental Rationale and Overview

As described in Chapter 2, we have developed and characterized a set of mass-tagged, chemically cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins), wherein the biotin moiety was linked to the nucleotide through a mass tagged, azido based linker. These nucleotide analogs have proven to be good substrates for the enzyme Thermo Sequenase, and are efficiently cleaved by TCEP in a procedure that is compatible with the downstream desalting step. This allows purification of DNA products under mild conditions, which facilitates and simplifies sample handling. The mass tags on these analogs were designed to enhance the base discrimination for MS analysis, especially within the higher mass range. In this chapter, these nucleotide analogs were further evaluated by their application in solid phase capture sequencing. The overall scheme of the cleavable solid phase capture sequencing strategy is described in Fig. 3.3. A Sanger sequencing reaction consisting of DNA template, sequencing primer, natural deoxynucleotides (dNTP), biotinylated dideoxynucleotides (ddNTP-N_3-biotins) and Thermo Sequenase is first carried out to generate biotin terminated DNA-sequencing fragments. These biotinylated sequencing products are isolated from the reaction mixture containing excess primers, false stopped fragments, enzyme and salts via solid phase capture using streptavidin beads. These sequencing products are then released from the beads by cleavage of the linker between the biotin and the nucleotides, leaving biotin still bound to the bead surface. Thereafter, the sequencing products are subjected to MALDI-TOF MS analysis for sequence identification, without interference from excess

primers or falsely stopped fragments. This reversible solid-phase-capture sequencing approach was implemented as described in this chapter to sequence assorted DNA fragments.

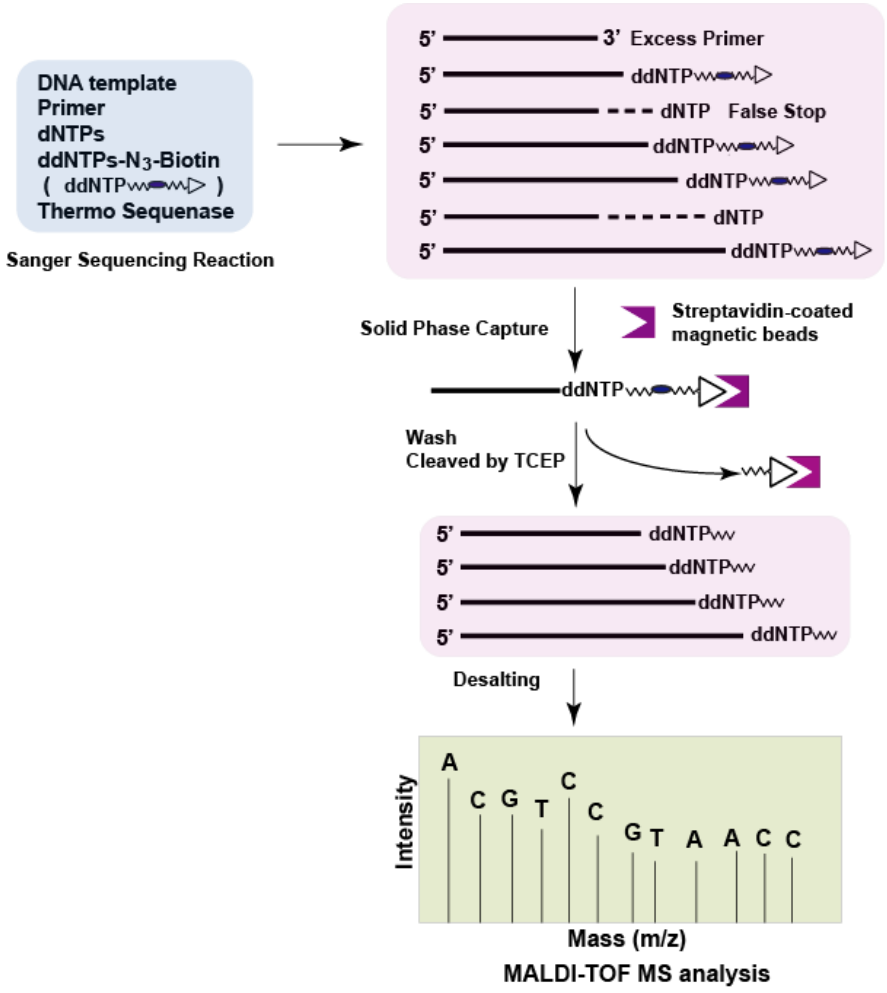


Fig. 3.3. Scheme for purification of DNA-sequencing fragments for MALDI-TOF MS analysis. DNA-sequencing fragments are isolated from the sequencing solution containing excess primers, falsely stopped fragments and salts by streptavidin-coated magnetic beads. Then the sequencing fragments are cleaved from the beads with TCEP for MALDI-TOF MS analysis, leaving the biotin moiety still bound to the surface.

3.3 Results and Discussion

The use of MALDI-TOF MS for the analysis of Sanger dideoxy base termination sequencing reactions has proven more difficult than fluorescence-based analysis of sequencing products due to the limited amounts of termination products typically produced. Therefore, the ratio of dideoxynucleotides to deoxynucleotides is critical, and the amount of DNA template and primers as well as number of sequencing cycles will also effect the sequencing product generation. Hence, different parameters have been tested and optimized to obtain reliable sequencing results.

3.3.1 DNA sequencing on synthetic template

We first investigated the application of the cleavable biotinylated dideoxynucleotides in sequencing of a synthetic single stranded DNA template. The resulting mass spectrum is shown in Fig. 3.4. The first peak in the spectrum is the primer peak plus the first nucleotide that is complementary to the corresponding nucleotide in the DNA template. The mass difference between each peak and the prior dNTP-extended primer can be measured to determine the identity of the base at each position, as each dideoxynucleotide with its attached half linker has a unique molecular weight corresponding to that peak. The read-length of over 37 bases is an improvement on the previous non-cleavable biotinylated dideoxynucletide-based sequencing results, which is mostly due to the higher cleavage efficiency and the lower product loss during the desalting step.

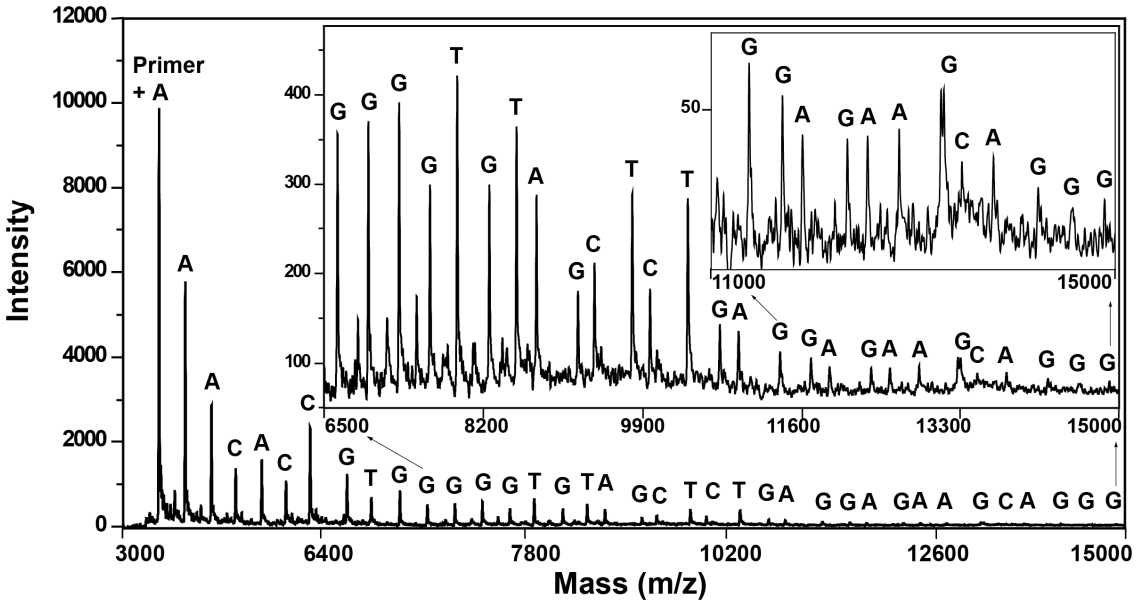


Fig. 3.4. Mass-sequencing spectrum generated using ddNTP-N_3-biotins on a synthetic template. The nested insets show increasing magnifications of the lower intensity region.

3.3.2 DNA sequencing on biological template

To further investigate its biological application, this reversible SPC-sequencing approach has been verified by sequencing a PCR product. A portion of the RHOD gene was first amplified by PCR on an anonymous sample, the result of which is shown in Fig. 3.5. To remove the excessive primers and dNTPs, exonuclease/alkaline phosphatase treatment and column purification were performed. After carrying out sequencing steps, a read-length of 32 bases was achieved, as shown in Fig. 3.6. The appearance of a few extra peaks might have been caused by incomplete purification of the PCR products since the spectrum obtained with the synthetic template was free of such extraneous peaks; however, they do not interfere with the sequence determination, since they do not correspond to any obvious mis-extended products based on the known sequence of the

RHOD gene. At the same time, traditional Sanger sequencing was performed on these PCR products to confirm the sequencing results, the result of which confirmed the accuracy of the MS sequencing (Fig. 3.7). As a proof-of-principle experiment, the sequencing of this PCR product proves the potential of using these nucleotides for sequencing biological samples. With improvements in post-PCR cleanup and further optimization of sequencing conditions, longer read-lengths with lower background signals might be achieved.

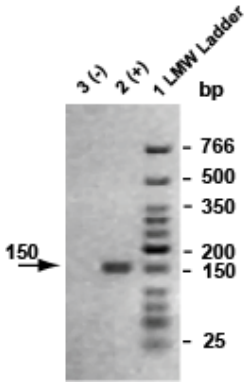


Fig. 3.5. Gel electrophoresis of PCR amplification within RHOD gene. The expected fragment size is 150 bp. 1. Low molecular weight ladder (LMW); 2, PCR reaction products (+); 3, Negative control (no DNA template, -).

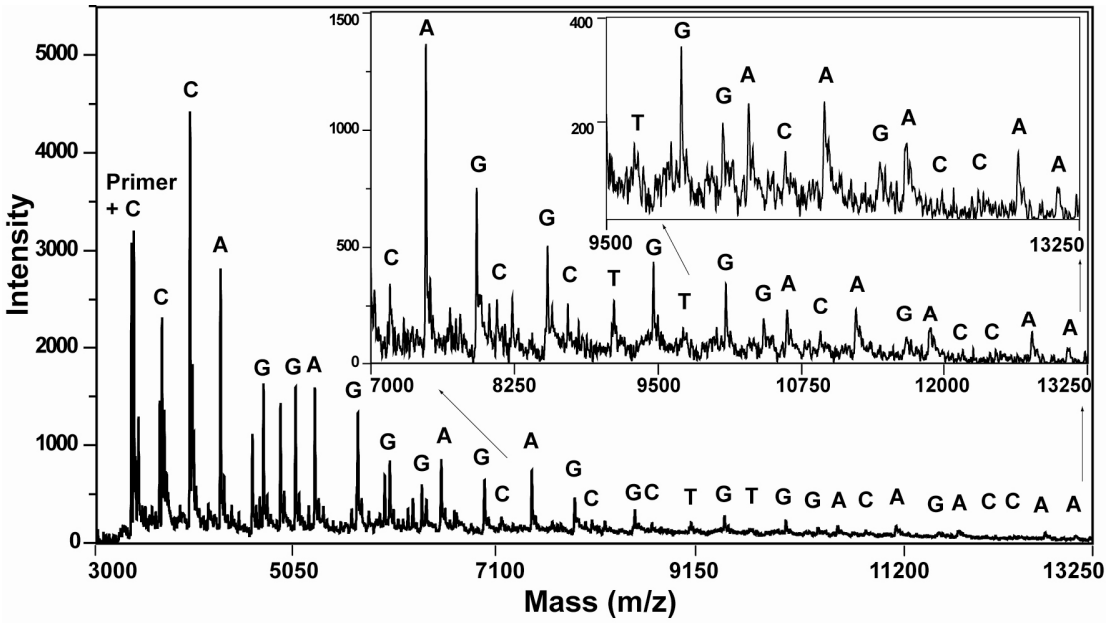


Fig. 3.6. Mass-sequencing spectrum generated using ddNTP-N_3-biotins on a PCR product. The nested insets show increasing magnifications of the lower intensity region.

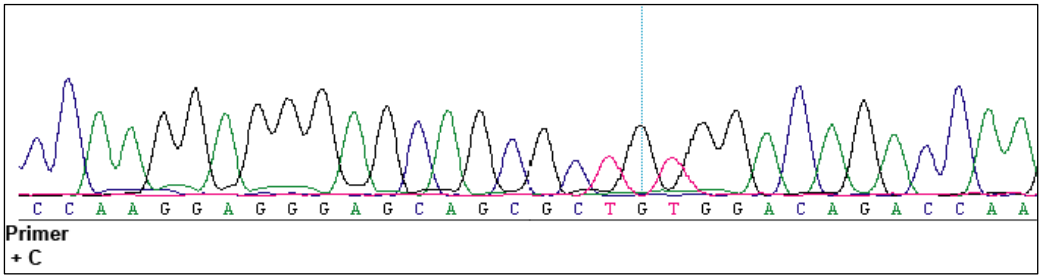


Fig. 3.7. Sanger sequencing for RHOD gene fragment, the result of which is consistent with that obtained by the MS-based sequencing.

3.4 Materials and Methods

3.4.1 Sanger DNA sequencing reaction

A synthetic 90mer template with sequence related to a portion of the RHOD (Ras Homologous Family D) gene: 5'-TGCCTCTCCGAAGCCTCCTCACACCCTCCCC-CGCCCTGCTTCTCCTCAGAGCTACACCCCCACGGTGTTTGAGCGGTACATGGT-

CAACCT-3', and the corresponding primer 5'-TGTACCGCTC-3', were first used to test the sequencing method using ddNTP-N_3-biotins. Another template, a 150 bp double stranded PCR product that contains a portion of the RHOD gene 5'-CAACT-ACGCTCCACTGACCCCCAAGGAGGGAGCAGCGCTGTGGACAGACCAAGT-CC CCAGTGCCTCTCCGAAGCCTCCTCACACCCTCCCCCGCCCTGCTTCTCC-TCAG AGCTACACCCCCACGGTGTTTGAGCGGTACATGGTCAACCT-3' was generated by PCR.

3.4.2 PCR reactions for generating biological template

The primers for the PCR reaction were: forward primer 5'-CAACTA-CGCTCCACTGACCC-3' and reverse primer 5'-AGGTTGACCATGTACCGCTC-3'. The PCR reaction was carried out in a 50 μl PCR cocktail mixture containing 10 ng template (DNA extracted from anonymous sample), 10 nmol dNTPs, 30 pmol reverse and forward primers, 1 U of JumpStart^{TM} REDTaq® DNA Polymerase (Sigma-Aldrich) and 1X corresponding polymerase reaction buffer. PCR products from this reaction were re-amplified to generate a higher yield of PCR fragments if desired. The amplification was performed under the following conditions: incubation at 94^{o}C for 2 min, followed by 36 cycles of 94^{o}C for 20 s, 58^{o}C for 30 s, 72^{o}C for 30 s and final extension at 72^{o}C for 5 min. The PCR product was subsequently treated with ExoSAP-IT (USB) and further purified using the MinElute PCR Purification Kit (Qiagen) before conducting the sequencing reaction.

3.4.3 DNA sequencing on biological template

The corresponding primer for sequencing the PCR product was 5'-CCACTGACCC-3'. Sanger sequencing reactions contained 2000 pmol each of dATP, dGTP, dCTP and dTTP; 20 pmol each of ddATP-N_3-biotin, ddGTP-N_3-biotin, ddCTP-N_3-biotin and ddUTP-N_3-biotin; 6 U of Thermo Sequenase; 1X Thermo Sequenase reaction buffer; 60 pmol synthetic DNA template (or ~ 2 μg PCR products); 300 pmol primer (200 pmol for PCR product) in a total volume of 20 μl. The sequencing reactions were subjected to 60 cycles of 94^{o}C for 30 s, 30^{o}C for 1 min (36^{o}C for PCR product) and 65^{o}C for 30 s in preparation for solid phase capture and MS analysis. For comparison, traditional Sanger sequencing using fluorescently labeled terminators was performed using standard protocols (see Materials and Methods in next Chapter for details).

3.4.4 Solid-phase purification of DNA-sequencing products for mass spectrometry measurements

The scheme for the solid-phase purification of DNA sequencing fragments is included in Fig. 3.3. The detailed procedure is described as follows. 20 μl DNA-sequencing products were combined with 40 μl streptavidin-coated magnetic beads that were prewashed with 1X B/W buffer and then re-suspended in 20 μl of 2X B/W buffer, and allowed to incubate for 1 h at room temperature. After solid phase capture, the beads containing the biotinylated DNA fragments were washed three times with 1X B/W

buffer, and three times with deionized water. Then the beads were suspended in 30 μl TCEP and incubated at 65^{o}C for 25 min. In this way, the biotin moiety was removed from the dideoxynucleotides and different lengths of DNA-sequencing fragments were released from the magnetic beads. The supernatant containing DNA sequencing fragments was desalted twice with a ZipTip and characterized by MS.

3.5 Conclusion

Taking advantage of our newly synthesized mass-tagged, cleavable biotinylated dideoxynucleotides, we have developed a reversible-solid-phase-capture MS sequencing approach and demonstrated its feasibility by obtaining a read-length of 37 bases on a synthetic template and a similar read-length on a biological sample. This approach has great potential in DNA sequencing by MS, especially for decoding short difficult stretches of DNA containing polymorphisms where fluorescence-based Sanger sequencing is inadequate, and next generation sequencing approaches are unnecessarily expensive. Considering its read-length, this MS based sequencing can also be used in RNA sequencing, such as miRNA sequencing, wherein identification of only a few bases is required. In addition, this approach has advantages over the traditional Sanger approach when sequencing in sites containing mini-indels or when sequencing through regions containing SNPs in pooled samples. Moreover, with the development of mass spectrometry, it is believed that MS sequencing will achieve much longer read-lengths.

References

1. Murray KK. DNA sequencing by mass spectrometry. *Journal of Mass Spectrometry*, **1996**, *13*, 1203-1215.

2. Pieles U, Zürcher W, Schär M, Moser HE. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res*, **1993**, *21*, 3191-3196.

3. Kirperkar F, Douthwaite S, Roepstorff P. Mapping posttranscriptional modifications in 5S ribosomal RNA by MALDI mass spectrometry. *RNA,* **2000,** *6*, 296-306.

4. Castleberry CM, Chou C-W, Limbach PA. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of oligonucleotides. *Current Protocols in Nucleic Acid Chemistry*, **2008**, 10.1.1-10.1.21.

5. Fitzgerald MC, Zhu L, Smith LM. The analysis of mock DNA sequencing reactions using matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **1993**, *7*, 895-897.

6. Roskey MT, Juhasz P, Smirnov IP, Takach EJ, Martin SA, Haff LA. DNA sequencing by delayed extraction-matrix-assisted laser desorption/ionization time of flight mass spectrometry. *Proceedings of the National Academy of Science of the United States of America,* **1996**, *93*, 4724-4729.

7. Mouradian S, Rank DR, Smith LM. Analyzing sequencing reactions from bacteriophage M13 by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, **1996**, *10*, 1475-1478.

8. Taranenko NI, Allman SL, Golovlev VV, Taranenko NV, Isoia NR, Chen CH. Sequencing DNA using mass spectrometry for ladder detection. *Nucleic Acids Research*, **1998**, *26*, 2488-2490.

9. Köster H, Tang K, Fu D-J, Braun A, Boom D, Smith CL, Cotter RJ, Cantor CR. A strategy for rapid and efficient DNA sequencing by mass spectrometry. *Nature Biotechnology*, **1996**, *14*, 1123-1128.

10. Edwards JR, Itagaki Y, Ju J. DNA sequencing using biotinylated dideoxynucleotides and mass spectrometry. *Nucleic Acids Research*, **2001**, *29*, e104.

11. Kim S, Ruparel HD, Gilliam C, Ju J. Digital genotyping using molecular affinity and mass spectrometry. *Nature Reviews Genetics***, 2003,** *4*, 1001-1008.

Chapter 4 SNP Genotyping of Mitochondrial DNA by MALDI-TOF MS using Cleavable Biotinylated Dideoxynucleotides

4.1 Introduction

The abundant sequencing data produced by the current sequencing platforms provides the opportunity to look intelligently at the genome with the fine interrogation of sequence variations, to explore the connections between genes and phenotypes, and to discover biomarkers for subtle differences among individuals or populations. As the predominant genetic variations, single nucleotide polymorphisms (SNPs) are of particular interest, and the development of an accurate, precise, time-efficient and cost-effective SNP genotyping method has emerged as one of the most highly desired genomic analysis tools. Various technologies have been developed for SNP genotyping for the human nuclear genome, yet accurate detection of SNPs with multiplex potential remains a challenge, especially in the mitochondrial genome due to its own complex genetics.

Mitochondria contain a rich number of vital enzymes, and are pivotal in cellular energy metabolism and the regulation of programmed cell death, or apotosis.[1] As shown in Fig. 4.1, the human mitochondrial genome is a double stranded, circular DNA molecule of 16569 bp, which encodes 13 structural proteins, two ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs).[2, 3] Unlike the nuclear DNA (nDNA), the

mitochondrial DNA (mtDNA) is maternally inherited, has high copy number per cell and appears as a mosaic of mutant and wild type genomes (heteroplasmy).[4] So far, more than 200 pathogenic point mutations in tRNA and rRNA genes as well as protein coding regions have been reported, and the hypervariable region of mtDNA is also highly polymorphic.[5, 6] All these features demand precise interpretation of mtDNA mutation data for disease treatment, population association studies and forensic investigations.[3, 7-9] Nonetheless, compared to the nuclear genome, it requires even higher sensitivity and accuracy SNP detection.

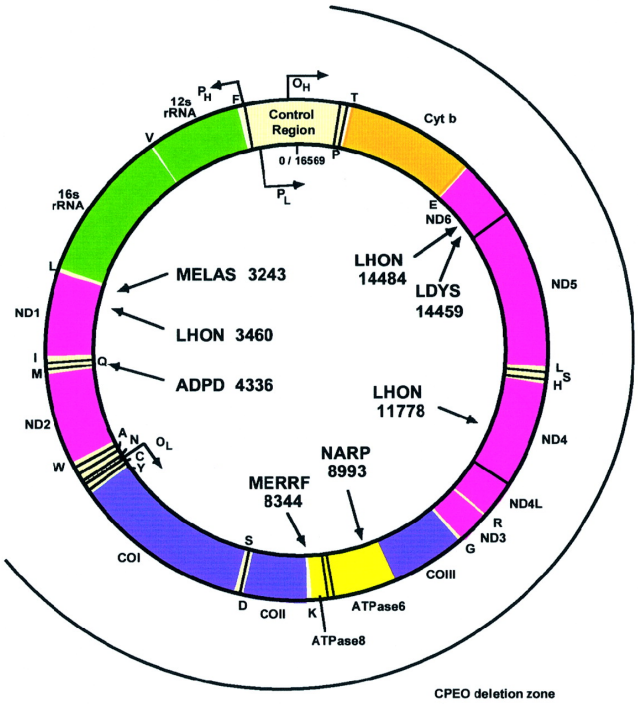


Fig. 4.1. The human mtDNA map, showing the location of selected pathogenic mutations within the 16,569-base pair genome. Genes are designated by the abbreviations outside the ring, and mitochondrial syndrome abbreviations and key mutation sites are displayed inside the ring (e.g., MERRF mutation at position 8344).

A variety of technologies have been developed for mtDNA genotyping, including the MitoChip,[10] PCR-RFLP analysis,[11] the TaqMan real-time genotyping assay,[12] and direct Sanger sequencing,[10, 13] yet each has some drawbacks.[14, 15] Most are time consuming and fail to detect low heteroplasmy levels (<10%); some, such as PCR-RFLP and the TaqMan assay, are labor intensive when detecting mutiplex SNPs; and false positives often exist due to the limitations of the detection methods. While next-generation sequencing is becoming popular for genotype analysis,[14-16] considering its cost and work flow, it is more suitable for genome-wide studies than targeted SNP genotyping. In addition, its relatively high error rate contributes to false positives and inadequate detection of low level heteroplasmy. Therefore, it is very important to develop a rapid, accurate, sensitive and cost-effective method for SNP genotyping of mitochondrial DNA.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) allows rapid and accurate sample measurement, and provides high resolution and sensitivity. It has been employed for SNP genotyping coupled with hybridization,[17] strand invasion dependent cleavage of oligonucleotides[18] and single base extension (SBE);[19-23] the latter, in particular, has emerged as a very powerful method for multiplex SNP detection. Generally, in single base extension, SBE primers designed to anneal adjacent to targeted SNP sites are extended by dideoxynucleotides that are complementary to the nucleotides at the polymorphic positions. The nucleotides can be identified by measuring the molecular weight of primer extension products. In this MS-based analysis, it is critical to isolate and stringently purify DNA primer extension

products from excess primers in the reaction mixture for high-fold multiplex SNP analysis and unambiguous detection. Hence, in previous work aimed at multiplex genotyping, we have developed a solid phase capture step for single base extension (SPC-SBE) by using biotinylated dideoxynucleotides to generate 3' biotin dideoxynucleotide terminated DNA strands, thereby allowing solid phase isolation of SBE products on streptavidin-coated beads.[23-26] However, the release of biotin terminated primers from streptavidin coated surfaces requires harsh conditions, such as treatment with formamide at a high temperature, to break the biotin-streptavidin bond. This complicates downstream procedures due to the requirement for ethanol precipitation steps which are tedious and can result in sample loss. Furthermore, the relatively small mass differences between each dideoxynucleotide reduces the effective peak resolution, which poses challenges for low level heteroplasmy detection due to potential peak overlaps.

The introduction of our newly synthesized mass-tagged, cleavable biotinylated dideoxynucleotides, ddNTP-N_3-biotins (ddATP-N_3-biotin, ddGTP-N_3-biotin ddCTP-N_3-biotin, and ddUTP-N_3-biotin), as described in Chapter 2, provides a potential way to address all these challenges. The presence of the azido group in the linker permits cleavage of the biotin to be accomplished by tris(2-carboxyethyl)phosphine (TCEP) under DNA-friendly aqueous conditions, and the length of the linker is adjusted to increase the mass differences among these nucleotide analogs. The combination of these changes will lead to improved sensitivity, accuracy and efficiency of multiplex SPC-SBE

genotyping.

In this study, we validated our chemically cleavable SPC-SBE approach in testing mtDNA samples for the myoclonic epilepsy with ragged red fibers (MERRF) syndrome. MERRF syndrome is a maternally inherited multisystemic disorder, as is often the case for mitochondrial diseases.[27, 28] Some of the most commonly used point mutations for clinical diagnosis of MERRF are m.8344A>G (A8344G),[29] m.8356T>C (T8356C)[30] and m.8363G>A (G8363A)[31] in the mitochondrial MT-TK gene encoding $tRNA^{Lys}$, and m.3243A>G (A3243G)[32] and m.3255G>A (G3255A)[33] in the mitochondrial MT-TL1 gene encoding $tRNA^{Leu}$. Among these, A8344G is the most common mutation in MERRF, present in over 80% of affected individuals,[27] while A3243G and G3255A are mutations shared with MELAS (myopathy, encephalopathy, lactic acidosis and stroke) syndrome.[33-35]

A 5-plex genotyping assay was developed to simultaneously identify variants at all five of these sites in mtDNA samples. It was demonstrated that the SPC-SBE approach using cleavable biotinylated ddNTPs was able to quantify heteroplasmy levels and detect as low as 2.5% heteroplasmy. This enables the rapid and accurate identification of mtDNA SNP variants at high sensitivity and specificity, offering great potential for mitochondrial disease diagnosis and other genetic testing demands.

4.2 Experimental Rationale and Overview

In this Chapter, the set of mass tagged, chemically cleavable biotinylated

dideoxynucleotides synthesized in our lab has been introduced into our previously developed solid-phase-capture (SPC)-single-base-extension (SBE) approach for SNP genotyping, taking into account the advantages of these nucleotide analogues, mild cleavage conditions compatible with the downstream processes leading to higher sample recovery, and enlarged mass differences for higher nucleotide discrimination. As shown in Fig. 4.2, our cleavable SPC-SBE SNP genotyping method is described as follows. The target regions of genomic DNA spanning the SNPs were first amplified by PCR to generate sufficient template for detection of SNPs. Single base extensions with cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins), SBE primers designed to anneal adjacent to the SNP loci, and Thermo Sequenase were carried out to produce biotinylated end labeled extension products. Solid phase capture was performed to purify the extension fragments from the reaction mixture, and SBE products were then released from the streptavidin surface by treatment with TCEP and subjected to MALDI-TOF MS analysis after a direct desalting step, leaving the biotin moiety still bound to the surface.

To employ this cleavable SPC-SBE method, the mitochondrial DNA carrying SNPs leading to MERRF were chosen as the genomic DNA target for the evaluation, as mitochondrial genotyping requires higher sensitivity and accuracy. Both a uniplex and 5-plex genotyping method was developed, with particular focus on the A8344G locus for evaluating the features of our protocol in low heteroplasmy detection, and quantification of heteroplasmy.

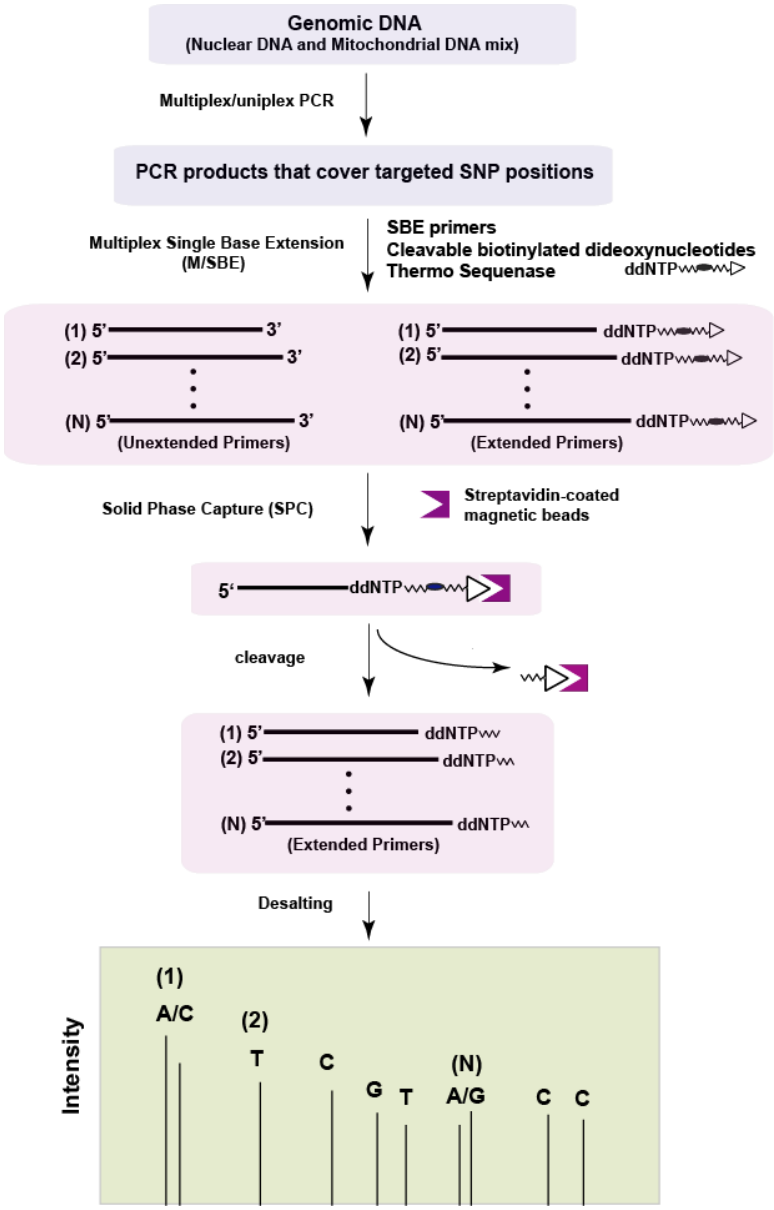


Fig. 4.2. The SPC-SBE approach for multiplex genotyping by MALDI-TOF MS using cleavable biotinylated dideoxynucleotides. DNA fragments that contain target SNP positions are generated by uniplex or multiplex PCR, and serve as templates for single base extension reactions. A set of SBE primers that are adjacent to SNP sites are used to generate single base extension products which are then isolated from the reaction mixture containing unextended primers, salts and other contaminants by streptavidin-coated magnetic beads. SBE products are then cleaved from the beads with TCEP in preparation for MALDI-TOF MS analysis, leaving the biotin moiety still bound to the bead surface.

4.3 Results and Discussion

4.3.1 PCR amplification of targeted region in mitochondrial DNA

For the PCR reaction, stringent annealing conditions were chosen to prevent non-specific amplification. Different amplification conditions were tested after the initial design of the PCR primers. According to the melting temperature of the PCR primers, three different annealing temperatures were tested, 53^{o}C, 55^{o}C and 58^{o}C, with the same amount of starting DNA sample. Two different amounts of starting DNA sample were used at the same annealing temperature of 55^{o}C. As shown in Fig. 4.3A, the desired fragments were successfully amplified under all conditions. The highest temperature of 58^{o}C was chosen for all the PCR reactions in this study.

Though mitochondria have much higher copy numbers, the DNA extract is a mixture of mitochondrial DNA and the 6 orders of magnitude larger nuclear genome. To verify the specific amplification of mitochondrial DNA, PCR amplification of a mitochondrial DNA-negative cell (rho^{o} cell) [28, 36] was performed under the same condition. Both 10 ng and 100 ng of rho^{o} DNA were tested. The results shown in Fig. 4.3 b confirmed specific amplification from the mitochondrial genome, since the mitochondrial DNA negative sample failed to show any amplification bands. For most PCR reactions in this study, 58^{o}C was set as the annealing temperature and 100 ng of DNA template was used. The uniplex PCR results for different mitochondrial samples are shown in Fig. 4.4. These conditions were used for PCR amplification of an artificial mixture for quantitative analysis, as will be described in a later section.

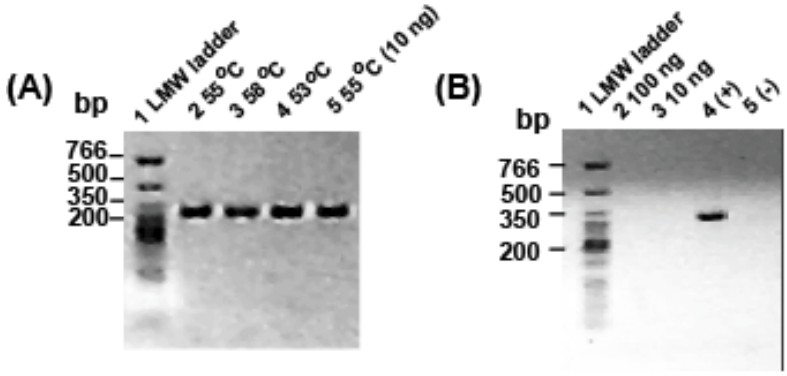


Fig. 4.3. Gel electrophoresis of PCR condition testing. A. PCR on mitochondrial DNA at different annealing temperatures and template concentrations: 1. Low Molecular Weight (LMW) ladder, 2, 55°C (100ng), 3, 58°C (100ng), 4, 53°C (100ng), 5, 55°C (10ng); B. PCR on mitochondrial DNA-negative sample: 1. LMW ladder, 2, 100ng, 3, 10ng, 4, positive control (mitochondrial DNA sample, +), 5, negative control (no template, -)

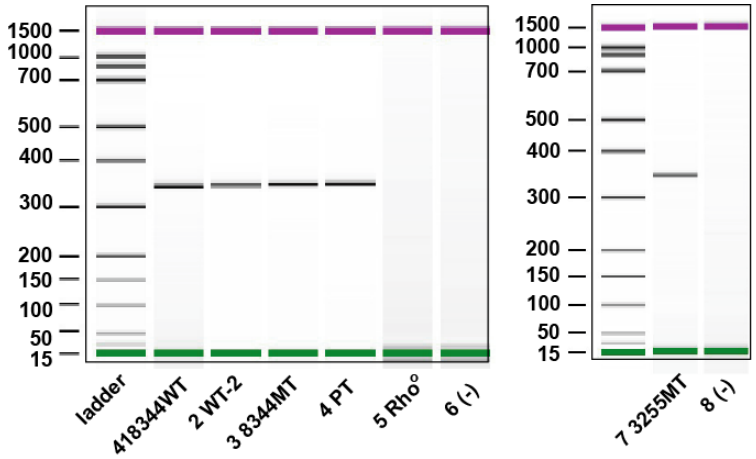


Fig. 4.4. PCR amplification of region 1 containing MERRF SNP at mitochondrial genome position 8344 on different samples: channel 1, 8344 wild type sample (8344WT), 2, anonymous healthy donor (WT-2), 3, 8344 mutant sample (8344MT), 4 patient sample (PT), 5, mitochondrial DNA-negative cell line (Rho^0), 6, negative control (-), 7, 3255 mutant sample (3255MT), 8, negative control.

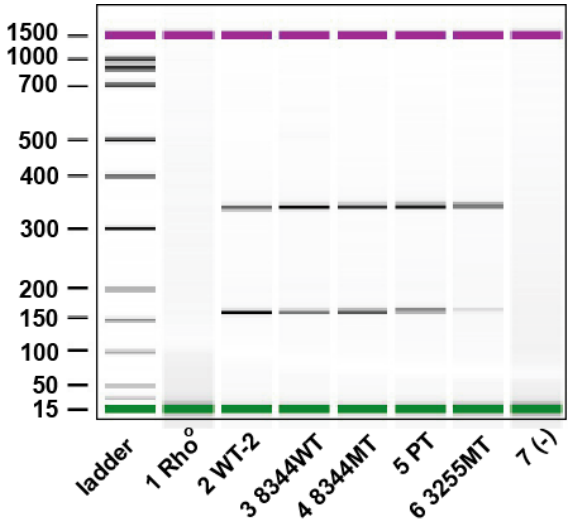


Fig. 4.5. 2-plex PCR on different samples: channel 1, mitochondrial DNA-negative (Rho^0) cell line (Rho^0); 2, anonymous healthy donor (WT-2); 3, 8344 wild type sample (8344WT); 4, 8344 mutant sample (8344MT); 5 patient sample (PT); 6, 3255 mutant sample (3255MT); 7, negative control.

Based on these conditions, we further developed a 2-plex PCR strategy to amplify the region covering all the target SNPs of interest. As shown in Fig. 4.5, two bands corresponding to two regions in the mitochondrial DNA were accurately amplified, with the expected sizes of 157bp and 339 bp. Again, there was no non-specific amplification of a mitochondrial DNA negative sample.

4.3.2 A8344G uniplex genotyping

A8344G is one of the most frequent mutations in the MERRF syndrome. Genotyping by the SBE-SPC approach using cleavable biotinylated dideoxynucleotides was first validated for this site in 6 different samples: homoplasmic wild type 8344WT, homoplasmic mutant type 8344 MT, and homoplasmic mutant type 3255MT cybrids,

patient sample 8344PT, anonymous healthy sample 8344WT-2 and a mitochondrial DNA-negative cell line (rho^0). As shown in Fig. 4.6, the nucleotide at position 8344 was

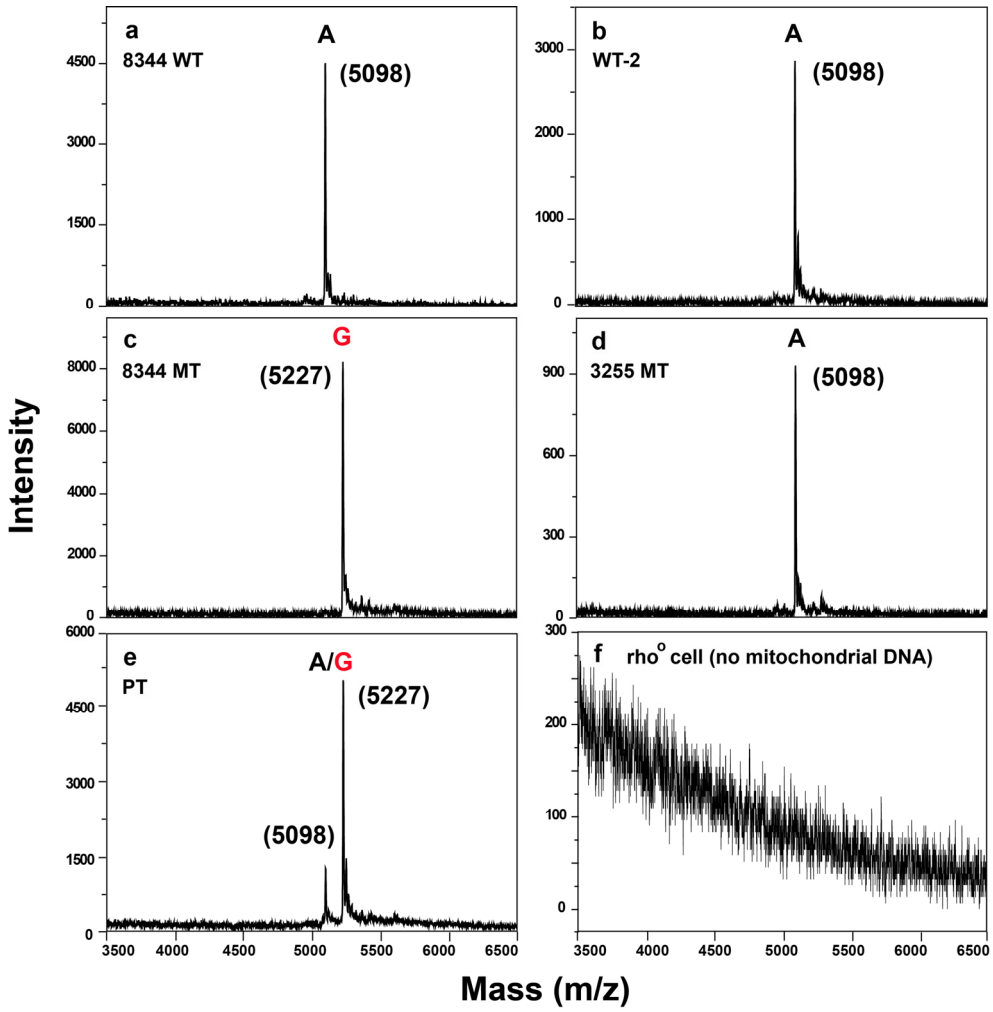


Fig. 4.6. Mass spectrometric detection of nucleotide variation at mtDNA locus 8344 from (a) 8344 wild-type sample; (b) anonymous healthy subject; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample; (f) mitochondrial DNA-negative cells. Letters in red refer to the mutant type.

identified correctly for each sample, and no false-positive peak appears for the rho^0 DNA sample. The molecular weight of the 8344 SBE primer is 4648 Da, therefore, the peak at 5098 m/z corresponds to an extension product with ddATP-N_3-biotin (WT) after cleavage,

while the peak at 5227 m/z corresponds to an extension product with ddGTP-N_3-biotin (MT) after cleavage. Fig. 4.6 3a and b are indicative of the homoplasmic wild type form (A) at locus 8344, Fig. 4.6 3c shows the homoplasmic mutant type form (G) at locus 8344, Fig. 4.6 3d shows the wild-type and Fig. 4.6 3e confirms the presence of heteroplasmy (A/G), which was in good concordance with a separately performed RFLP assay. Sanger sequencing of each DNA sample further confirmed the accuracy of this method (Fig. 4.7)

We further evaluated the specificity of this chemically cleavable SPC-SBE approach by analyzing a mitochondrial DNA-negative cell line (rho^0). No single base extension products were detected by mass spectrometry (Fig. 4.6 3f).

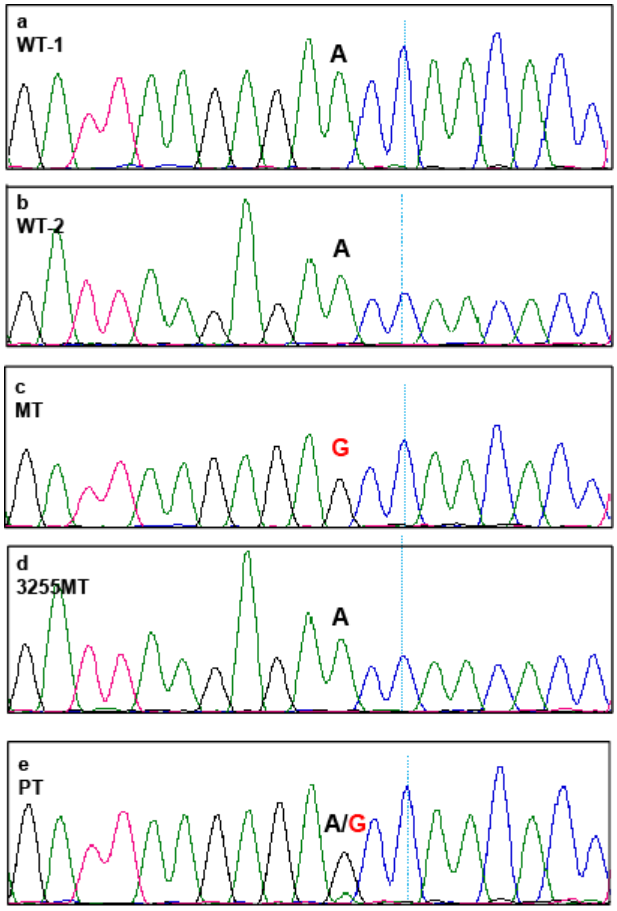


Fig. 4.7. Sanger sequencing results on different samples. (a) 8344 wild-type sample; (b) anonymous healthy subject; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample.

4.3.3 5-plex genotyping

Since multiple SNP sites are routinely screened in the clinical diagnosis for the MERRF syndrome, a 5-plex genotyping assay was developed. As shown in Fig. 4.8, nucleotides at these five SNP sites were identified unambiguously for all the samples, while the negative result in the mass spectrum for the rho^0 cells further confirms the specificity of the assay. The molecular weights of primer peaks and their potential extension products are listed in Table 4.1 in the Materials and Methods section of this chapter. To ensure sufficient mass differences between the various primers, reverse SBE

primers that anneal to the complementary strand were selected for mutation sites G3255A,[33] T8356C,[30] and G8363A.[31] Therefore, in the mass spectrum for the mutant type, transitions from C to U at locus 3255, A to G at locus 8356 and C to U at locus 8363 were expected. As shown in Fig. 4.8, in accordance with the uniplex genotyping results, only 8344MT and PT have A to G mutations at locus 8344. Though lacking the mutation at position 8344, 3255MT has a C to T mutation at position 3255 (Fig. 4.8 d), consistent with previous information. No false positives were detected on the rho^0 DNA sample (Fig. 4.8 f). According to Table 4.1 and the mass spectrometric results (Fig. 4.8), samples 8344MT, PT and 3255MT were diagnosed as MERRF, while 8344 WT and WT-2 were determined to be clear of the disease, which is consistent with known information for these samples.

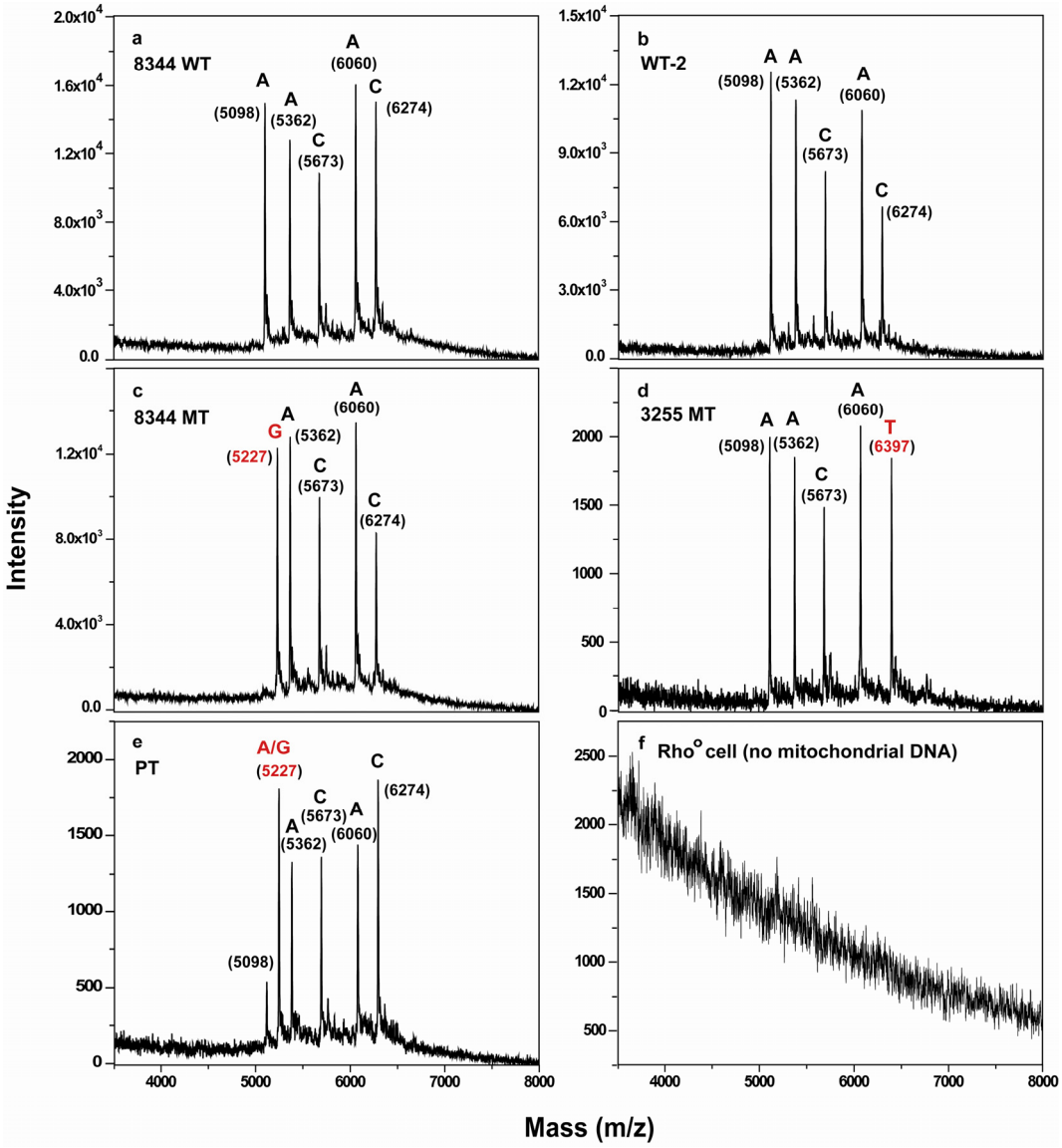


Fig. 4.8. Mass spectrometric results of 5-plex SNP genotyping for MERRF syndrome. (a) 8344 wild-type sample; (b) anonymous healthy donor; (c) 8344 mutant sample; (d) 3255 mutant sample; (e) patient sample; (f) mitochondrial DNA-negative cell line. Red labels refer to the mutated forms.

4.3.4 Quantitative mutation analysis for mitochondrial sample

In the mitochondrial genome, DNA mutations often coexist with wild-type DNA. In contrast to heterozygous mutations in nuclear DNA, mutated and wild-type forms of

mtDNA rarely occur at 1:1 ratios. Instead, heteroplasmy (ratio of wild-type to mutant mitochondrion in cell or tissue source) can occur at any ratio between 0:100 and 100:0. And this proportion will drift in time and space following cell division.[7] Different levels of heteroplasmy are strongly associated with different clinical manifestations, and the disease phenotype becomes evident when mutated forms of the mitochondrial DNA exceed a threshold level.[37] Hence, for better understanding of the disease, quantitative analysis is of particular interest. Taking A8344G as an example, the capability of performing quantitative analysis for mtDNA heteroplasmy was evaluated. Before mixing homoplasmic WT and MT samples in ratios to mimic heteroplasmic states, real time PCR was first performed to estimate the amount of mitochondrial DNA in these samples. The results shown in Fig. 4.9 indicated that the 8344MT and 8344WT samples contained the same amount of mitochondrial DNA. Nonetheless, considering possible preferential amplification of wild-type and mutated forms, purified homoplasmic PCR products were combined to generate another set of artificial mixtures for quantitative analysis. Regardless of whether the DNA was mixed before or after PCR to generate these standard curves, the results indicated that the PT sample contains ~ 80% to 90% heteroplasmy at locus 8344. Quantitative results for the post-PCR mixing experiment and the direct sample mixing experiment is shown in Fig. 4.11 A and B, respectively. The graph was generated from the normalized signal intensities of MT and WT peaks in the mass spectra. By comparing the mass spectrum of the patient sample (right-most bars) with the resulting standard curve, it was suggested that its heteroplasmy falls within the

range of 90% to 80%, which is consistent with the information provided for this patient. The results demonstrated that this cleavable SPC-SBE approach was able to estimate the heteroplasmy level from the quantitative calibration curve.

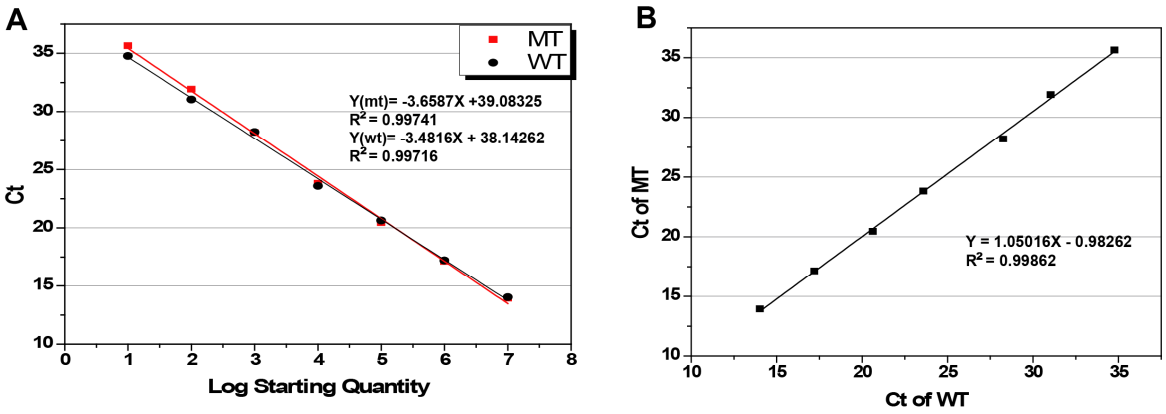


Fig. 4.9. Real-time PCR for quantitative analysis of A8344G homoplasmic wild type and mutant type samples. A. The Ct value versus the concentration of DNA for both MT and WT samples; B. Direct comparison between MT and WT using linear regression.

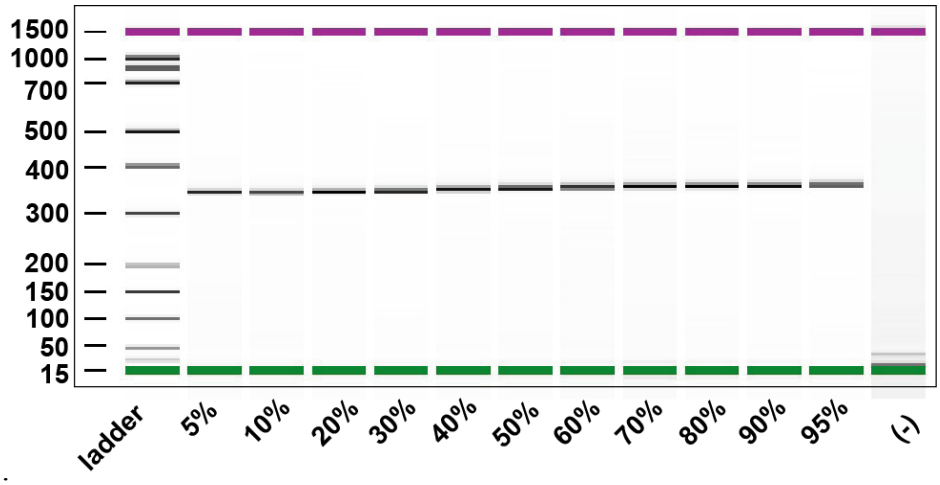


Fig. 4.10. PCR result for different levels of artificial heteroplasmy (%).

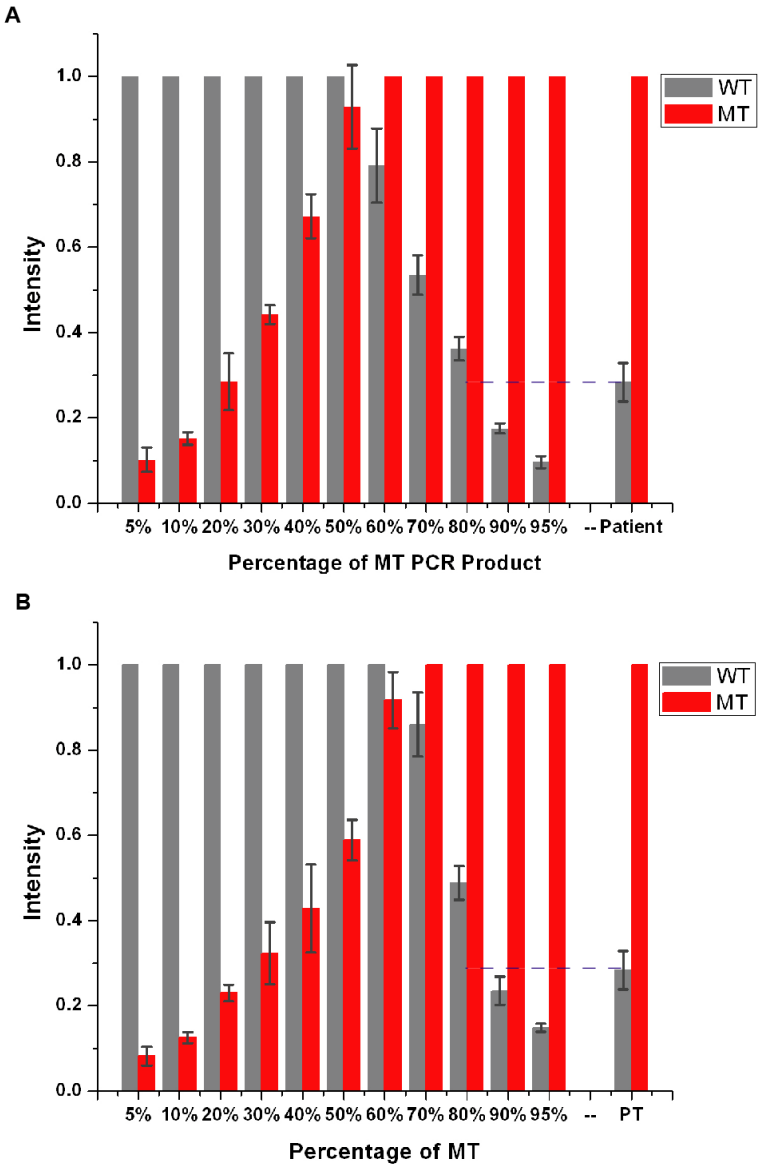


Fig. 4.11. Quantitative analysis of position 8344 heteroplasmy in mitochondrial DNA based on various mixture ratios of purified PCR products (A) and original biological samples (B).

4.3.5 Sensitivity of SPC-SBE MALDI-TOF MS for detecting low heteroplasmy of mitochondrial DNA

For early detection, it is necessary to discriminate a very small amount of mutated mtDNA in a majority of wild-type background. The large mass differences between our

new modified nucleotides were designed to generate sufficient distance between wild-type and mutated SBE product peaks in the mass spectrum. Our solid phase capture approach and efficient cleavage chemistry removed interference around the major product peak. These modifications make it possible to detect low level heteroplasmy with very high resolution. Taking 8344 mutation detection as an example, the sensitivity of this assay was validated. With sample mixing based on two types of information (real time PCR or measurement of PCR product concentration), as low as 2.5% heteroplasmy was correctly identified, as shown in Fig. 4.12 A and B. Similarly, as low as 2.5% wild-type in a mutant background was detectable. This further confirms the accuracy of the assay. For comparison, Sanger sequencing and PCR-RFLP assays were also performed. As seen in Fig. 4.12 C, neither 5% of mutant (a) nor normal mtDNA (b) were convincingly detected by Sanger sequencing analysis. In a PCR-RFLP assay, a 127 bp PCR product was digested with the restriction endonuclease *Hae*III. The use of mismatched primer in the PCR reaction enabled the production of extra restriction site for *Hae*III if A is mutated to G at position 8344. Therefore, for wild type mitochondrial DNA sample, *Hae*III cut PCR products into two fragments: 69 and 57 bp. While for PCR product from mutant type mitochondrial DNA, the 69 bp fragment is further cleaved into a 48 bp and a 21 bp fragment. Therefore, the final digested fragments for normal DNA are 57 and 69 bp, whereas for mutant DNA are 57, 48 and 21 bp. As shown in Fig. 4.12 D, this ethidium bromide (EB)-stained PCR-RFLP assay also failed to detect 5% heteroplasmy. Thus the MS-based chemically-cleavable SPC-SBE approach apparently

outperformed these two approaches in terms of sensitivity.

Since mtDNA mutations might accumulate with time, the high sensitivity and accuracy of low heteroplasmy level detection will be very helpful for routine health screening in terms of disease prevention. Although the pathogenic threshold is usually above 70%,[7] dominant mutations also exist. For example, it has been reported that the pathogenic threshold of a particular mitochondrial disorder was only 4 to 8%,[38] in which case low heteroplasmy detection becomes critical. Radio-labeled RFLP or real time PCR assays might meet the need; nonetheless, the mass-based chemically-cleavable SPC-SBE approach outperforms them in terms of simplicity, lower cost, reduced labor and the potential for higher multiplexing.

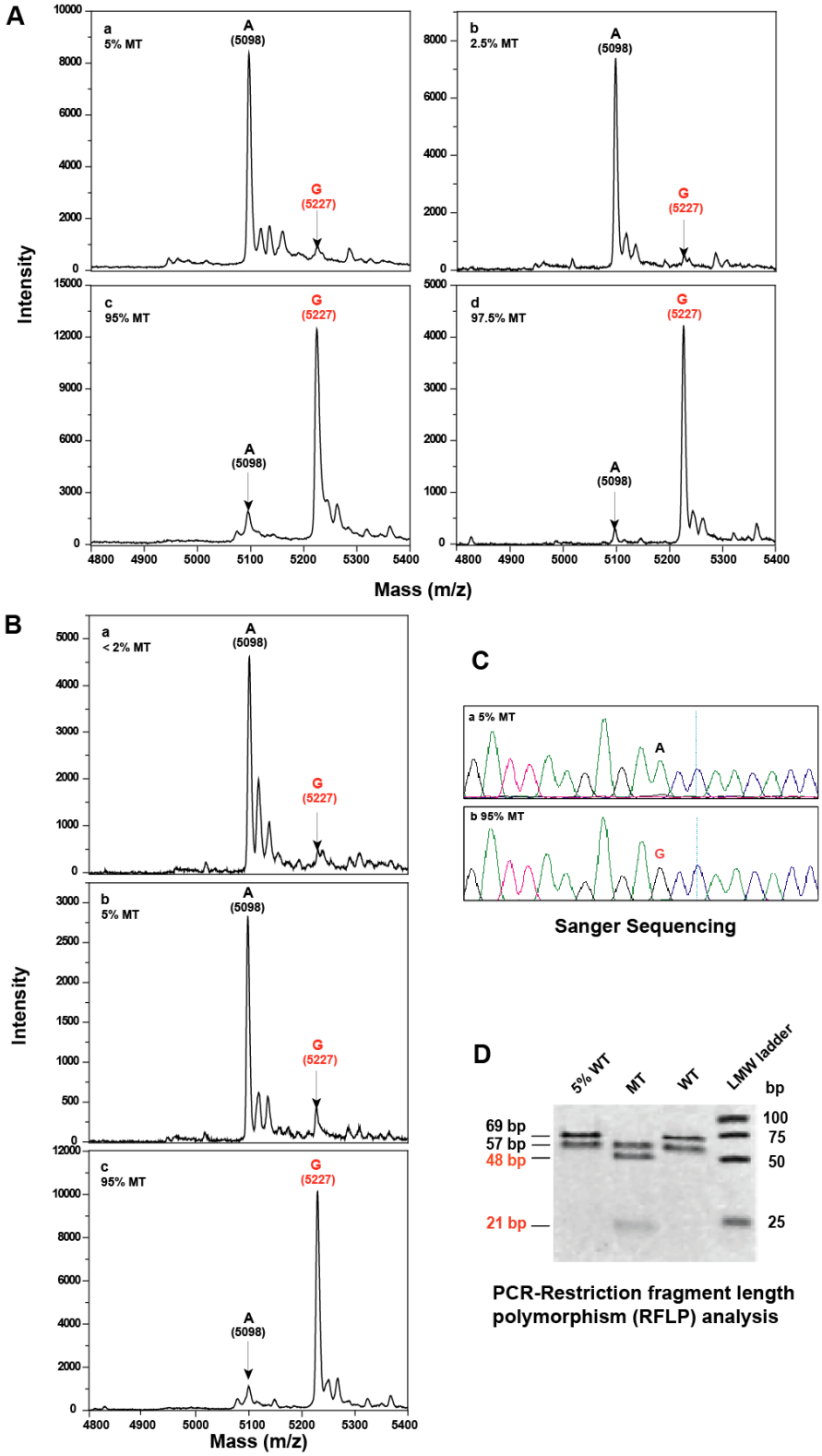


Fig. 4.12. Detection of low 8344 heteroplasmy levels for sensitivity testing. (A) Detection of 5% and 2.5% heteroplasmy, based on original sample DNA mixing, by MALDI-TOF MS; (B)

Detection of 5% and 2% heteroplasmy, based on purified PCR products mixing, by MALDI-TOF MS (A) and (B) shows the reproducible detection of <5% heteroplasmy. (C) Sanger sequencing analysis for 5% and 95% mixtures; (D) PCR-RFLP analysis for wild-type，mutant type，5% heteroplasmy and anonymous patient sample. LMW refers to low molecular weight.

4.4 Materials and Methods

General Information. DNA samples were extracted from homoplasmic 8344 wild type (8344WT), homoplasmic 8344 mutant type (8344 MT), and homoplasmic 3255 mutant type (3255MT) cybrids, a mitochondrial DNA-negative cell line (rho^0), a MERRF patient with the 8344 point mutation (8344PT), and an anonymous healthy donor (8344WT-2). The cleavable dideoxynucleotides were synthesized in our laboratory.

4.4.1 PCR amplification

The PCR primers were initially designed with Primer3 (http://frodo.wi.mit.edu/primer3) but further adjusted after BLAST search in order to avoid non-specific amplification of nuclear DNA. A uniplex PCR reaction was performed to amplify region 1 containing three of the SNP sites: A8344G, T8356G and G8363A, while a 2-plex PCR reaction was used to amplify regions 1 and 2 simultaneously, the latter of which covers the other two SNP sites: A3243G and G3255A. The primers were 5'-GACCGGGGGTATACTACGGT-3' (region 1, forward), 5'-GGAGGTAGGTGGT-AGTTTGTG-3' (region 1, reverse), and 5'-GACCGGGGGTATACTACGGT-3' (region 2, forward), and 5'-GGAGGTAGGTGGTAGTTTGTG-3' (region 2, reverse). Therefore,

the DNA fragments that are amplified from the 2-plex PCR reaction include all 5 of the above frequently mutated SNP sites used for the clinical diagnosis of MERRF syndrome. The uniplex PCR reaction was carried out in 30 μl of a PCR cocktail mixture containing 100 ng mitochondrial DNA, 10 nmol dNTPs, 8 pmol each of reverse and forward primers (region 1, forward and reverse) (oligonucleotides purchased from Integrated DNA Technologies, Coralville, IA or Eurofins MWG Operon, Huntsville, AL), 0.5 U JumpStart™ REDTaq® DNA Polymerase (Sigma-Aldrich, St. Louis, MO) and the corresponding 1X polymerase reaction buffer. The 2-plex PCR reaction consisted of 100 ng mitochondrial DNA, 12 nmol dNTPs, 7 pmol of each reverse and forward primer, 0.5 U JumpStart™ REDTaq® DNA Polymerase and 1X polymerase reaction buffer in a total volume of 30 μl. Amplification was performed under the following conditions: incubation at 94°C for 2 min, followed by 25 cycles of 94°C for 20 s, 58°C for 30 s, 72°C for 30 s and final extension at 72°C for 3 min. After PCR, 1 μl of PCR product from each sample was run on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to visualize the product quality. To remove excess dNTPs and primers, PCR products were subsequently treated by ExoSAP-IT (USB/Affymetrix, Cleveland, OH), with incubation at 37°C for 15 min and enzyme deactivation at 80°C for 15 min.

4.4.2 Quantification of mtDNA in the samples

The amounts of mitochondrial DNA in the homoplasmic samples, 8344MT and 8344WT, were quantified by TaqMan real-time PCR. Region 1, which is specific to the

mitochondrial genome, was chosen to compare the relative quantities of mtDNA between these two samples. Ten-fold serial dilutions from 10 ng/µl to10^{-5} ng/µl were carried out on both 8344MT and 8344WT samples before real-time PCR. The real time PCR reaction cocktail consisted of 1 µl of diluted DNA sample, 4 pmol each of reverse and forward primer and 10 µl of diluted SYBR green mixture (Applied Biosystems/Life Technologies, San Diego, CA) in a total volume of 20µl. The reaction was performed under the following conditions: incubation at 50^{o}C for 2 min and 95^{o}C for 2 min, followed by 40 cycles of 95^{o}C for 15 s, 60^{o}C for 30 s. Ct values for each dilution of both samples were collected and the graph of Ct values against concentration was plotted to determine the relative amount of mtDNA in samples 8344MT and 8344WT. Based on the quantification results, the 8344MT and 8344WT samples were mixed in a series of ratios (2.5%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97.5% mutant type) for heteroplasmy quantitation and sensitivity analysis, targeting the SNP at position A8344G.

Taking into account the potential for preferential PCR amplification, purified homoplasmic PCR products instead of the original DNA sample were used to produce the artificial mixtures. Briefly, 8344MT and 8344WT mtDNA samples were first amplified by uniplex PCR reactions to separately generate homoplasmic mutant and wild-type PCR products. Then PCR products were treated with ExoSAP-IT and purified twice using the MinElute PCR Purification Kit (Qiagen, Valencia, CA). The PCR product concentrations were measured with a NanoDrop 2000 spectrometer (Thermo Scientific,

Wilmington, DE, USA), after which they were combined in the above ratio series.

4.4.3 Single base extension using cleavable biotinylated dideoxynucleotides for MALDI-TOF MS detection

The primers for each SNP site are listed in Table 4.1. SBE reactions with the uniplex PCR products included 3 μl ExoSAP-IT treated PCR products (or ~80ng purified PCR products), 10 pmol of 8344 primers, 15 pmol ddATP-N_3-Biotin, 45 pmol ddGTP-N_3-Biotin, 4 U Thermo Sequenase (GE Healthcare, Piscataway, NJ), and 1X Thermo Sequenase reaction buffer in 20 μl total volume. For 5-plex genotyping, multiple single base reactions (MSBE) included 4 μl ExoSAP-IT treated 2-plex PCR products, 36 pmol 8344Primer, 20 pmol 8356Primer, 26 pmol 8363Primer, 26 pmol 3243Primer, 34 pmol 3255Primer, 67.5 pmol ddATP-N_3-biotin, 90 pmol ddGTP-N_3-biotin, 90 pmol ddCTP-N_3-biotin, 90 pmol ddUTP-N_3-biotin, 6 U Thermo Sequenase, and 1X reaction buffer in a 20 μl total volume. All reaction mixtures underwent 30 cycles of the extension reaction in a single tube at 94^{o}C for 20s, 42^{o}C for 45 s, 68^{o}C for 90 s and a final extension at 68^{o}C for 3 min. The SBE products were then purified by solid phase capture on streptavidin-coated magnetic beads to remove unextended primers, salts and other contaminants. Briefly, 20 μl SBE or MSBE products were combined with 20 μl streptavidin-coated magnetic beads (Dynabeads®MyOne^{TM} Streptavidin C1, Invitrogen/Life Technologies) that were prewashed with 1X B/W buffer, then resuspended in 20 μl of 2X B/W buffer, and allowed to incubate for 1 h at room

temperature. After solid phase capture, the beads containing the biotinylated DNA fragments were washed three times with 1X B/W buffer, and three times with deionized water. The beads were suspended in 20 µl tris(2-carboxyethyl)phosphine (TCEP) (Sigma-Aldrich) aqueous solution (pH 9.0, adjusted with ammonium hydroxide) and incubated at 65°C for 15 min to cleave the biotin. The supernatant containing SBE fragments was desalted with a ZipTip (Millipore, Bellerica, MA) and characterized with a Voyager DE™ MALDI-TOF mass spectrometer (Applied Biosystems). The entire process is shown in Fig. 4.1.

Table 4. 1SBE sites and corresponding SBE primers to generate DNA extension products.

Primer (SNP sites)	Sequences (5'-3')	Primer mass (Da)	Mass of single base extension product (Da)			
			ddATP-N_3-biotin	ddGTP-N_3-biotin	ddCTP-N_3-biotin	ddUTP-N_3-biotin
8344	TTAAAGATTAAGAGA	4648.2	**5098.2**	**5227.2**	5075.2	5189.2
8356	GGGGCATTTCACTGTA	4912.3	**5362.3**	**5491.3**	5339.3	5453.3
8363	ATTTAGTTGGGGCATTT	5246.5	5696.5	5825.5	**5673.5**	**5787.5**
3255	TAAAGTTTTAAGTTTTATG	5846.9	6296.9	6425.9	**6273.9**	**6387.9**
3243	GGTTTGTTAAGATGGCAG	5609.7	**6059.7**	**6188.7**	6036.7	6150.7

Note: Reverse SBE primers were used for sites 8356, 8363, and 3255. The masses of the extension products shown in bold black refer to wild-type; those in bold red refer to the mutant type.

4.4.4 Direct Sanger DNA Sequencing and PCR-RFLP Assay

The Sanger sequencing reaction mixture included 0.5 µl of BioDye Terminator 3.1 cycle sequencing mix (Applied Biosystems), 1X buffer 3.1, and 5 pmol forward or reverse primer. The reaction was carried out as follows: 25 cycles of 94°C for 20 s, 55°C

for 45 s, 68°C for 90 s and final extension at 68°C for 3 min. After clean-up of sequencing products by sodium acetate/ethanol precipitation and solubilization in formamide at 94°C, Sanger sequencing was performed on an ABI3730xl DNA analyzer (Applied Biosystems).

The primers used in the PCR-RFLP assay were 5'-TTAAGTTAAAGATTAAGA-GG-3' (forward), and 5'-TATTTTTAGTTGGGTGATGAGGAA-3' (reverse), wherein the 3' end of the forward primer is mismatched to locus 8343 in the mitochondrial genome, generating a restriction site for *Hae*III if position 8344 is mutated from A to G. The PCR reaction mixture consisted of 100 ng mitochondrial DNA, 10 nmol dNTPs, 10 pmol each of reverse and forward primers, 0.5 U JumpStartTM REDTaq® DNA Polymerase and the corresponding 1X polymerase reaction buffer. The amplification was carried out with 30 cycles of 94°C for 1 min, 50°C for 1 min, 72°C for 1 min. Then 20µl of PCR product was treated by adding 10 U *Hae*III (New England Biolabs, Ipswich, MA), 4 µl of 10X buffer4 and 15 µl H_2O at 37°C. The product size was checked by separation in a 15% polyacrylamide gel (Bio-Rad, Hercules, CA).

4.5 Conclusion

Expanding on our previously established SPC-SBE method, we introduced our new synthesized chemically cleavable biotinylated dideoxynucleotides into the SBE reactions, which enabled highly effective isolation and purification of the SBE products and achieved higher resolution through larger mass differences between the nucleotides and

their corresponding SBE products. This SPC-SBE approach coupled with MALDI-TOF mass spectrometry enables rapid, accurate and sensitive analysis of mitochondrial DNA single nucleotide polymorphisms and different levels of heteroplasmy. This methodology is highly targeted and cost-effective, and is compatible with automated micro/nanofluidic systems, enabling high-throughput SNP screening. As a genetic testing tool, it will have applications in various areas, including disease treatment, haplotype analysis, genetic barcoding, evolutionary studies and forensic investigations.

References

1. Kiberstis PA. Mitochondria make a come back. *Science*, **1999**, *283*, 1475.

2. DiMauro S, Moraes CT, Mitochondrial Encephalomyopathies, Archives of Neurology, 1993, 50, 1197-208.

3. Wallace DC. Mitochondrial diseases in man and mouse. *Science*, **1999**, *283*, 1482-1488

4. Wallace DC, Mitochondria as Chi, *Genetics*, **2008**, *179*, 727-735

5. Wong LC Pathogenic mitochondrial DNA mutations in protein-coding genes, *Muscle & Nerve*., **2007**, *36*, 279-93.

6. MITOPMA: a human mitochondrial genome database. http://www.mitomap.org; 2010.

7. Schon EA, DiMauro S, Hirano M, Gilkerson R. Therapeutic prospects for mitochondrial disease. *Trends in Molecular Medicine*, **2010**, *16*, 268-276.

8. Köhnemann S, Pfeiffer H. Application of mtDNA SNP analysis in forensic casework, *Forensic Science International: Genetics*, **2010**, doi: 10.1016/j.fsigen.2010.01.015.

9. Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA: Applications, debates, and Foundations. *Annual Review of Genomics and Human Genetics*, **2003**, *4*, 119-141.

10. Hartmann A, Thieme M, Nanduri LK, Stempfl T, Moehle C, Kivisild T, Oefner PJ. Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Human Mutation*, **2009**, *30*, 115-122.

11. Holt IJ, Harding AE, Petry RK, Morgan-Hughes JA. A new mitochondrial disease associated with mitochondrial DNA heteroplasmy, *The American Journal of Human Genetics*, **1990**, *46*, 428-433.

12. Bai RK, Wong LJ. Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach. *Clinical Chemistry*, **2004**, *50*, 996-1001.

13. Choi BO, Hwang JH, Cho EM, Jeong EH, Hyun YS, Jeon HJ, Seong KM, Cho NS, Chung KW, Mutational analysis of whole mitochondrial DNA in patients with MELAS and MERRF diseases. *Experimental and Molecular Medicine*, **2010**, *42*, 446-455.

14. Tang S, Huang T. Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. *Biotechniques*, **2010**, *48*, 287-296.

15. Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high throughput sequencing of complete human mitochondrial DNA genomes. *The American Journal of Human Genetics*, **2010**, *87*, 237-249.

16. He Y, Wu J, Dressman DC, lacobuzio-Donahue C, Markowitz SD, Velculescu VE, Jr. Diaz LA, Kinzler KW, Volgelstein B, Papadopoulos N. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, **2010**, *464*, 610-614.

17. Stoerker J, Mayo JD, Tetzlaff CN, Sarracino DA, Schwope I, Richert C. Rapid genotyping by MALDI-monitored nuclease selection from probe libraries. *Nature Biotechnology*, **2000**, *18*, 1213-1216.

18. Lyamichev V, Mast AL, Hall JG, Prudent JR, Kaise MW, Takova T, Kwiatkowski RW, Sander TJ, de Arruda M, Arco DA, Neri BP, Brow MA. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probe. *Nature Biotechnology*, **1999**, *17*, 292-296.

19. Tang K, Fu DJ, Julien D, Braun A, Cantor CR, Köster H. Chip-based genotyping by mass spectrometry. *Proceedings of the National Academy of Science of the United States of America*, **1999**, *96*, 10016-10020.

20. Li J, Butler JM, Tan Y, Lin H, Royer S, Ohler L, Shaler TA, Hunter JM, Pollart DJ, Monforte JA, Becker CH. Single nucleotide polymorphism determination using primer extension and

time-of-flight mass spectrometry. *Electrophoresis*, **1999**, *20*, 1258-1265.

21. Ross P, Hall L, Smirnov IP, Haff L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature Biotechnology*, **1998**, *16*, 1347-1351.

22. Xiu-Cheng FA, Garritsen HS, Tarhouny SE, Morris M, Hahn S, Holzgreve W, Zhong XY. A rapid and accurate approach to identify single nucleotide polymorphisms of mitochondrial DNA using MALDI-TOF mass spectrometry, *Clinical Chemistry and Laboratory Medcine,* **2008**, *46,* 299-305.

23. Kim S, Ruparel HD, Gilliam C, Ju J, Digital genotyping using molecular affinity and mass spectrometry. *Nature Reviews Genetics*, **2003**, *4*, 1001-1008.

24. Kim S, Edwards JR, Deng L, Chung W, Ju J. Solid phase capturable dideoxynucleotides for multiplex genotyping using mass spectrometry. *Nucleic Acids Research*, **2002**, *30*, e85.

25. Kim S, Ulz ME, Nguyen T, Li CM, Sato T, Tycko B, Ju J. Thirtyfold multiplex genotyping of the p53 gene using solid phase capturable dideoxynucleotides and mass spectrometry, *Genomics*, **2004**, *83*, 924-931.

26. Misra A, Hong JY, Kim S. Multiplex genotyping of cytochrome p450 single-nucleotide polymorphisms by use of MALDI-TOF mass spectrometry. *Clinical Chemistry*, **2007**, *52*, 933-939.

27. DiMauro S, Hirano M, MERRF, In: Pagon RA, Bird TC, Dolan CR, Stephens K (Eds) GeneReviews [internet]. Seattle (WA): University of Washington, Seattle, 1993-2003.

28. Pallotti F, Baracca A, Hernandez-Rose E, Walker WF, Solaini G, Lenaz G, Melzi D'Eril GV, DiMauro S, Schon EA, Davidson MM. Biochemical analysis of respiratory function in cybrid cell lines harbouring mitochondrial DNA mutations, *Biochemistry Journal*, **2004**, *384*, 287-293.

29. Masucci JP, Davidson M, Koga Y, Schon EA, King MP. In vitro analysis of mutations causing myoclonus epilepsy with ragged-red fibers in the mitochondrial tRNA(Lys)gene: two genotypes produce similar phenotypes. *Molecular and cellular biology*, **1995**, *15(5)*, 2872-2881.

30. Silvestri G, Moraes CT, Shanske S, Oh SJ, DiMauro S. A new mtDNA mutation in the tRNA(Lys) gene associated with myoclonic epilepsy and ragged-red fibers (MERRF). *The American Journal of Human Genetics*, **1992**, *51(6)*, 1213-1217.

31. Santorelli FM, Mak SC, El-Schahawi M, Casali C, Shanske S, Baram TZ, Madrid RE, DiMauro S. Maternally inherited cardiomyopathy and hearing loss associated with a novel mutation in the

mitochondrial tRNA(Lys) gene (G8363A). *The American Journal of Human Genetics*, **1996**, *58 (5)*, 933-93

32. Fabrizi GM, Cardaioli E, Grieco GS, Cavallaro T, Malandrini A, Manneschi L, Dotti MT, Federico A, Guazzi G. The A to G transition at nt 3243 of the mitochondrial tRNALeu(UUR) may cause an MERRF syndrome. *Journal of Neurology, Neurosurgery & Psychiatry,* **1996**, *61*, 47-51.

33. Nishigaki Y, Tadesse S, Bonilla E, Shungu D, Hersh S, Keats BJ, Berlin CI, Goldberg MF, Vockley J, DiMauro S, Hirano M. A novel mitochondrial tRNA(Leu(UUR)) mutation in a patient with features of MERRF and Kearns-Sayre syndrome. *Neuromuscular Disorders*, **2003**, *13(4)*, 334-340.

34. Choi BO, Hwang JH, Cho EM, Jeong EH, Hyun YS, Jeon HJ, Seong KM, Cho NS, Chung KW, Mutational analysis of whole mitochondrial DNA in patients with MELAS and MERRF diseases, *Experimental and Molecular Medicine*, **2010**, *42*, 446-455.

35. Campos Y, Martin MA, Lorenzo G, Aparicio M, Cabello A, Arenas J, Sporadic MERRF/MELAS overlap syndrome associated with the 3243 tRNA(Leu(UUR)) mutation of mitochondrial DNA, *Muscle & Nerves*, **1996**, *19*, 187-190.

36. King MP, Attardi G. Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. *Science*, **1989**, *246*, 500-503.

37. Rossignol R, Faustin B, Rocher C, Malgat M, Mazat JP, Letellier T, Mitochondrial threshold effects, *Biochemistry Journal*, **2003**, *370*, 751-762.

38. Sacconi S, Salviati L, Nishigaki Y, Walker WF, Hernandez-Rosa E, Trevisson E, Delplace S, Desnuelle C, Shanske S, Hirano M, Schon EA, Bonilla E, De Vivo DC, DiMauro S, Davidson MM, A functionally dominant mitochondrial DNA mutation, *Human Molecular Genetics*, **2008**, *17*, 1814-1820.

Chapter 5 Exploration of the Integrated Microdevice for SNP Genotyping by MALDI-TOF Mass Spectrometry

5.1 Introduction

Single nucleotide polymorphisms (SNPs) are of significant interest in the life sciences, as they provide the genetic fingerprint of an individual as well as potential biomarkers for disease diagnosis and predisposition. The detection of SNPs, SNP genotyping, has become one of the most powerful and indispensable tools in genomic analysis. A plethora of methods have been applied for SNP genotyping using different detection platforms, among which mass spectrometry-based single base extension has been shown to be a promising method due to its utilization of the intrinsic property of molecular mass for nucleotide discrimination, ease of data acquisition, capability for quantitation, improved accuracy and multiplex capacity. The MS-based SBE genotyping strategy was further improved by our cleavable solid phase capture (SPC) single-base-extension (SBE) approach with the use of cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins), as described in Chapter 4, allowing the rapid, accurate, sensitive and quantitative detection of SNPs. However, the use of manual benchtop procedures limits its applications, especially for large scale sample handling in a time efficient manner. True implemention of our reversible SPC-SBE SNP genotyping method requires a state-of-the-art integrated platform capable of high-speed, high-throughput, and automatic detection at low cost.

Lab-on-a-chip (LOC) technologies for the development of biological/chemical analysis tools hold great promise. Though not developed until the 1990s, lab-on-a-chip technology soon revolutionized bio-analytical protocols since this miniaturization system offers great advantages in terms of integration, speed, efficiency and portability. Moreover, the miniaturized microfluidic device allows the use of small sample volumes, and hence lower reagent consumption and energy requirements.[1, 2]

Genetic analysis microfluidic systems have been extensively explored to integrate DNA analysis steps, including DNA extraction,[3, 4] PCR amplification[1, 5] and capillary electrophoretic (CE) separation[6] at the microliter to nanoliter scale. Various types of genotyping microdevices based on different mechanisms have been reported, including restriction fragment analysis, DNA sequencing, hybridization, allele-specific amplification or heterduplex analysis.[7] For example, Choi *et al.* have developed an integrated microdevice for short tandem repeat analysis combined with capillary electrophoresis.[8] However, to our knowledge, no such device has been constructed for a SPC-SBE genotyping approach with mass spectrometic detection. This is mostly because in the previous reported SPC-SBE method,[9] the release of extension products required harsh treatment (formamide) and ethanol precipitation, which are hard to integrate into a microdevice. The introduction of cleavable nucleotide analogs in our SNP genotyping approach addresses these problems, which, together with recently developed lab-on-a-chip techniques, makes it promising to construct a fully integrated SNP genotyping microfluidic platform.

In this chapter, a proof-of-concept microfluidic device for SNP genotyping using cleavable biotinylated nucleotides is presented, thanks to the collaboration with Dr. ThaiHuu Nguyen and Jing Zhu in Dr. Qiao Lin's microfabrication laboratory. This device primarily consists of a micro-reaction chamber for single base extension and cleavage reactions with an integrated micro heater and temperature sensor for on-chip temperature control, a microchannel loaded with streptavidin magnetic beads for solid phase capture, and a microchannel packed with C18-modified reversed-phase silica particles as a stationary phase for desalting before MALDI-TOF analysis. We have evaluated the microdevice by performing each functional step, including single base extension, solid phase capture and cleavage, and desalting, the results of which demonstrate the feasibility of developing a fully integrated cleavable SPC-SBE based microfluidic SNP genotyping device.

5.2 Experiment Rationale and Overview

As shown in Fig. 5.1, our typical cleavable SPC-SBE genotyping approach can be divided into four major steps, where step 1 and step 3 require temperature control, while step 2 and step 4 are essentially solid phase extraction and purification.

Single base extension with non-reversible terminators is similar to PCR amplification, except that it is a unidirectional linear amplification reaction with the primer growing by only one base (though more bases can be added in a random fashion when both dNTPs and ddNTPs are included). The cleavage step simply requires

incubation at 65°C. These two steps can be carried out in a single chamber, with similar design to a PCR microdevice. As shown in Fig 5.2, the heater was designed in the form of a closed circle to provide adequate heating for the entire chamber.

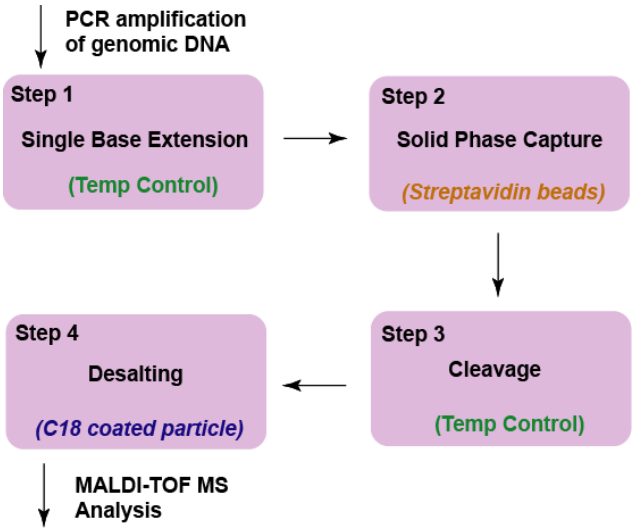


Fig. 5.1. Overall scheme of cleavable SPC-SBE approach for SNP genotyping.

The mobility of magnetic beads under the magnet provides the flexibility in the design. As shown in Fig 5.2, streptavidin magnetic beads will be placed in a loading channel and moved into the central chamber using the magnet when required. The solid phase capture is designed to take place inside the central chamber, followed by bead washing and the temperature controlled cleavage reaction.

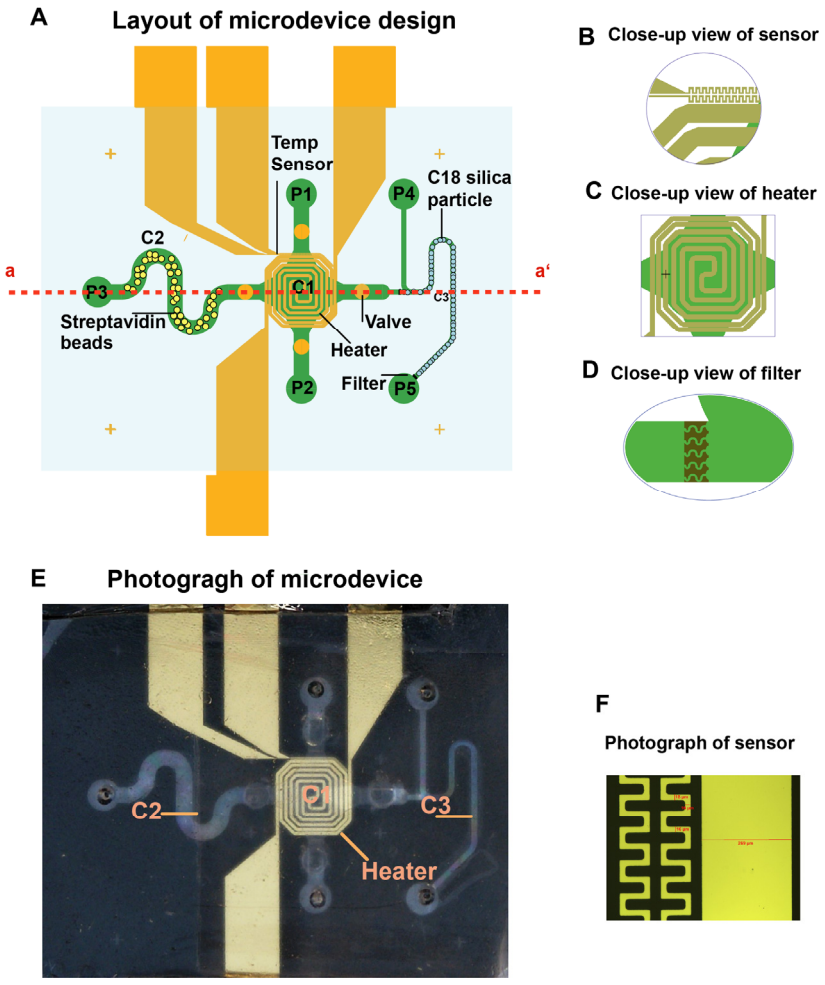


Fig. 5.2. (A) Schematic of the microfluidic SNP genotyping device. C1 is the central chamber for the single base extension and cleavage reaction, C2 is the streptavidin bead loading channel, and C3 is the C18 reverse-phase channel. P1 to P5 represent the outlet/inlet ports for sample loading and waste collection. The light blue rectangle indicates the PDMS layer. (B) A close-up view of the temperature sensor; (C) A close-up view of the central heater; (D) A close-up view of a filter; (E) Photograph of microdevice. (F) A light microscopic blow up view of the temperature sensor.

Our previous experience with the desalting step is to utilize the commercially available Millipore ZipTips, in which the pipet tips are loaded with octadecyl carbon chain coated (C18) silica particles as the micro-column reverse-phase to facilitate the sample recovery.[10, 11] The basic principle is the affinity of DNA molecules for these C18

coated particles. When DNA molecules flow through the C18 reversed-phase channel, DNA molecules will be adsorbed onto the C18 surface whereas the salts will not, and the DNA molecule can then be eluted from the C18 surface using an organic solvent (e.g., 50% acetonitrile). Based on this principle, we designed a C18 silica particle packed channel for desalting, with filters in between to prevent particle loss (Fig 5.2. A, D).

To fully implement the entire process and enable the controlled continuous microflow, each design component needs to be connected in turn to an on/off switch . The most straightforward way to accomplish this is to employ a valving system. Due to the ease of fabrication and amenability for miniaturization, elastomeric polydimethylsiloxane (PDMS) has been exploited as a microvalve material with different actuation mechanisms.[12-14] The principle is that the thin PDMS membrane will deform under pressure, hence press down into the microfluidic channel to modulate fluid flow. It has been reported that filling the PDMS control channel with air could act upon the underlying fluidic channel to form such a valve.[14] In a similar strategy, we designed our PDMS valves by air deflection of a thin PDMS film on a glass substrate. Small holes were drilled in the glass substrate corresponding to the valve position and connected to air supply tubing, allowing these sections of the thin PDMS film on the glass substrate to deform and thereby close the valve.

In summary, as shown in Fig. 5.2, our SNP genotyping microdevice is composed of a central reaction chamber for single base extension and cleavage (C1), a streptavidin bead loading channel (C2) and a C18 reverse-phase desalting channel (C3), with 5 inlet/outlet

ports for sample injection and waste collection. Around the central chamber, four PDMS film based pressure valves were designed to control the microfluidic flow, and the filters were set at the two ends of the C18 microchannel for maintaining the C18 stationary phase. In this study, as the proof of concept, the microdevice was evaluated by performing separate experiments for each step shown in Fig. 5.1. Briefly, single base extension (SBE) was first carried out in the central reaction chamber, followed by solid phase capture of SBE products in the same chamber via the incubation with streptavidin beads moving from the bead loading channel. Next, the cleavage reaction was performed in the central chamber, and finally cleavage products were flowed through the C18 reversed-phase microchannel for desalting.

5.3 Results and Discussion

5.3.1 Temperature Sensor Calibration

In our microdevice system, the resistance temperature detector (RTD) configuration was used. The relationship between temperature and resistance of a metal is described in the Callendar-Van Dusen equation (Equation 5.1),[15] wherein the resistance at temperature T (R(T)) is the sum of the resistance at 0^oC (R_0) , A R_0T and BR_0T^2, with A and B coefficients characteristic of the material.

$$R(T) = R_0(1 + AT + BT^2) \quad \text{Equation 5.1}$$

By replacing the resistance at 0^oC with the reference temperature (e.g., room temperature), the temperature coefficient of resistance (TCR) is given by equation 5.2.

$$TCR = \frac{1}{R}\frac{(R - R_0)}{(T - T_0)}$$ Equation 5.2

Therefore, the temperature coefficient of resistance (TCR) value can be determined by calculation of a temperature sensor calibration curve with temperature on the x-axis and resistance on the y-axis, as shown in Fig 5.3. The TCR value of $2.4\times10^{-3}/^{o}C$ was obtained by calculation of the slope after linear regression fitting.

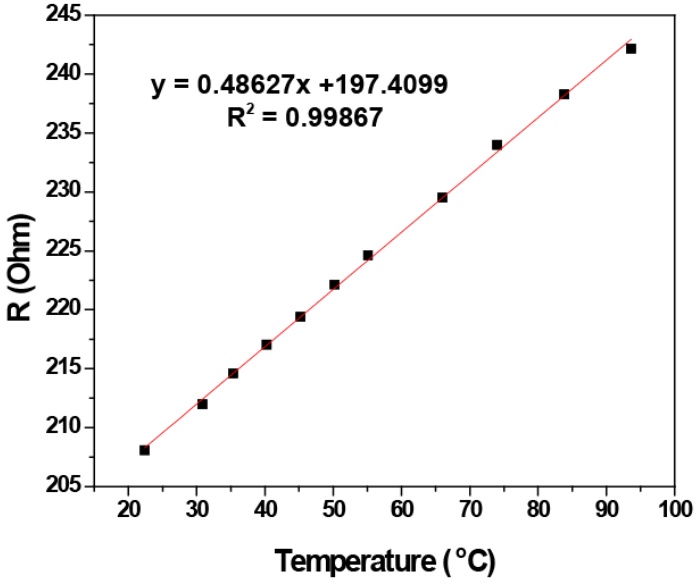


Fig. 5.3. Temperature sensor calibration. The formula shown is the linear regression fitting result, with y referring to resistance (Ohms) and x referring to temperature (^{o}C).

5.3.2 On-chip testing

To verify the feasibility of implementing our cleavable SPC-SBE approach in this microfluidic platform, the microdevice was evaluated by performing individual SNP genotyping steps separately, with the hope of thoroughly understanding the characteristics of each part before full integration.

5.3.2.1 On-chip single base extension

SBE was first carried out to evaluate the feasibility of performing single base extension in our microdevice, with a standard dideoxynucleotide, ddTTP, as shown in Fig. 5.4 A. 100% incorporation was confirmed with MALDI-TOF mass spectrometry by observing the total disappearance of the primer peak (5163 m/z) and the emergence of the ddTTP extension product (5453 m/z) (Fig. 5.4 B). This demonstrated heat transfer capability and the accuracy of temperature control of our microdevice, which confirms the feasibility of on-chip SBE reaction. Though the experiment was performed under the same thermo-cycling conditions as in a conventional thermal cycler, the number of cycles and the time for each step (denaturation, annealing, extension) might be further reduced, because the large surface-to-volume ratio in the microdevice will facilitate the heat transfer and temperature equilibration, enabling swift thermal responsiveness of the sample to the controller during the SBE process. .

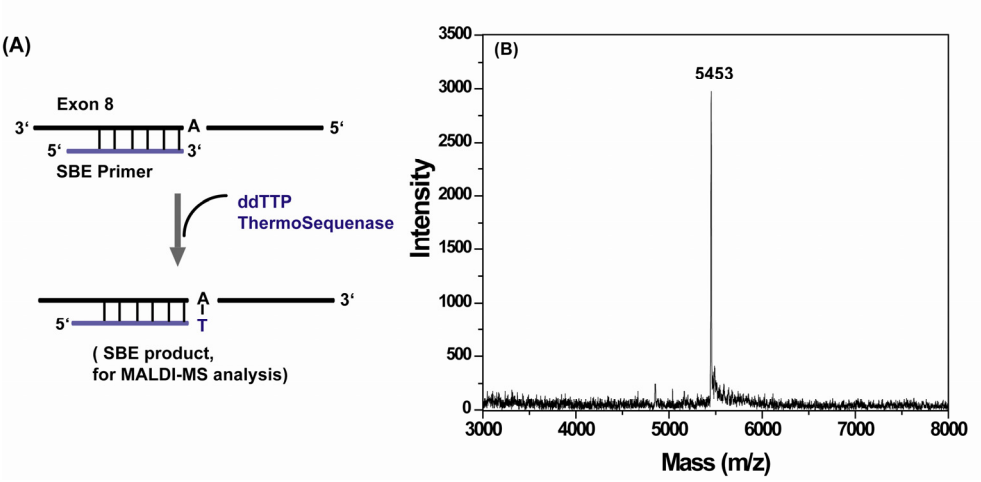


Fig. 5.4. Results of on-chip single base extension reaction. (A) Scheme of SBE reaction with ddTTP; (B) MALDI-TOF mass spectrum of the SBE extension products.

5.3.2.2 On-chip solid phase capture

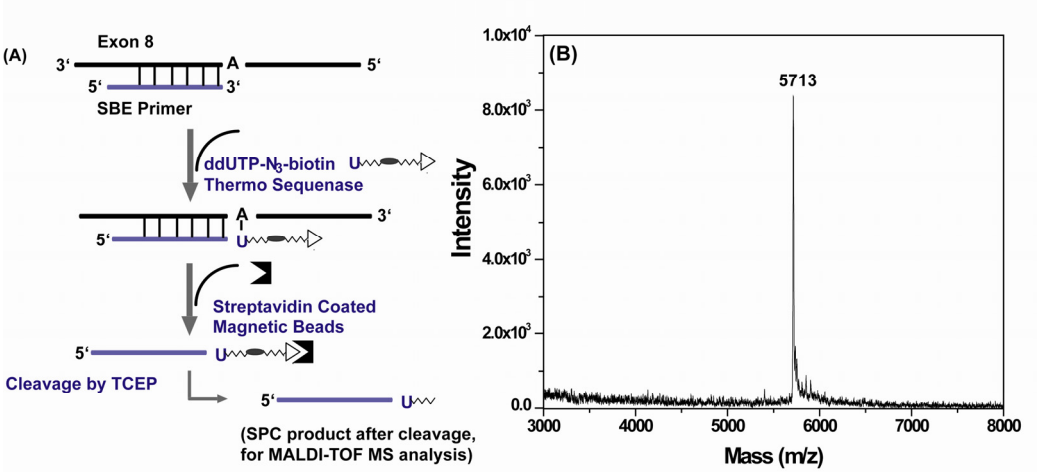


Fig. 5.5. On-chip solid phase capture and cleavage result. (A) Scheme of the whole process including SBE, SPC and cleavage; (B) MALDI-TOF mass spectra of the products.

To evaluate the ability to perform on-chip solid phase capture and cleavage, the SPC and cleavage experiments were carried out on a single ddUTP-N_3-biotin extended product. As shown in Fig. 5.5 A, biotin terminated SBE products with a molecular weight at 6186 m/z were generated by SBE reaction using ddUTP-N_3-biotin in a thermal cycler using a standard protocol. This SBE product was then incubated with streptavidin beads for extracting extended products. The magnet was able to freely move the magnetic beads into the chamber for mixing the beads and SBE solution, which facilitated effective functional surface interaction with DNA molecules. The cleavage reaction by TCEP released the extension products from the solid bead surface while leaving the biotin moiety on the surface, the molecular weight of which was 5713 m/z. The observation of a mass peak at 5713 m/z (Fig. 5.5 B) indicated that the magnetic beads loaded into the channel were capable of capturing enough products and releasing the

cleavage products, which further verified the feasibility of on-chip SPC and cleavage.

5.3.2.3 On-chip desalting

The C18 reversed phase channel is very important for MS-based SPC-SBE analysis; its purpose is to remove of salts from the cleavage solution while maintaining enough pure sample for mass spectrometric analysis. Theoretically, the capture efficiency and sample recovery rate increases with an increasing number of C18 particles, yet the back pressure will build up beyond the operational range due to tight packing. On the other hand, while the particle size is inversely related to the dynamic capacity, in order to reduce the back pressure, the particle size needs to be relatively large.[10] Therefore, relatively larger size C18 silica particles (40 to 75 μm) were first chosen for the preliminary test to see if it they could achieve the required sensitivity for mass spectrometry. The portion of the channel packed with C18 particles is shown in Fig. 5.6 A. Different amounts of DNA in TCEP solution were tested to evaluate desalting efficiency. The results indicated that the microchannel was able to recover 0.5 pmol, 1pmol, 10 pmol and 100 pmol DNA sample and gave clean spectra without evidence of salt disturbance (the result with the 0.5 pmol DNA/TCEP solution is shown in Fig 5.6. B). The fact that as little as 0.5 pmol of DNA sample was able to be recovered from the C18 channel and generate a strong signal intensity in mass spectrometric analysis demonstrated its promise in detecting low levels of mutation, as this is much lower that the amount typically used for SBE (10-20 pmol).

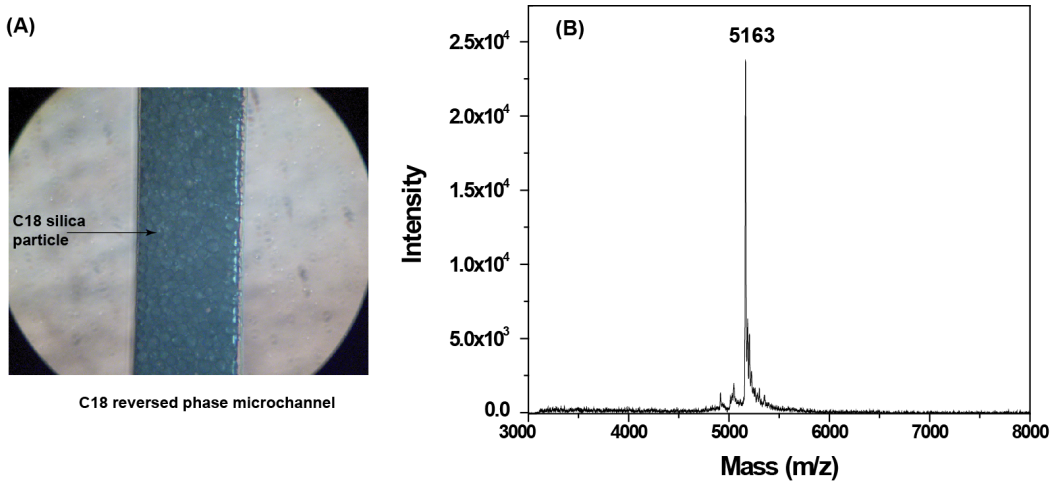


Fig. 5.6. (A) The microchannel packed with C18 coated particles. (B) Mass spectrum after on-chip desalting of 0.5 pmol DNA/TCEP solution.

5.4 Materials and Methods

General information. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. The cleavable biotinylated dideoxynucleotides were synthesized in our laboratory. Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). Thermo Sequenase was from GE Healthcare (Piscataway, NJ). Streptavidin coated magnetic beads (Dynabeads® MyOne™ Streptavidin C1) were obtained from Life Technologies (San Diego, CA). Photo resist S1818, SU-8 2150, and SU-8 2010 were from Microchem (Newton, MA), and *Sylgard* 184 poly(dimethylsiloxane) (PDMS) was from Dow Corning. Microscope grade glass slides (2" x 3") were from Fisher Scientific. C18 particles, C18 Spherical Silica Gel with size distribution from 40 μm to 75 μm, were from Sorbent Technologies (Norcross, GA). Primary instruments used for photolithography were a Süss MicroTec MJB8 Mask Aligner, Edwards BOC/Auto 306 Thermal Evaporator, and a Diener Plasma Etch System.

Device temperature control was performed using integrated resistive heating and sensing elements connected to a DC power supply (Agilent E3631A), a digital multimeter (Agilent 34410A), and a proportional-integral-derivative (PID) controlled LabVIEW program (National Instruments).

5.4.1 Microfluidic Device fabrication

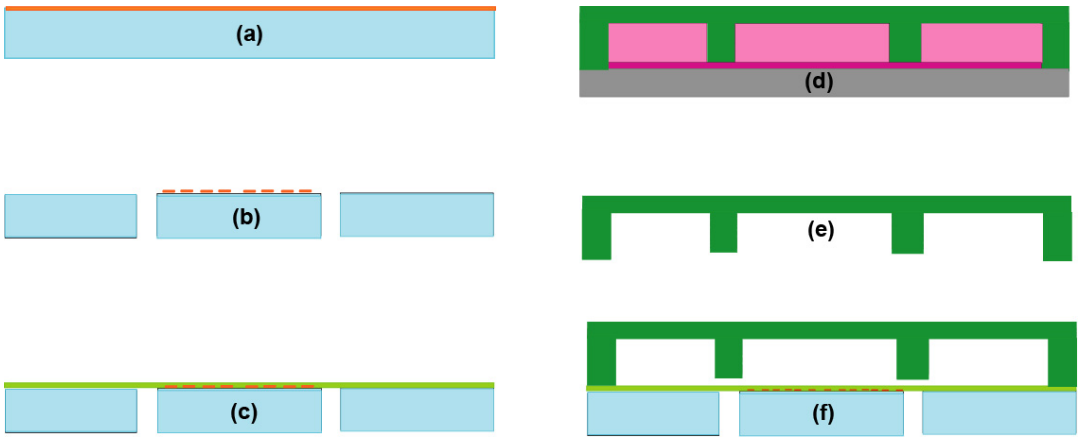


Fig. 5.7. A simplified device process flow referring to cross section a-a' in Fig. 5.2. (a) Metal Cr/Au deposition on glass slide. (b) Lithography for metal patterning and hole drilling in the slide. (c) Thin PDMS layer bonding. (e) PDMS channel fabrication. (f) PDMS channel demolding. (f) Bonding and packaging.

The temperature control chip was fabricated by standard semiconductor fabrication techniques. The glass substrate were first diced to the size of 2x2 cm and cleaned by immersing in Nano Strip (MicroChem) overnight. Subsequently, 5 nm chromium and 100 nm gold films were deposited via thermal evaporation (Fig 5.7 a) and then patterned by photolithography and wet etch. The glass slide was drilled with four holes

corresponding to the four pressure valve locations (Fig 5.7 b). The metal films, including heater and sensor, were then passivated with a thin PDMS film (~100 nm), which also serves as the valve under the air actuation (Fig 5. 7 c).

The PDMS microfluidic chamber was fabricated using soft lithography techniques. Negative photoresist SU-8 2050 was spin-coated onto a 4 inch silicon wafer to create a mold master of 50 μm thickness. After a soft bake, the pattern of the mask was transferred to the SU-8 coated silicon wafer by mask aligner (MA-6, Karl Suss GmbH, Germany) followed by hard-baking and development. Then SU-8 2150 was spin-coated onto the same wafer to create 375 μm thick patterns/features through similar procedures. As depicted in Fig.5.7.d, a PDMS pre-polymer solution was mixed at a volume ratio of 10:1, distributed on the silicon mold and degassed by vacuum for ~30 min, followed by curing at 65^{o}C for 40 min, to generate the PDMS sheet with defined channels shown in Fig.5.10 e. Finally, the PDMS layer was removed from the wafer mold and aligned and bonded to a thin passivation film covered temperature control chip (Fig.5.7 c) following O_2 plasma treatment of the bonding interface for 15s.

Following the above procedures, the complete device shown in Fig. 5.2.E was fabricated. The volume of the central microchamber is around 12 μl (6 mm in diameter and 425 μm in height), and volumes of the streptavidin bead loading channel and C18 reversed-phase channel are around 15 and 5.5 μl respectively. An integrated Cr/Au resistive heater and temperature sensor were used in conjunction with off-chip programming (Labview) to control the temperature.

The PDMS microfluidic chamber has its advantages in the ease of fabrication, low cost and robustness, however, it will also cause some critical problems, including sample evaporation and bubble formation during heating, as well as non-specific adsorption. To address these problems, parylene coating was used as it was reported to have lower permeability to moisture and long-term stability.[16] The coating was achieved by depositing Parylene C Dimer (di-chloro-di-para-xylylene) into the channels by chemical vapor deposition using a PDS 2010 LABCOTER deposition system.

For the complete microdevice package, the streptavidin magnetic beads were washed with and suspended in 2X binding buffer, and directly loaded into the channel with the magnet on the bottom holding the beads. The C18 particles were washed with methanol three times, loaded into the channel and dried by heating at 80^{o}C overnight.

5.4.2 Experimental setup

An illustration of the experimental setup is presented in Fig. 5.8. The integrated heater and temperature sensor are connected to off-chip PC programming for temperature control. The sample loading is controlled by the syringe pump at a stable fluid rate. The fluid flow through the microchannel is controlled by the four PDMS pressure valves in the microdevice, which are connected to the valving system (Fig. 5.8.C).

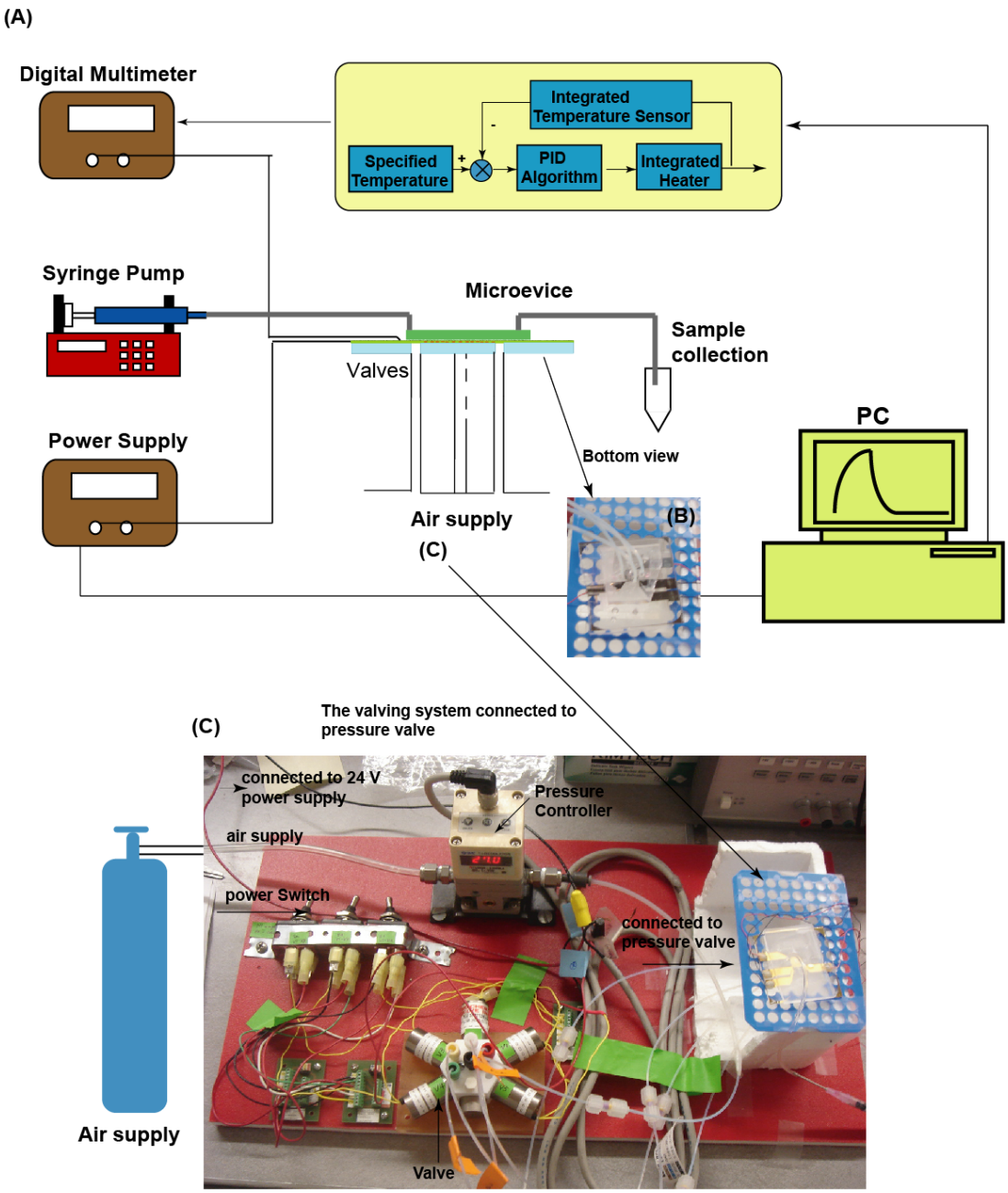


Fig. 5.8. (A) Microfluidic experimental test setup with closed-loop feedback temperature control. (B) Bottom view of microdevice; (C) Valving system set-up.

As referred to in Fig. 5.8 C, with the help of Mirkó Palla in our laboratory, the complete valving system was constructed as follows. Four tubes were inserted into the holes of the glass substrates (Fig. 5.8 B), each connected to one port of a solenoid valve,

the outlet of which was connected to a gas supply with a pressure controller. The solenoid valve was controlled by three mechanical power switches, allowing selective actuation of subsets of the PDMS valves. Theoretically, the operation of the valve is dependent on the thickness of the PDMS layer, the height of the microchannel and the air pressure.

5.4.3 Microdevice performance testing

5.4.3.1 Temperature Sensor Calibration

The temperature coefficient of resistance (TCR) value for the microdevice was obtained by calibration using the temperature sensor. The device was place in an oven and the temperature was allowed to increase in approximately 5^{o}C intervals from room temperature to 95^{o}C. The steady state temperature and the resistance of the sensor were recorded to generate the temperature sensor calibration curve, and the TCR value was determined by calculation of the slope of the curve after linear regression fitting.

5.4.3.2 Single base extension

In the single base extension test, dideoxynucleotides instead of biotinylated nucleotides analogs were used. The magnetic beads and C18 particle loading channels were sealed by PDMS. 20 pmol of synthetic DNA template, Exon8 of the p53 gene, (5'-GAAGGAGACACGCGGCCAGAGAGGGTCCTGTCCGTGTTTGTGCGTGGAG TTCGACAAGGCAGGGTCATCTAA**TGGTGATGAGTCCTATC**CTTTTCTCTTCGT TCTCCGT-3', the letters in bold indicating the primer annealing site), 40 pmol of primer

(5'-GATAGGACTCATCACCA-3'), 60 pmol of ddTTP, 1X Thermo Sequenase reaction buffer and 2 unit of Thermo Sequenase were degassed in vacuum before being gently injected into the reaction chamber. The use of excess reagent was used to fill in the inlet and outlet channels to avoid generation of air bubbles, and two drops of mineral oil were applied to the outlets and inlets to prevent evaporation. The on-chip reaction was performed in the central reaction chamber for 20 cycles of 90^{o}C for 15 s, 40^{o}C for 1 min, and 70^{o}C for 30s, and the products were removed from the reaction chamber and analyzed using MALDI-TOF MS.

5.4.3.3 Solid phase capture and cleavage test

To mimic solid phase capture, SBE reaction products from a thermal cycler were used. 10 μl of reaction mixture consisting of 10 pmol template, 20 pmol ddTTP, 30 pmol ddUTP-N_3-biotin, 1X Thermo Sequenase reaction buffer and 2 U Thermo Sequenase were carried out under the following condition: 20 cycles of 90^{o}C for 15 s, 40^{o}C for 1 min, and 70^{o}C for 30 s. Streptavidin magnetic beads were preloaded into the bead loading channel (C2). After injecting the post SBE reaction mixture into the central chamber (C1), the magnetic beads were brought into the central chamber by movement of the magnet incubation allowed to proceed for around 30 min. Then beads were washed with H_2O and treated in TCEP solution at 65^{o}C for 15 min. The cleavage products were collected and analyzed by MALDI-TOF MS.

5.4.3.4 C18 reversed phase desalting

To test the desalting capability of the C18 reversed phase microchannel on the microdevice, 0.03 g of C18 particles were prewashed with methanol 3 times and resuspended in methanol for injection. The packed C18 particles were dried by heating at 80^{o}C overnight. To mimic the cleavage products, the oligonucleotides (17mer, the primer for the above SBE reactions) were dissolved in 20 μl TCEP solution to generate a series of diluted solutions, using 1 pmol, 10 pmol and 100 pmol oligonucleotides. The C18 channel was first washed with 50% acetonitrile (ACN) at a flow rate of 20 μl/min for 5 min, followed by 0.1 M Triethylammonium acetate buffer, pH 5.5 (TEAA) washing for 10 min. Then the "cleavage products" were slowly flowed through the C18 channel at 0.5 μl/min. The washing steps were performed consecutively with 0.1M TEAA and dH_2O at a flow rate of 50 μl/min. Then the purified "cleavage products" were slowly eluted from the C18 channel with 50% ACN at a flow rate of 1 μl/min. The final products were collected and analyzed by MALDI-TOF MS.

5.5 Conclusion

Following our successful demonstration of cleavable solid phase capture-single base extension SNP genotyping approach in detecting mitochondrial SNP at high accuracy and high sensitivity, we further explored the use of a SNP genotyping microfluidic lab-on-a-chip device with the potential for high throughput, miniaturization and automation. As a proof of concept, we have designed, constructed and evaluated our

prototype SNP genotyping microdevice by performing individual functional steps, which could be later integrated in the mature microdevice. We have demonstrated 100% single base incorporation on-chip, sufficient capture and release of the biotin terminated single base extension products, and high sample recovery from the C18 reverse-phase microchannel with as little as 0.5 pmol DNA molecules. This demonstrated the feasibility of the microdevice toward the goal of an integrated, miniaturized, and high throughput platform for SNP genoyping, and also for potential mass spectrometry-based sequencing by synthesis.

References

1. Zhang Y, Ozdemir P. Microfluidic DNA amplification-a review. *Analytica Chimica Acta*, **2009**, *638*, 115-125.

2. Beyor N, Yi L, Seo TS, Mathies RA. Integrated capture, concentration, polymerase chain reaction, and capillary electrophoretic analysis of pathogens on a chip. *Analytical Chemistry*, **2009**, *81*, 3523-3528.

3. Tian H, Huhmer AF, Landers JP. Evaluation of silia resins for direct and efficient extraction of DNA from complex biological matrices in a miniaturized format. *Analytical Biochemistry*, **2000**, *283*, 175-191.

4. Bienvenue JM, Duncalf N, Marchiarullo D, Ferrance JP, Landers JP. Microchip-based cell lysis and DNA extraction from sperm cells for application to forensic analysis. *Forensic Science*, **2006**, *51*, 266-273.

5. Lagally ET, Medintz I, Mathies RA. Single-molecule DNA amplification and analysis in an integrated microfluidic device. *Analytical Chemistry*, **2001**, *73*, 565-570.

6. Legally ET, Emirich CA, Mathies RA. Fully integrated PCR-capillary electrophoresis microsystem for DNA analysis. *Lab chip*, **2001**, *1*, 102-107.

7. Szántai E, Guttman A. Genotyping with microfluidic devices. *Electrophoresis*, **2006**, *27*,

4896-4903.

8. Choi JY, Seo TS. An integrated microdevice for high-performance short tandem repeat genotyping. *Biotechnology Journal*, **2009**, *4*, 1530-1541.

9. Kim S, Ulz ME, Nguyen T, Li CM, Sato T, Tycko B, Ju J. Thirtyfold multiplex genotyping of the p53 gene using solid phase capturable dideoxynucleotides and mass spectrometry, *Genomics*, **2004**, *83*, 924-931.

10. Pluskal MG. Microscale sample preparation. *Nature Biotechnology*, **2000**, *18*, 104-105.

11. Gaspar A, Salgado M, Stevens S, Gomez FA. Microfluidic "thin chips" for chemical separations. *Electrophoresis*, **2010**, *31*, 2520-2525.

12. Chen CF, Liu J, Chang C-C, DeVoe DL. High-pressure on-chip mechanical valves for thermoplastic microfluidic devices. *Lab Chip*, **2009**, *9*, 3511-3516.

13. Gaspar A, Piyasena ME, Daroczi L, Gomez FA. Magnetically controlled valve for flow manipulation in polymer microfluidic devices. *Microfluid Nanofluid*, **2008**, *4*, 525-531.

14. Unger MA, Chou H, Thorsen T, Scherer A, Quake SR. Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science*, **2000**, *288*, 113-116.

15. Quach Q. A parylene real time PCR microdevice. Ph.D. dissertation, California Institute of Technology, **2010**.

16. Shin YS, Cho K, Lim SH, Chung S, Park S-J, Chuang C, Han D-C, Chang JK. PDMS-based micro PCR chip with Parylene coating. *Journal of Micromechanics and Microengineering*, **2003**, *13*, 768-774.

Part II Strategies to Improve Sequencing by Synthesis with Cleavable Fluorescent Nucleotide Reversible Terminators

DNA sequencing plays a tremendously important role in biology. DNA sequencing by synthesis, the dominant method in second-generation sequencing platforms, and being explored as well for third generation sequencing, has emerged as one of the most important sequencing technologies. Nevertheless, to meet the goal of the $1000 Genome, the current short sequencing read-length is still a bottleneck. This section of the thesis will describe some strategies to improve the sequencing by synthesis from two perspectives, extending the sequence read-length by primer “walking”, and increasing throughput as well as overall coverage with the emulsion-bead-on-chip approach.

Chapter 6 Development of Primer "Walking" Strategy to Increase the Read-Length of Sequencing by Synthesis

6.1 Introduction

DNA sequencing by synthesis (SBS), with its enormously increased throughput relative to conventional Sanger sequencing, has gained tremendous interest for achieving the $1000 human genome goal. In particular, sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators (CF-NRTs) has offered the most promise, with its higher accuracy and lower fluorescent background as compared to other SBS technologies. Over a number of years, investigations in the Ju laboratory have built a collection of complete sets of nucleotide analogs for successful sequencing by synthesis.

The key to this development is the use of four reversible nucleotide terminators, one for each of the bases of DNA, such that in each cycle of SBS, (1) a polymerase reaction should add with precision only the next base thanks to the presence of a 3' OH blocking group on the nucleotides which prevents further extension, (2) detection of the incorporated base, and (3) conversion of the blocked (protected) nucleotide back to a normal nucleotide to allow further rounds of nucleotide incorporation. Ideally, each of the 4 nucleotide analogues should be tagged with distinguishable characteristics, exhibit efficient incorporation by DNA polymerases and efficient deprotection under mild conditions.[1] Earlier work in the literature exploring the SBS method was mostly focused on a cleavable chemical moiety linked to a fluorophore to cap the 3'-OH group of the

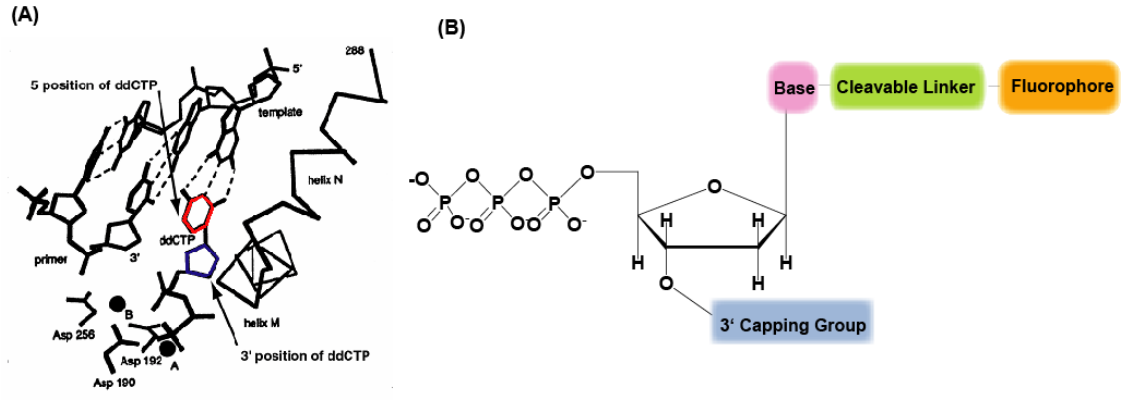


Fig. 6.1. Schematic illustration of rationales for cleavable fluorescent nucleotide reversible terminators (CF-NRTs). (A) Stereo diagram of the polymerase active site for incorporating an incoming ddCTP;[4] The ring in red refers to the base (cytosine), and the ring in blue refers to the sugar. The 5 position of the base and the 3' position of the sugar are indicated by arrows. (B) General structure of our design of CF-NRTs.

nucleotides, with the hope that 3'-OH would be regenerated to allow subsequent nucleotide addition after the fluorophore was removed.[2, 3] However, no success had been reported. As we look into the polymerase-DNA-nucleotide complex[4] (Fig. 6.1), it is apparent that the 3' position on the deoxyribose is very close to the amino acid residues in the active site of the polymerase so that any large modification would tend to inhibit polymerase function, while the 5 position of the nucleotide has a reasonable amount of space for introducing big molecules. These findings are also consistent with the report that some modified DNA polymerases are highly tolerant to nucleotides with extensive modifications with bulky groups at the 5-position of the pyrimidines (T and C) and the 7-position of the purines (G and A).[5, 6] Based on this rationale, our group first proposed the design and synthesis of cleavable fluorescent nucleotide terminators produced by

tethering a unique fluorescent dye to the 5-position of the pyrimidines (T and C) and the 7-position of the purines (G and A) via a cleavable linker, and attaching a small cleavable chemical moiety to cap the 3'-OH group (Fig. 6.1 B).[7]

Since this discovery, our lab has explored many sets of cleavable fluorescent nucleotide reversible terminators (CF-NRTs) and nucleotide reversible terminators (NRTs) or their analogues for sequencing by synthesis, including 2-nitrobenzyl linker based photocleavable fluorescent nucleotide analogues,[8] allyl linker based nucleotide analogues[9, 10] and azido based nucleotide analogues,[11, 12] with azido based nucleotide analogues (3'-*O*-N_3-dNTPs and 3'-*O*-N_3-dNTP-N_3-fluorophores) having shown the most promise so far. For NRTs, the 3' OH of the dNTPs was capped with the azidomethyl group (Fig. 6.2), while for CF-NRTs, each nucleotide was attached to a unique fluorophore via an azido linker and capped at the 3' position with the azidomethyl group (Fig. 6.3). The advantages of using an azido based linker lies in its mild cleavage conditions: the azido group can be efficiently converted into an amine group by tris(2-carboxyethyl) phosphine (TCEP), an odorless and stable agent often used to digest peptide disulfide bonds, and finally regenerate the OH group at the 3' position after hydrolysis. The process of sequencing by synthesis using this set of nucleotide analogs is described in Fig. 6.4. In brief, starting with a cluster of self-priming DNA templates immobilized on a solid surface, the incorporation step is begun with a mixture of 3'-*O*-N_3-dNTPs and 3'-*O*-N_3-dNTP-N_3-fluorophores. A small portion of the DNA molecules that incorporate CF-NRTs will give a sufficient fluorescent signal for base

identification, while the greater portion of the DNA templates incorporates NRTs. To synchronize all the templates, a capping step with NRTs is subsequently performed, followed by a cleavage step. Both the 3'-capping group (azido) and the fluorophores are cleaved away by TCEP to recover the 3'-OH group for the next round of signal generation. In this manner, as the sequencing cycle progresses, the bases will be identified sequentially one by one. Using this set of CF-NRTs/NRTs, our lab has been able to perform accurate DNA sequencing by synthesis.[12]

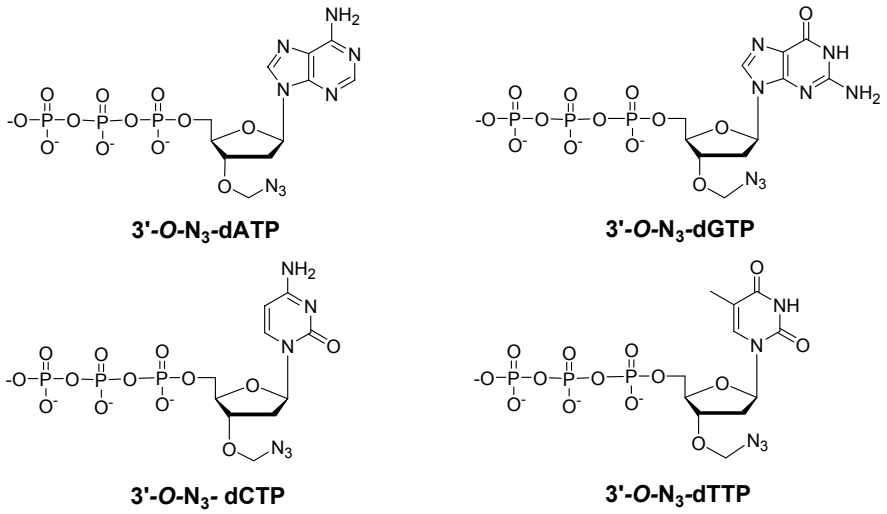


Fig. 6.2. Molecular structures of 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP, 3'-*O*-N_3-dCTP and 3'-*O*-N_3-dTTP.

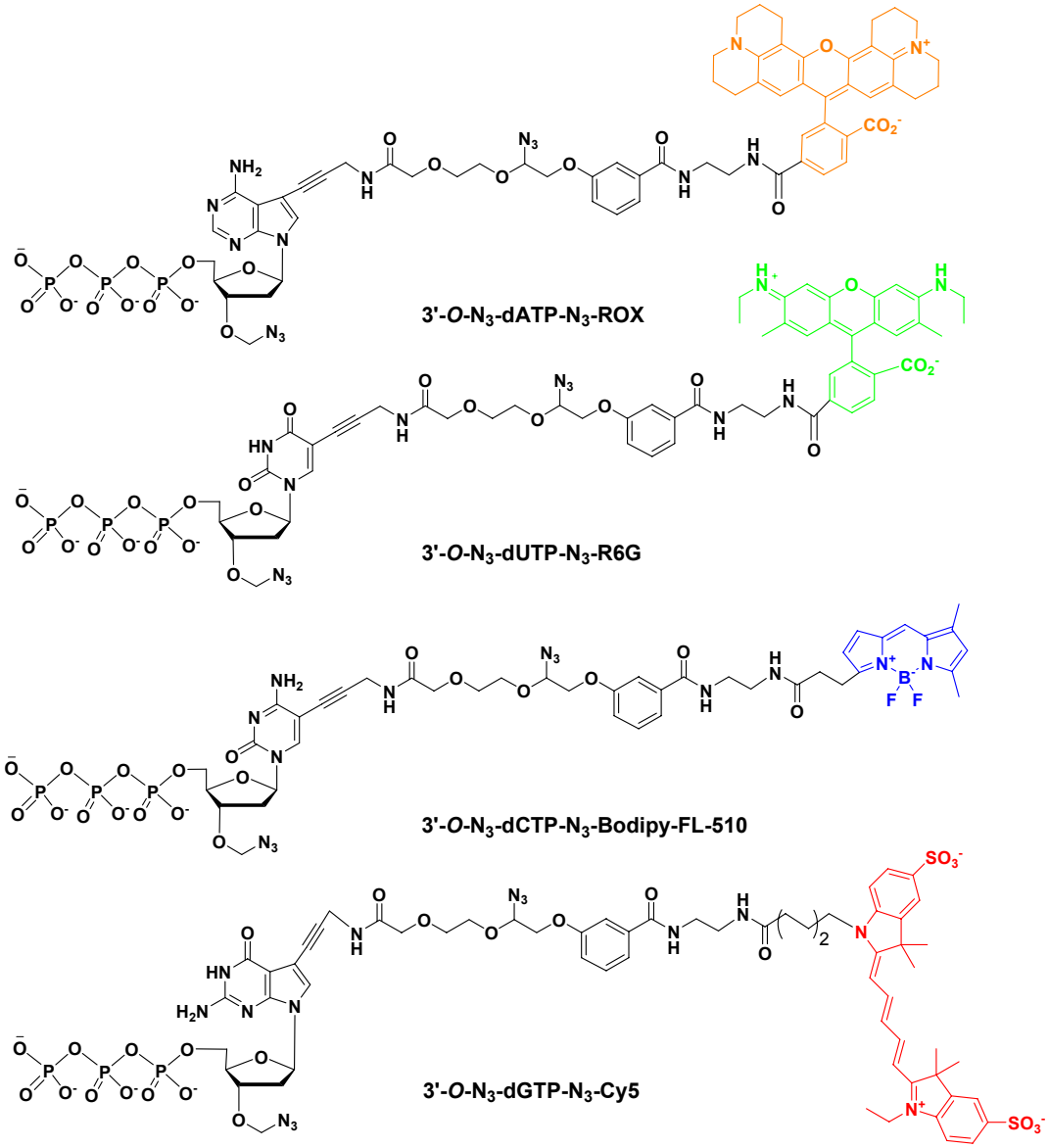


Fig. 6.3. Molecular structures of 3'-*O*-N_3-dATP-N_3-ROX, 3'-*O*-N_3-dUTP-N_3-R6G, 3'-*O*-N_3-dCTP-N_3-Bodipy-FL-510 and 3'-*O*-N_3-dGTP-N_3-Cy5.

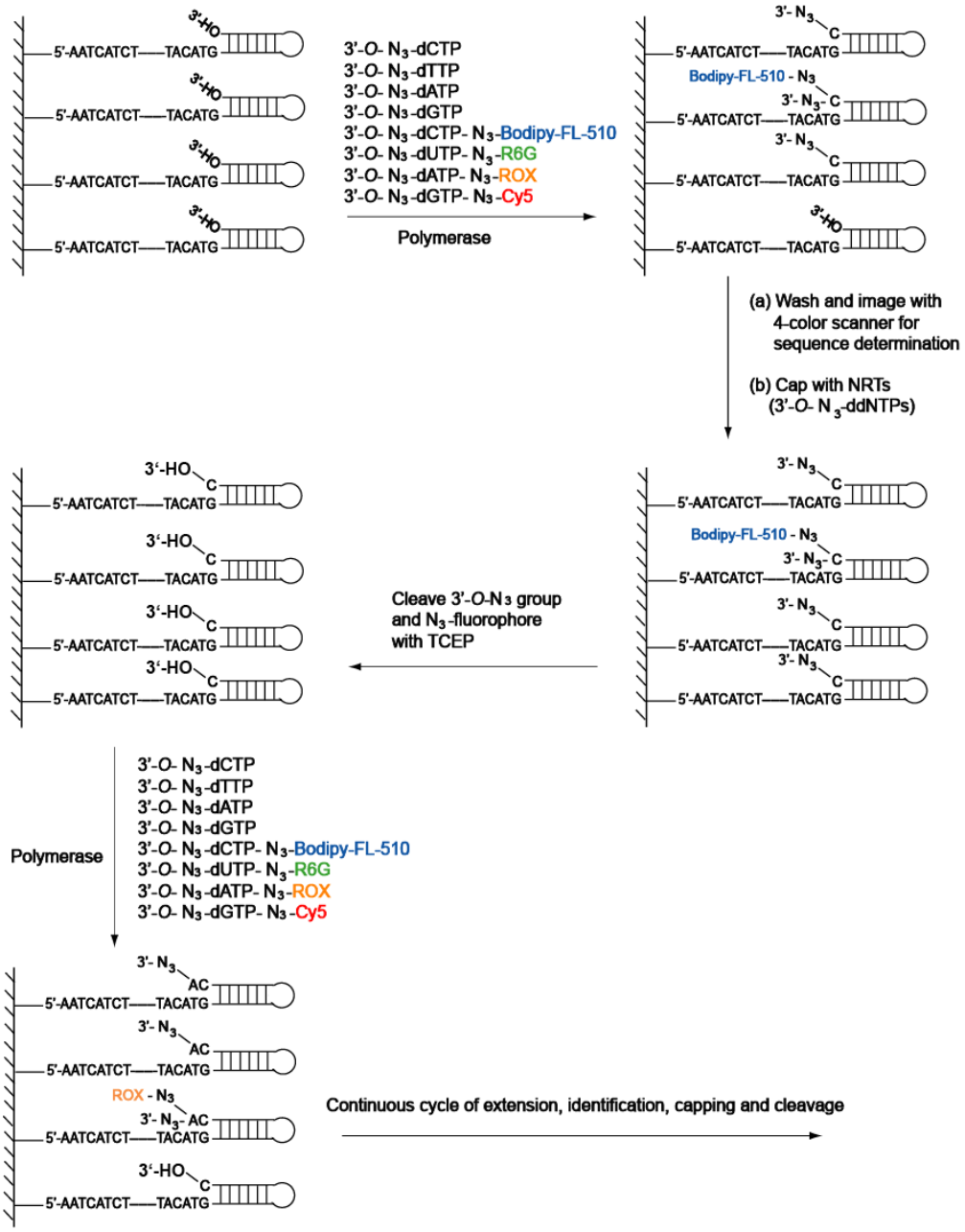


Fig. 6.4. Scheme of sequencing by synthesis using CF-NRTs (3'-*O*-N_3-dNTP-N_3-fluorophores) and NRTs (3'-*O*-N_3-dNTPs).

An alternative set of non-reversible cleavable fluorescent nucleotide terminators (ddNTP-N_3-fluorophoores) have also been developed[11] (Fig. 6.5), which have special utility for the hybrid Sanger-SBS approach to be described. This 4-color hybrid SBS has

intermediate properties between Sanger sequencing and the standard SBS reactions. The rationale is that the identity of the incorporated nucleotide is determined by the unique fluorescent emission from the four fluorescent dideoxynucleotide terminators (ddNTP-N_3-fluorophores, Fig. 6.5), while the role of the 3'-*O*-N_3-dNTPs is to further extend the DNA strand to continue the determination of the DNA sequence. In general, as shown in Fig. 6.6, sequencing reactions on a chip were initiated by extending the self-priming DNA templates using a solution containing four 3'-*O*-N_3-dNTPs, four ddNTP-N_3-fluorophores, and 9^{o}N DNA polymerase. The incorporated nucleotide is identified by fluorescence detection. Then a synchronization reaction using the four 3'-*O*-N_3-dNTPs in relatively high concentration was performed to reduce the amount of un-extended priming strands after the initial extension reaction. Both the fluorescent groups and 3' capping azidomethyl moiety are cleaved away by TCEP for the next sequencing cycles. Here, the DNA strand extended by ddNTP-N_3-fluorophores will no longer participate in subsequent polymerase reaction cycles because they are permanently terminated, while the DNA strand that incorporates the 3'-*O*-N_3-dNTPs will turn back into the natural nucleotide and continue with the next cycle. Therefore, the ratio between the amount of ddNTP-N_3-fluorophores and 3'-*O*-N_3-dNTPs during the polymerase reaction is crucial in determining how much of the ddNTP-N_3-fluorophores are incorporated and thus the corresponding fluorescent emission strength. With a finite amount of immobilized DNA template on a solid surface, as the sequencing cycles continue, the amount of the ddNTP-N_3-fluorophores needs to be gradually increased to

maintain the fluorescence emission strength for detection. The advantage of this strategy is that the extension with the mixture of 3'-*O*-N_3-dNTPs/ddNTP-N_3-fluorophores does not have a negative impact on the enzymatic incorporation of the next nucleotide analogue, because after cleavage to remove the 3'-OH capping group, the DNA products extended by 3'-*O*-N_3-dNTPs carry no modification groups. Guo *et al.*[11] in our lab performed hybrid SBS on a chip-immobilized DNA template using the 3'-*O*-N_3-dNTPs/ddNTP-N_3-fluorophores combination and obtained a read-length of around 30 bases on self-priming DNA templates.

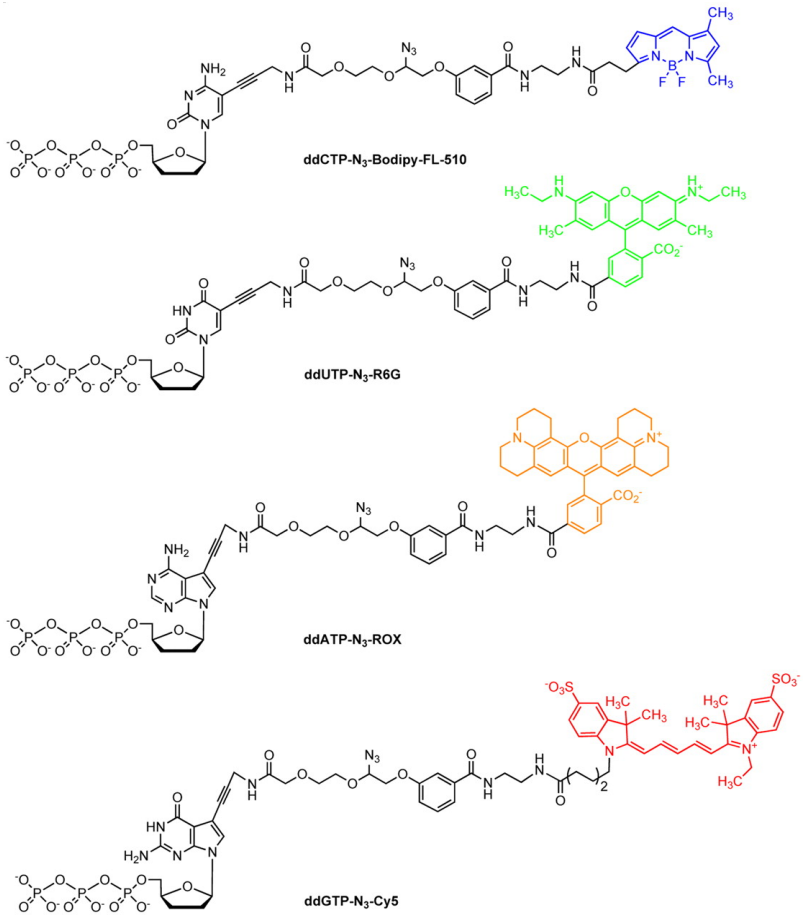


Fig. 6.5. Molecular structures of cleavable fluorescent dideoxynucleotide terminators: ddCTP-N_3-Bodipy-FL-510, ddUTP-N_3-R6G, ddATP-N_3-ROX and ddGTP-N_3-Cy5.

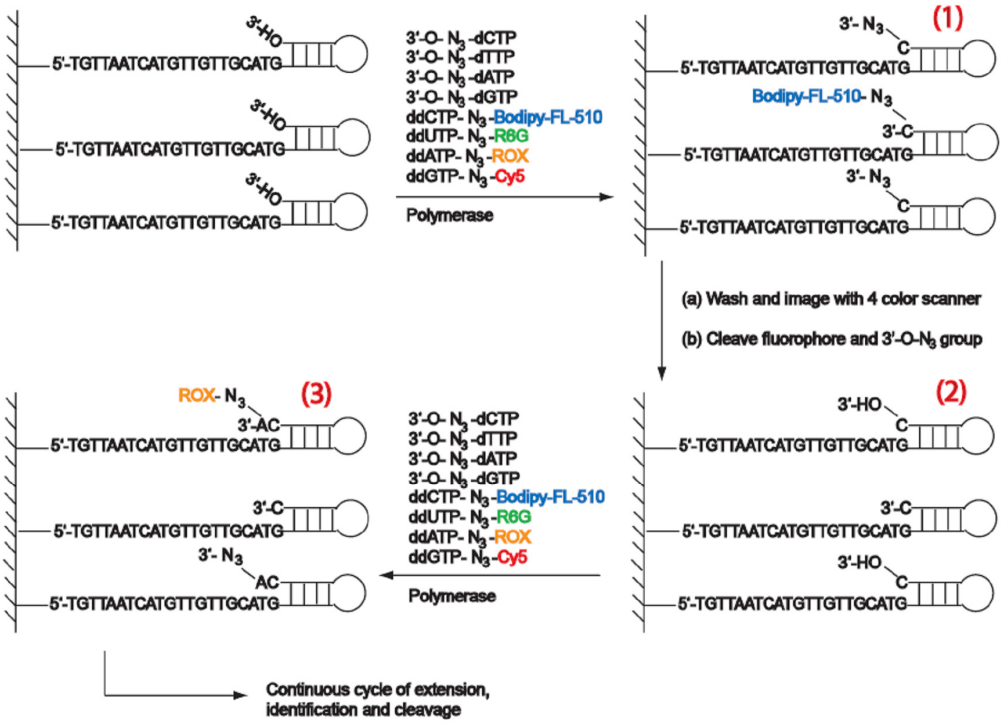


Fig. 6.6. A hybrid SBS scheme for 4-color sequencing on a chip using the nucleotide reversible terminators (3'-*O*-N_3-dNTPs) and cleavable fluorescent dideoxynucleotide terminators (ddNTP-N_3-fluorophores).

The above two strategies using azido based nucleotide analogues has demonstrated their great potential in DNA sequencing. However, the relatively short read-lengths obtained pose a challenge for DNA sequence assembly and *de novo* sequencing. Increasing read-length is of significance for fragment assembly, not only for sequencing by synthesis, but for all the other next-generation sequencing technologies. Despite the general progress in technology across different fields, including surface chemistry, nucleotide biochemistry, polymerase engineering, detection methodology and others, alternative strategies need to be developed to increase the read-length with current chemistries and methods.

To overcome the limitations of the short read-length, paired-end and mate-pair sequencing have been developed and are currently used in some commercial sequencing platforms. The paired-end approach developed by Illumina is to sequence both the forward and reverse strand of each molecule.[13] As shown in Fig. 6.7 A, after the first stage of sequencing, the synthesized sequence is washed away by chemical melting, and bridge amplification cycles are repeated to generate the reverse strand, after which the starting strand is selectively removed before annealing another sequencing primer for the second read. Using this paired-end sequencing approach, approximately twice the amount of data can be generated.

Another approach, called mate-pair sequencing, involves the preparation of special libraries. As shown in Fig 6.7 B, the long target DNA molecules are connected at each end to sticky-end adapters (referred to as internal adapters), allowing the ends of the molecule to come into close proximity by circularization. The circularized DNA molecule is then processed using restriction enzymes or fragmentation to generate the mate-pair segment, the two combined molecular ends, which are end ligated to the outer adapters and sequenced, theoretically from both types of adapters. This allows the sequence of both ends of the target DNA to be determined sequentially and in each direction, providing information about relative location and orientation of a pair of reads useful for assembly. For instance, if the starting fragments average 3 kb in length, the mate pair reads will also be approximately that distance apart, enabling one to scaffold genomic contigs generated by fragment sequencing using sophisticated computer

software.

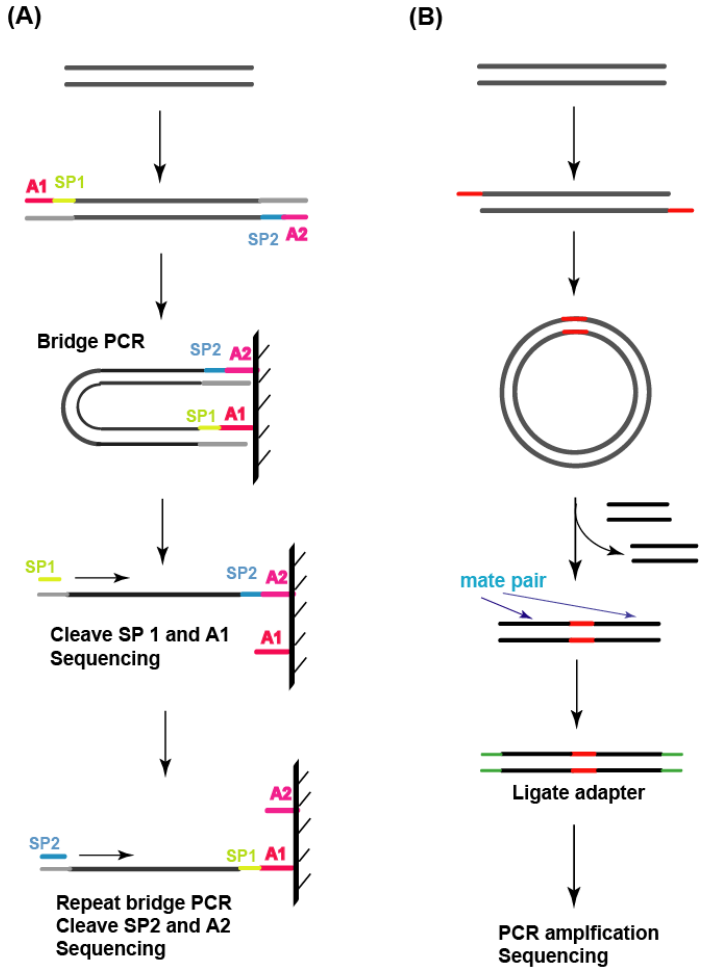


Fig. 6.7. Approaches to overcome short read-length. (A) Paired-end sequencing; (B) Library preparation for mate-pair sequencing

These two approaches have helped to improve the read-length and data analysis. However, the extra PCR cycles for the paired-end method and special library preparation for the mate-pair approach bring extra complexity to the process. It is desirable to develop a new strategy that is simple and effective to increase the read-length. In this chapter, we describe a SBS integrated primer walking strategy we have developed to

increase the read-length of SBS several fold. Here, the already-extended "sequenced" primer is stripped away after the first stage of SBS. Then a new primer is re-annealed to the DNA template and extended to the approximate end of the first round SBS with the use of three natural nucleotides (dATP, dCTP and dGTP) and one nucleotide reversible terminator (NRT, 3'-*O*-N_3-dGTP). This primer walking process results in a long natural DNA strand primed for the next stage of SBS. The repeated SBS and primer resetting will eventually increase the read-length of SBS several fold. As a proof of principle, 53 bp of read-length was achieved by the combination of SBS and primer walking, with the use of cleavable fluorescent nucleotide reversible terminators (instead of ddNTP-N_3-fluorophores) and nucleotide reversible terminators for the sequencing cycles.

6.2 Experimental Rationale and Overview

The fundamental rationale behind the primer walking strategy 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. In general, as shown in Fig. 6.8, six steps are involved in this approach: (1) annealing of the first primer, (2) performing a 1st round of 4-color SBS, (3) denaturing the sequenced section of the template to recover a single-stranded DNA for the second primer annealing, (4) annealing of a fresh primer identical to the first primer to the template, (5) walking of the primer by extending the primer with the walking nucleotides mixture to the approximate end of the first round sequence, (6) sequencing from that point. Theoretically, steps 1 through 6 can be carried out repeatedly until the

target DNA is sequenced in its entirety. The advantage of primer resetting lies in its ability to restore all the templates after the denaturation step, so the next stage of SBS can restart downstream from the first stage of SBS with potentially the same amount of sequenceable DNA as the previous round. Therefore, the read-length of sequencing is increased to the combination read-length of the first and second stages of SBS. The main challenges of the primer walking strategy are as follows: first, to obtain similar read-length on linear template compared to self-priming template, which is under the control of primer hybridization efficiency and stability; second, to strip away the extended primer after the first round of SBS under mild conditions, efficient enough but not damaging to the templates; third, to develop an effective walking method. The first two challenges can be addressed by optimizing the hybridization and denaturation conditions, while for the third one, there are three possible approaches: walking with nucleotide reversible terminators (NRTs), with three natural nucleotides, or with three natural nucleotides in combination with one NRT. The first approach only utilizes the NRTs, therefore, the extension stops after every single base extension and resumes after the cleavage reaction. This helps to prevent mis-incorporation, yet it requires the same number of repeated cycles as the first round SBS, which slows the walking process. The second approach is to employ two sets of natural nucleotide mixtures as substrates in rotation. For example, the first set of natural nucleotides is composed of dATP, dGTP, dCTP, and the second set consists of dATP, dGTP, dTTP. Using the first set, the polymerase reaction stops before base “A” in the template due to lack of the

complementary base dTTP, and will resume after bringing in the second mixture containing dTTP, resulting in a polymerase reaction that stops at the base "G" in the template due to the lack of dCTP. The idea of using natural nucleotides alone simplifies and speed up the whole process, but because of the presence of only three nucleotides, the enzyme sometimes incorporates an incorrect dNTP instead of stopping at the expected position. The third approach combines the advantages of NRTs and natural nucleotides. The incorporation is conducted using three natural dNTPs (e.g. dATP, dCTP, dTTP) and one NRT (3'-*O*-N_3-dGTP). Therefore, the polymerase extension will continue until it incorporates the NRT, and resumes after the cleavage reaction to regenerate the 3'-OH. Repeated cycles of such incorporation and cleavage will fill the gap between the first and second rounds of SBS at a relatively fast speed. In this study, only the second and third approaches were explored, since the first one was judged to be too time consuming.

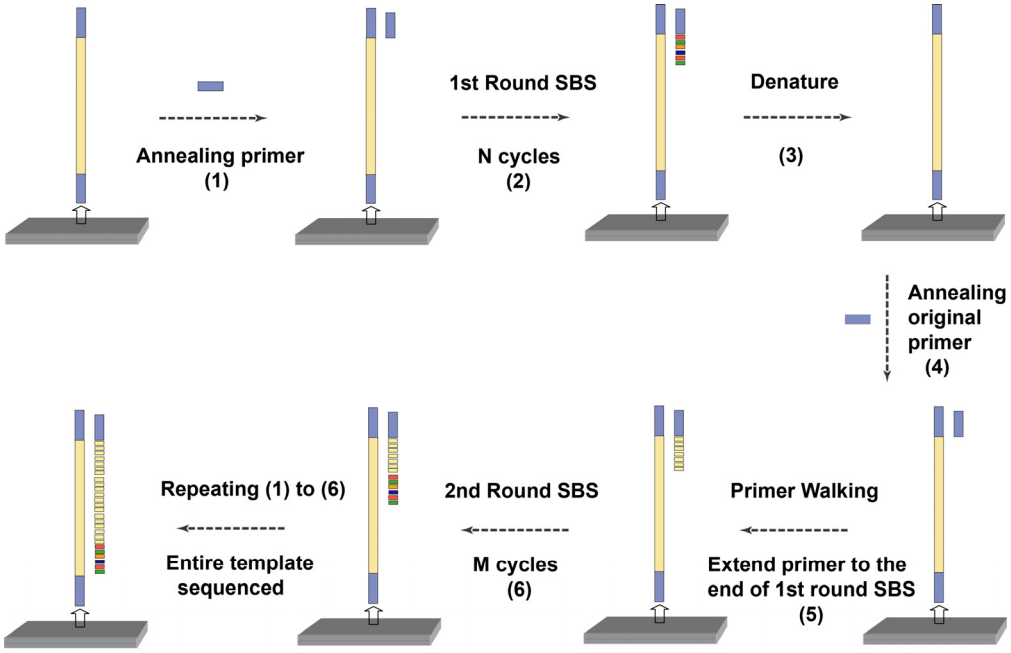


Fig. 6.8. General scheme of primer walking strategy for extending the read-length.

6.3 Results and Discussion

6.3.1 Optimization of primer annealing to DNA template on solid surface

Previous sequencing by synthesis performed in our lab was mostly carried out with a loop template, where the self-priming sequence inside the DNA template strand guaranteed that each template could be analyzed and the primers would not fall off as the cycles progressed, giving the maximum possible signals. However, to perform the SBS on a linear template, one requires external primers to anneal to the DNA template for the initiation of sequencing. This poses challenges with regard to the annealing efficiency of primer hybridization to the template as well as the stability of the primer-template duplex: in most cases steric hindrance will prevent primer hybridization to the surface-bound linear template in the first place and the primers tend to shear off after undergoing

multiple washing steps. To generate a sufficient signal, it is crucial to specifically anneal primers to as many of the templates in the ensemble as possible and for the interaction to be stable enough to survive many cycles. Different conditions were tested by varying the concentrations of the salt and primer, the length of primer, and annealing temperature. As shown in Fig. 6.9, the base calling signal strength is highly related to the concentration of primers during the hybridization, and the results with both the 37mer and 52mer primer indicated that use of 3.5 μM primers in a total volume of 10 μl generated the strongest signal. This concentration was then used for all the annealing steps. Though the data are not shown here, by setting up parallel experiments, it was demonstrated that higher salt concentration actually contributed to better hybridization efficiency, multiple incubations with fresh hybridization mixture also led to better hybridization results, and an annealing temperature of 65^{o}C gave a more specific signal with lower background compared to two lower temperatures, 60^{o} and 55^{o}C. After several cycles of sequencing by synthesis, it was also demonstrated that a primer length of 37 bases gave a continuous stable signal.

The amount of template on the surface is also a very important factor in performing SBS with a linear template. It was expected that CodeLink HD (HD) slides could bear more DNA linear templates than CodeLink (CL) slides, but this did not necessarily mean that there would be more primer annealed templates available for sequencing (sequenceable tempates) if the steric hindrance was greater due to the higher surface density. Furthermore, CodeLink slides are easier to visualize in our array scanner. Therefore, it was worthwhile to investigate the effects of surface-bound template

concentration. Parallel experiments were set up and the results demonstrated that HD slides gave better signals than CL slides in single base extension reactions (Fig. 6.10).

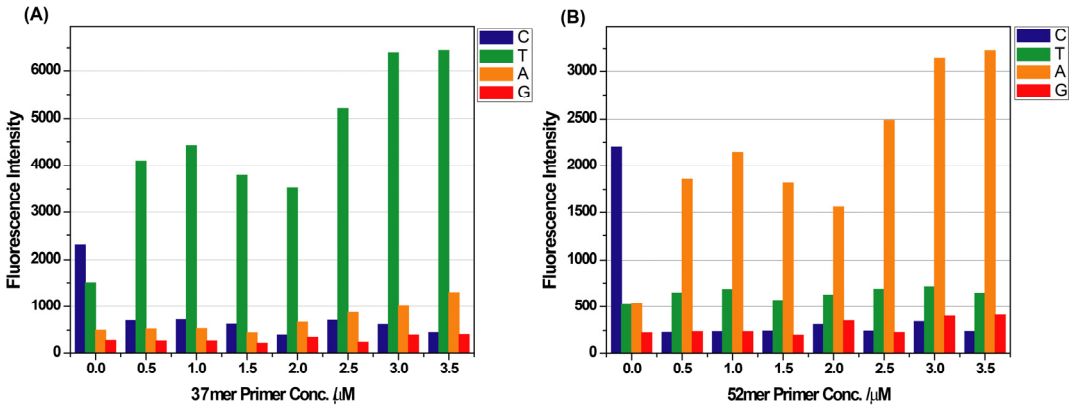


Fig. 6.9. The effects of primer concentration: one base extension with fluorescent nucleotide reversible terminators (ddNTP-N_3-fluorophores) after hybridization of primers at different concentrations. (A) 37mer primer (B) 52mer primer.

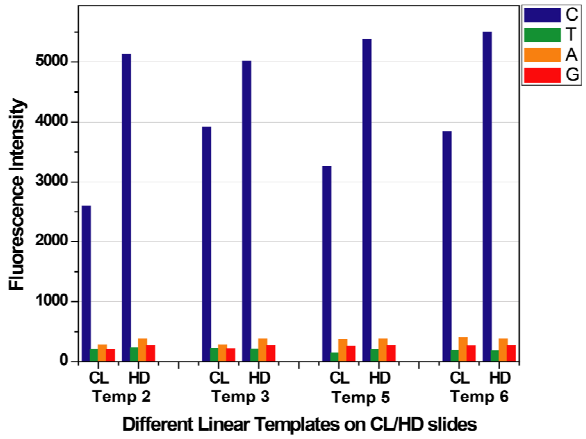


Fig. 6.10. Comparison of CL and HD slides: one base extension after primer hybridization.

6.3.2 Sequencing by synthesis on linear DNA template

Compared to sequencing on loop templates, the difficulties of sequencing a linear DNA template lies in the lower signal strength due to fewer sequenceable templates and the stability of the primer/template duplex over multiple cycles. This difficulty is

amplified in hybrid SBS sequencing, since in each cycle the incorporation of nucleotide terminators to the primers leads to loss of those templates from further reactions. In this study, hybrid SBS sequencing chemistry was used for testing linear template sequencing, as it has a higher requirement for stability of primer/template duplexes and the optimized conditions could be directly used for SBS with fluorescent nucleotide reversible terminators.

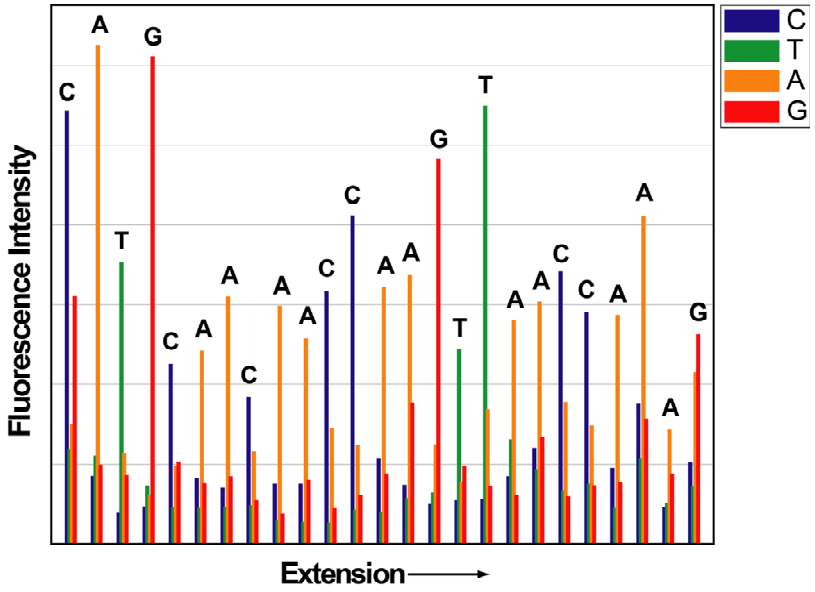


Fig. 6.11. 4-color hybrid SBS data on a linear template (Template1) using ddNTP-N_3-fluorophores and 3'-*O*-N_3-dNTPs.

With respect to the stability of the DNA duplex, temperature is the key factor. The lower the temperature, the more stable the duplex. However, the temperature should be high enough for the enzyme to achieve efficient incorporation. Previously, for sequencing on loop templates, 65°C was used for extension reactions. Based on this, different temperatures, 65°C, 60°C, 55°C and 50°C, were tested under the same conditions.

Though the data are not shown here, 55^{o}C turned out to give the longest read-length in hybrid SBS cycles. On the other hand, for hybrid SBS itself, the ratio of fluorescently labeled nucleotide terminators and reversible nucleotide terminators must be carefully controlled in order to reach reasonable read-lengths. As shown in Fig. 6.11, a total read-length of 25 bp was achieved, which is similar to the read-length obtained with the loop template.[11]

6.3.2 Primer resetting and walking for extending read-length

As described above, the primer walking strategy involves denaturing primers away from the templates after the first round of sequencing, re-annealing a fresh primer to the template and extending the primer to the approximate end of the first round of sequencing. Each step was tested and optimized in this study.

Different methods could be applied to denature the DNA duplex, such as high temperature (>90^{o}C), strong alkaline solution (e.g. 0.1-0.5 M NaOH), or formamide treatment. NaOH was excluded in the testing due to its possible chemical damage to the surface. Considering the potential damage to the DNA templates on the surface by high temperature or concentrated formamide, the combination of elevated but not high temperatures with dilute formamide (80%) was employed in the denaturation step. Different temperatures were tested in order to completely remove the first round sequencing primer. Since the first run of sequencing was stopped at the incorporation step prior to cleavage of the fluorophore, the denaturation efficiency should be reflected

by the amount of loss of fluorescence signal. However, the integrity of the DNA template on the surface could only be analyzed after primer resetting and SBS cycles. Hence, after denaturing the primer, the fresh primer was annealed and single base extension was performed. The integrity of the DNA template was compared with the fluorescence intensity after the first fluorescent nucleotide was incorporated. Different conditions were tested in this study to achieve better denaturation efficiency while keeping the surface interactions intact. The results demonstrated that 80% formamide incubation at 65^{o}C gave the best result. It was also discovered that replenishing the denaturation mixture multiple times at shorter intervals helped to achieve higher overall specific signals. To effectively remove the non-specific binding of the fluorophores from the surface, 0.1% SDS was added to the denaturation mixture.

The key of the primer walking strategy is to fill in the primer with natural nucleotides to the end of the first round of sequencing. As discussed in the following section, walking with three natural nucleotides and the combination of natural nucleotides with reversible terminators were both tested in this study.

6.3.2.1 Primer walking using three natural nucleotides

The rationale behind primer walking using three natural nucleotides is that the DNA polymerase reaction will stop due the lack of a complementary nucleotide and will resume in the presence of a complementary nucleotide. For example, if the downstream sequence after the primer is 5'-ATGCTAG---3', in the present of three natural

nucleotides (dATP, dTTP and dGTP), the primer is only extended to 5'-ATG. Polymerase stops incorporating due to the absence of dCTP. With the introduction of the alternative nucleotide mixture (dCTP, dATP, dTTP), the primer will be continually extended to 5'-ATGCTA up to the G site. By alternating the two incorporation mixtures, the primer was extended ("walked") to the end of the first round of sequencing.

The high fidelity and accuracy of the DNA polymerase is crucial for carrying out template walking with three natural nucleotides. Most importantly, the polymerase must not incorporate a miscellaneous nucleotide where the complementary nucleotide is missing. Various DNA polymerases were tested to find the most suitable ones for this application. Starting with a single base extension reaction, a self-priming template (SP26dT, M.W. 7966) and a single natural nucleotide, dTTP, were tested. It was expected that primer extension would stop right after the polymerase incorporated dTTP due to the absence of the next complementary nucleotide, dGTP. Hence, the expected molecular weight of the SP26T after a single base extension should be 8270. Four different commercially available DNA polymerases were tested: 9^{o}N Therminator II, Pfu DNA Polymerase (with 3'-5' exonuclease activity), Tgo DNA polymerase and Thermo Sequenase, and the extension products were analyzed by MALDI-TOF MS. As shown in Fig 6.12, the four polymerases behaved differently during the single base extension reactions. 9^{o}N enzyme was not able to stop after incorporating the first dTTP but mis-incorporated an extra dTTP to yield a DNA product with two Ts (M.W. 8606). Instead of displaying incorporation activity, both the Pfu and Tgo polymerases cleaved

the loop primer from the 3' to 5' direction due to their exonuclease activity. Though not presented completely in Fig. 6.12 C, Tgo polymerase digested more than 15 bases off the primer. Thermo Sequenase was the only polymerase to correctly incorporate a single dTTP and promptly stopped ahead of a mismatch at the second, displaying a single peak at 8270. Therefore, Thermo Sequenase was chosen for further testing.

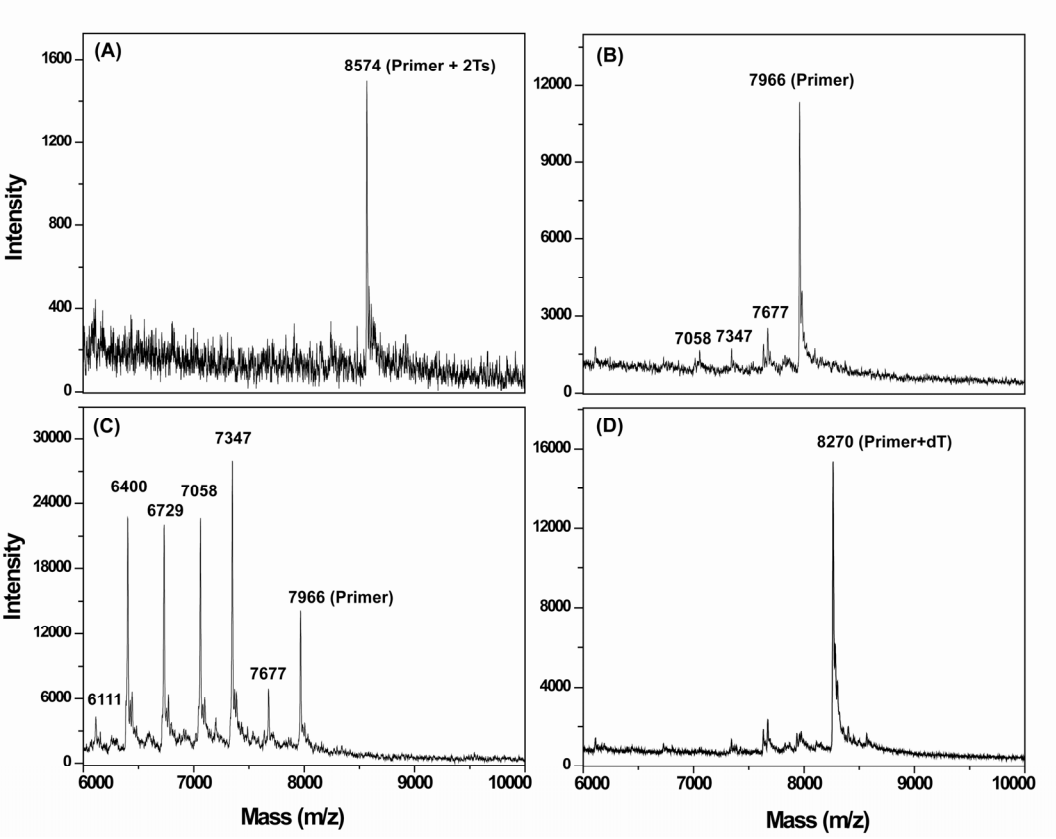


Fig. 6.12. MALDI-TOF MS spectra of single base extension products using self-priming SP26T template and dTTP with DNA polymerase (A) Therminator II (9^{o}N), (B) Pfu DNA polymerase, (C) Tgo DNA polymerase and (D) Thermo Sequenase.

In order to fulfill the requirement of walking, Thermo Sequenase was further evaluated for its ability to correctly incorporate three natural nucleotides. In primer

walking, the walking rate depends on polymerase fidelity in the presence of all but one natural nucleotide. The same self-priming template, SP26T was used, together with three nucleotides, dATP, dGTP and dTTP. Since the sequence immediately downstream of the primer is 5'-TGAC-3', the polymerase was expected to stop incorporation at the third base (A). However, the MALDI-TOF MS spectrum result indicated that mis-incorporation of the fourth base occurred, as shown in Fig. 6.13 A, the extra peak at 8217 corresponding to the mis-incorporation of a dTTP. Although the majority of extension products had the correct three nucleotides incorporated (m/z 8912), this protocol still could not be used for walking. We hypothesized that the presence of excessive dNTPs in the reaction mixture during the extension reaction lowered the fidelity of Thermo Sequenase, resulting in the incorporation of the extra base. Hence, a PyroE solution, which contains Apyrase, an enzyme used to eliminate un-reacted dNTPs during pyrosequencing, was added to the extension mixture to digest the excess dNTPs. The result, as shown in Fig. 6.13 B, demonstrated that the combination of Thermo Sequenase and PyroE solution enabled the correct incorporation of the nucleotides.

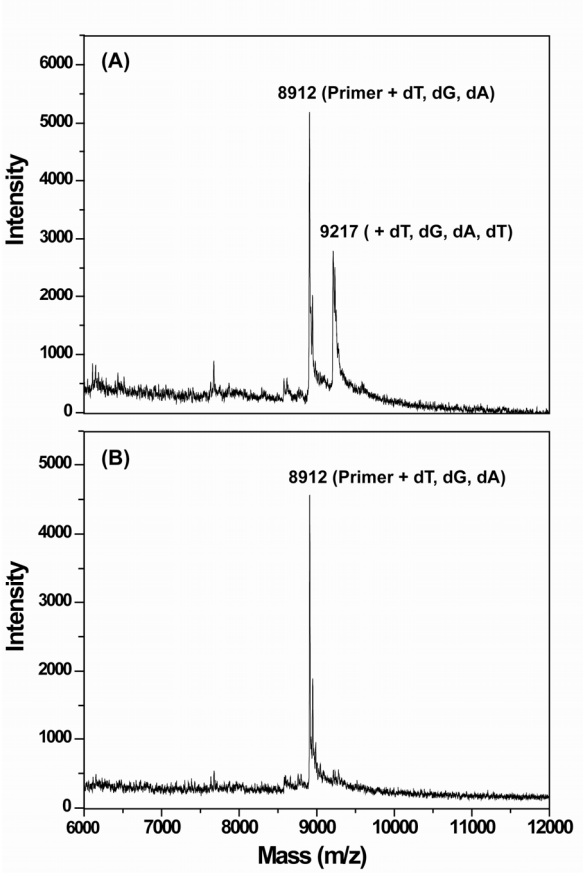


Fig. 6.13. MALDI-TOF MS spectra of self-priming SP26T template extended with dATP, dGTP and dTTP using (A) Thermo Sequenase (B) Thermo Sequenase and PyroE solution.

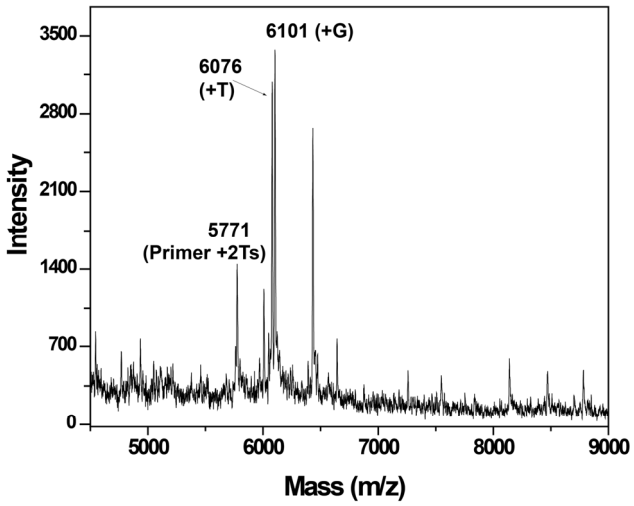


Fig. 6.14. MALDI-TOF MS spectrum of Primer5163 extended with dTTP, dGTP and dCTP using Thermo Sequenase and PyroE solution

To truly integrate primer walking with natural nucleotides in the SBS platform, the extension reaction needs to be carried out on linear templates instead of self-priming templates, in order to remove the primers for the next round of sequencing. This was tested using a synthetic linear template, Exon8, and related primer5163. Three nucleotides, dTTP, dGTP, dCTP, were used. The first several bases after the priming site are 5'-TTAGA-3', therefore, the use of nucleotides dTTP, dGTP, dCTP will enable the primer to walk to the second base (T) and stop due to the lack of dATP. The molecular weight of the primer is 5163 Da, and after extension with two Ts, the mass of the primer should have increased to 5771. However, the result (Fig. 6.14) showed that Thermo Sequenase failed to stop incorporation in the absence of complementary nucleotides. Instead, it mis-incorporated either a dTTP or dGTP at the position that is supposed to be A, generating peaks of 6076 and 6101 respectively. Though additional peaks are not labeled in the figure, it is apparent that Thermo Sequenase continued incorporating incorrect bases far beyond the A position. Though different modifications were tested, including varying the temperature, the amount of enzyme, nucleotide concentration, the amount of PyroE solution and others, we were not able to obtain the correct number of incorporated nucleotides. Hence, we decided to explore the next option, that is, primer walking with three natural nucleotides and one nucleotide reversible terminator.

6.3.2.2 Primer walking using three natural nucleotides and one reversible nucleotide terminator

The strategy of using three natural nucleotides in combination with one nucleotide reversible terminator is to regulate the primer walking by halting the primer extension with nucleotide reversible terminators (NRTs, 3'-*O*-N_3-dNTPs). When a primer extends with an NRT, primer walking will stop. With the cleavage of the 3'-azidomethyl capping moiety on the NRT, the 3'-OH group is recovered and ready for the next round of walking. Unlike template walking with three natural nucleotides, the use of reversible terminators effectively stops the DNA polymerization, hence increasing the accuracy of the DNA polymerase and reducing the chance of mis-incorporation. For example, if a walking nucleotide mixture, consisting of three natural nucleotides (dATP, dCTP and dTTP) and one nucleotide reversible terminator (NRT, 3'-*O*-N_3-dGTP) is used, all walking primers should be synchronized at the G nucleotide after the incorporation of the NRT, which should facilitate the analysis of the sequence.

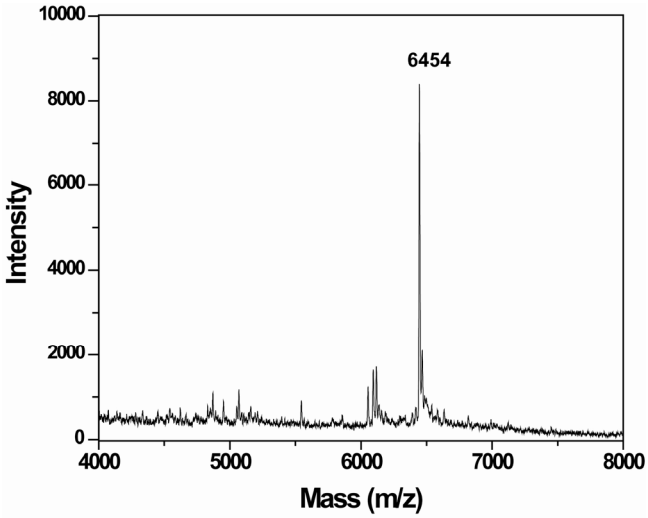


Fig. 6.15. MALDI-TOF MS spectrum of Primer5163 extended with dTTP, dGTP, dCTP and 3'-*O*-N_3-dATP using 9^oN polymerase

MALDI-TOF MS was first used to test the primer walking reactions in solution and investigate the incorporation kinetics for condition optimization. A walking mixture of three natural nucleotides (dATP, dCTP and dTTP) and one NRT (3'-*O*-N_3-dGTP) was used, as the sequence immediately downstream of the primer is 5'-TTAG-3'; therefore, the expected mass after walking should be 6454 m/z. As shown in Fig. 6.15, it was demonstrated that the incorporation of 3'-*O*-N_3-dGTPs completely stopped the primer extension, though there were some incomplete incorporation products before this major peak. This can be addressed by increasing the dNTPs to 3'-*O*-N_3-dGTP ratio.

To keep the test simple, direct sequencing after three walking cycles was first explored, without the involvement of first round sequencing and denaturation. Again, the hybrid SBS chemistry was used for the test. As shown on Fig. 6.16 A, primers were first hybridized to the linear templates on the chip, and then extended by polymerase reaction

using three dNTPs (dCTP, dATP, dTTP) and 3'-*O*-N_3-dGTP. The extension process stopped after the first G. To synchronize the un-extended primers, a capping step with 3'-*O*-N_3-dGTP was performed. Upon cleavage of the 3'-*O*-azidomethyl group, the 3'-OH of the primer was regenerated and ready for the next cycle of walking. Here, 3 walking cycles were performed, which traversed 25 bases. After walking, 24 bases of DNA sequence were correctly identified after carrying out the "second" round hybrid SBS, as shown in Figure 6.16 B. This demonstrated the feasibility of the primer walking strategy.

To further explore the feasibility of the primer walking strategy, collaborating with Dr. Yu in our lab,[5] the complete process, including first round SBS sequencing, primer resetting and walking, and second round SBS sequencing, was carried out by SBS chemistry using cleavable fluorescently labeled nucleotide reversible terminators (CF-NRTs, dNTP-N_3-fluorophores). As shown in Fig. 6.17, the first 30 basses were sequenced during the 1st round of SBS. After stripping away the sequencing primer, the original primer was reannealed to the template, followed by four cycles of primer 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, which, together with the 30 bases in the first round, achieved an overall read-length of 53 bases. In this experiment, we used reversibly blocked CF-NRTs (3'-*O*-N_3-dTNP-N_3- fluorophores) in place of ddNTP-N_3-fluorophores.

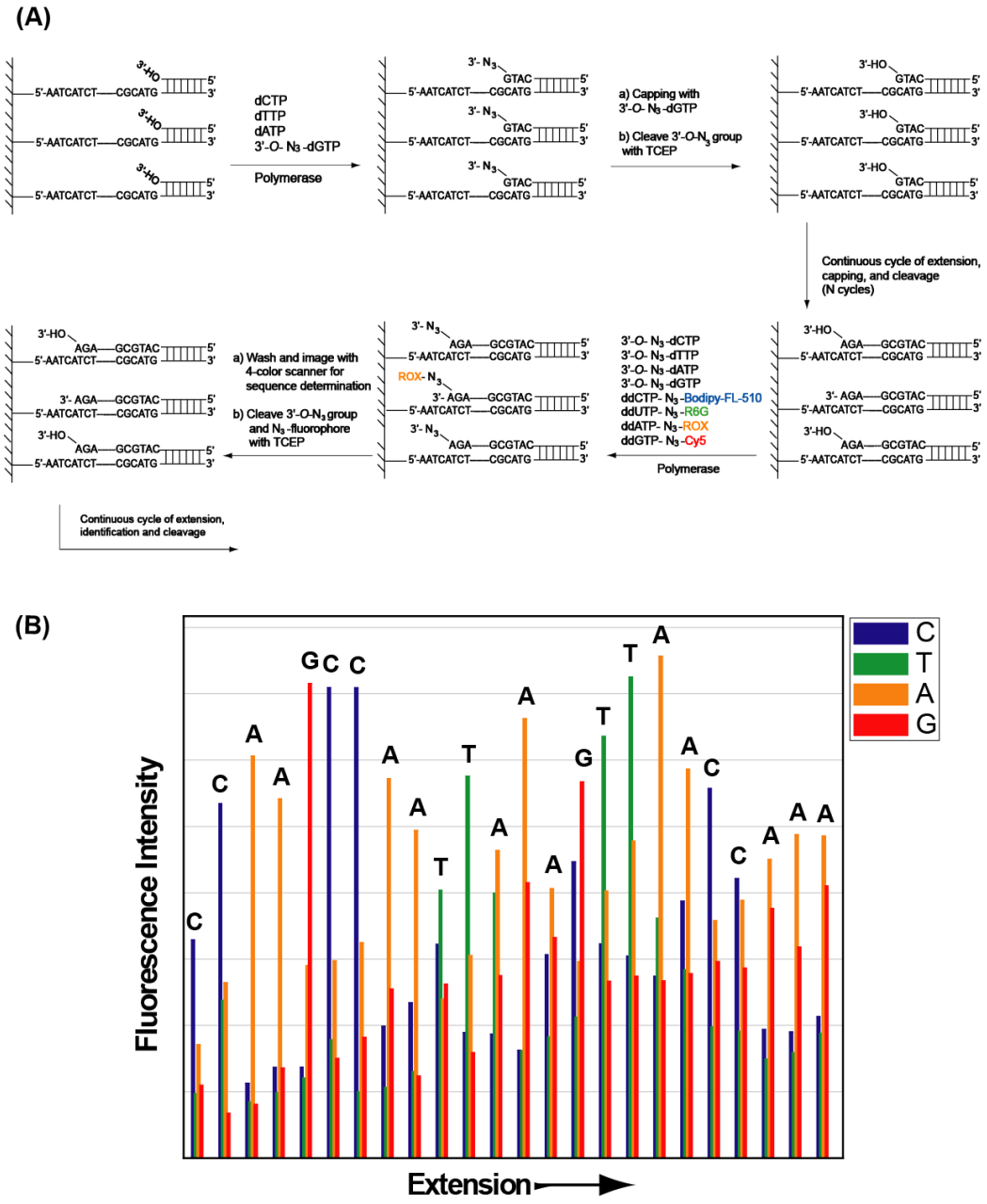
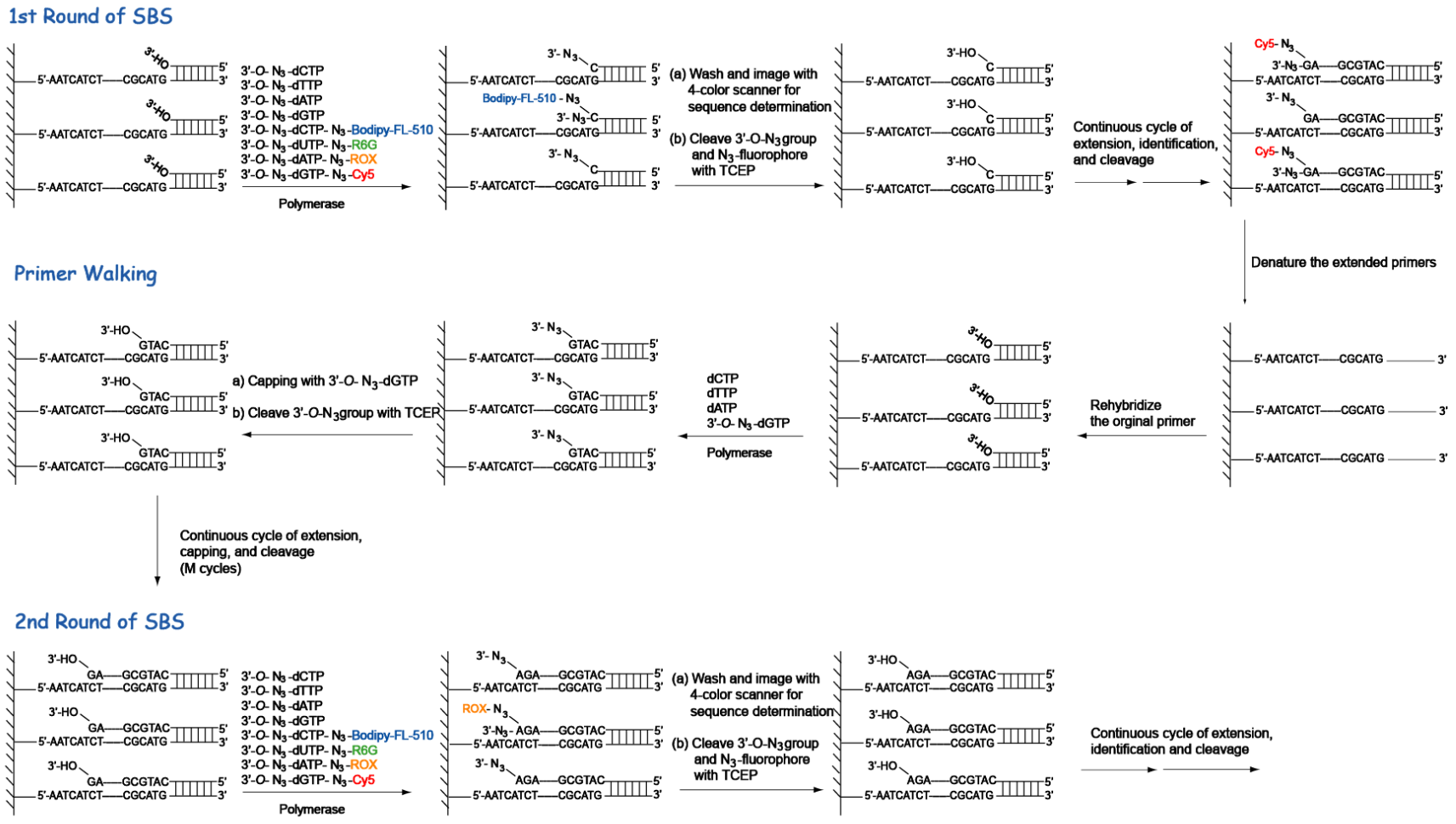


Fig. 6.16. Four-color DNA sequencing by hybrid SBS after primer "walking". (A) A scheme for primer walking followed by hybrid SBS sequencing. The primer hybridized to the linear templates was extended by polymerase using three dNTPs (dCTP, dATP, dTTP) and 3'-*O*-N_3-dGTP, and extension stopped after the first G. After a capping step to achieve synchronization, the 3'-OH of the extended primer was regenerated by removing the 3'-*O*-N_3 group. Hybrid SBS sequencing started after 3 walking cycles. (B) Four-color sequencing data obtained by hybrid SBS sequencing after primer "walking" 25 bases.

(A)


1st Round of SBS
3'-O- N3 -dCTP
3'-O- N3 -dTTP
3'-O- N3 -dATP
3'-O- N3 -dGTP
3'-O- N3 -dCTP- N3 -Bodipy-FL-510
3'-O- N3 -dUTP- N3 -R6G
3'-O- N3 -dATP- N3 -ROX
3'-O- N3 -dGTP- N3 -Cy5
Polymerase
Bodipy-FL-510 - N3
(a) Wash and image with 4-color scanner for sequence determination
(b) Cleave 3'-O-N3 group and N3 -fluorophore with TCEP
Continuous cycle of extension, identification, and cleavage
Cy5- N3
Denature the extended primers
Primer Walking
a) Capping with 3'-O- N3 -dGTP
b) Cleave 3'-O-N3 group with TCEP
dCTP
dTTP
dATP
3'-O- N3 -dGTP
Polymerase
Rehybridize the orginal primer
Continuous cycle of extension, capping, and cleavage (M cycles)
2nd Round of SBS
3'-O- N3 -dCTP
3'-O- N3 -dTTP
3'-O- N3 -dATP
3'-O- N3 -dGTP
3'-O- N3 -dCTP- N3 -Bodipy-FL-510
3'-O- N3 -dUTP- N3 -R6G
3'-O- N3 -dATP- N3 -ROX
3'-O- N3 -dGTP- N3 -Cy5
Polymerase
ROX- N3
(a) Wash and image with 4-color scanner for sequence determination
(b) Cleave 3'-O-N3 group and N3 -fluorophore with TCEP
Continuous cycle of extension, identification and cleavage

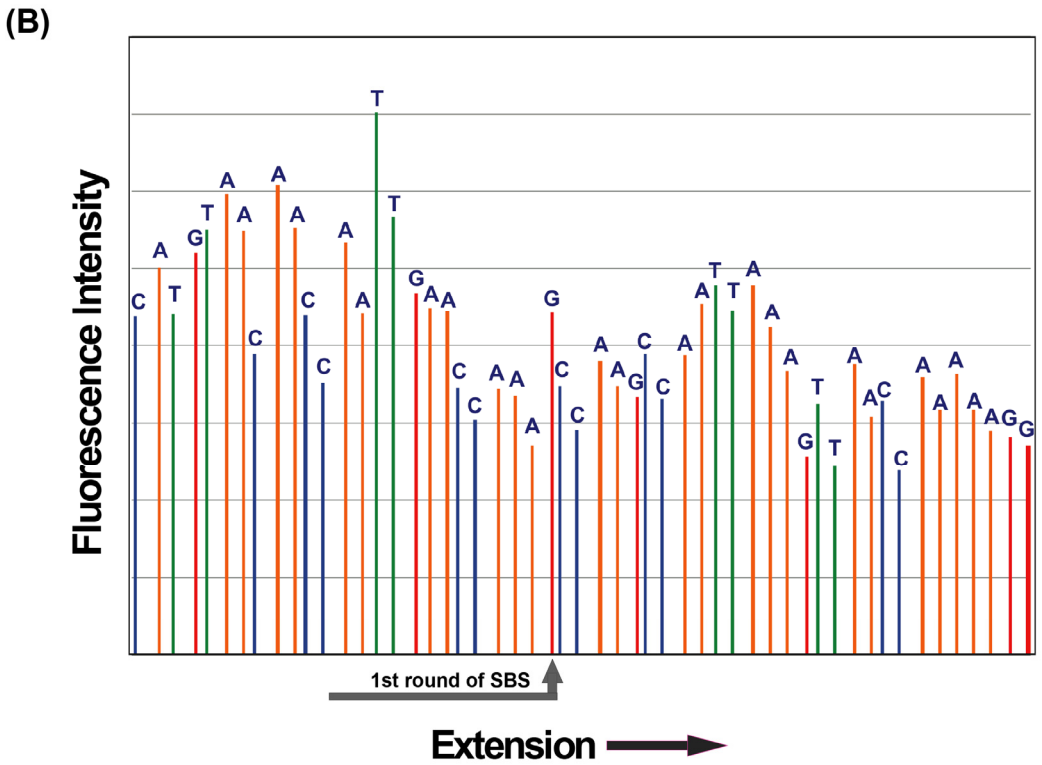


Fig. 6.17. Implementation of primer walking strategy for extending the read-length. (A) Scheme of SBS integrated primer walking. Not shown are capping in between each sequencing incorporation cycle. (B) 4-color SBS sequencing data obtained by combining 1st and 2nd round SBS using CF-NRTs. The first 30 bases were sequenced during the 1st round of SBS. 23 bases were sequenced in the second round, which generated the overall read-length of 53 bases.

6.4 Materials and Methods

General Information. All chemicals were purchased from Sigma-Aldrich (St. Louis, MO) unless otherwise specified. Cleavable fluorescent nucleotide reversible terminators (CF-NRTs, dNTP-N_3-fluorophores), nucleotide reversible terminators (NRTs, 3'-*O*-N_3-dNTPs) and fluorescent dideoxynucleotide terminators (ddNTP-N_3-fluorophores) were synthesized in our lab. Natural deoxynucleotides (dNTPs) and dideoxynucleotides

(ddNTPs) were purchased from Sigma-Aldrich. Oligonucleotides used for primers and templates were from Integrated DNA Technologies (Coralville, IA). Therminator™ II DNA polymerase (9°N) was purchased from New England Biolabs (Ipswich, MA), Pfu DNA polymerase from Agilent Technologies (Santa Clara, CA), Tgo DNA polymerase from Roche Applied Science (Indianapolis, IN), and Thermo Sequenase from GE Healthcare (Piscataway, NJ). CodeLink® Activated Microarray Slides and CodeLink® HD Activated Microarray Slides for DNA template immobilization were obtained from GE Healthcare as well. Seven different DNA templates were spotted on the slides in multiple replicates, one row for each type of template, with three rows of loop templates as the positive control. The sequence information for the DNA templates is shown in Table 6.1. Mass measurement of DNA was performed on a Voyager DE™ MALDI-TOF mass spectrometer (Applied Biosystems by Life Technologies, Carlsbad, CA). The DNA array was analyzed under a four-color ScanArray Express Scanner (PerkinElmer Life Sciences, Boston, MA).

Table 6.1 Sequence information for the DNA templates on slide

Template 1	5'-CAC TCA CAT ATG TTT TTC ATG GTA CCG TCA TAG TCA **GTG ACA TGC GAC TTA AGG CG**C TTG **CGC CTT AAG TCG CAT GTC AC**-3'
Template 2	5'-AAT GAA TGT ATC ACT ACT CAG ATC GTT ACT GAA CAT CTG CAT GGT TCA CCT CGC TGT GAC **GTG CCC ATG CGA GTG CGA GTG CAC GTG GCG CAG CAG GTC A**
Template 3	5'-GCA TGG TTC AAA TGA ATG TAT CAC TAC TCA GAT CGT TAC TGA ACA TCT CCT CGC TGT GAC **GTG CCC ATG CGA GTG CGA GTG CAC GTG GCG CAG CAG GTC A**-3'
Template 4	5'-CAC TCA CAT ATG TTT TTT AGC TTT TTT AAT TTC TTA ATG ATG TTG TTG CAT G**CG ACT TAA GGC G**CT TG**C GCC TTA AGT CG**-3'

Template 5	5'-CTG AAC ATC TGC ATG GTT CAA ATG AAT GTA TCA CTA CTC AGA TCG TTA CCT CGC TGT GAC **GTG CCC ATG CGA GTG CGA GTG CAC GTG GCG CAG CAG GTC A**-3'
Template 6	5'-CAG ATC GTT ACT GAA CAT CTG CAT GGT TCA AAT GAA TGT ATC ACT ACT CCT CGC TGT GAC **GTG CCC ATG CGA GTG CGA GTG CAC GTG GCG CAG CAG GTC A** -3'
Template 7	5'-ATC ATG TCA TGA ATC ACA CTC ACA TAT GTT TTT CAT GGT ACC GTC ATA GTC A**CG ACT TAA GGC G**CT TG**C GCC TTA AGT CG**-3'

Note: Template 1, 4 and 7 are self-priming loop templates. The letters in black bold indicate the primer hybridization site, and the letters in orange bold indicate the self-priming site.

6.4.1 Primer hybridization

To optimize the primer hybridization, different conditions were tested. The annealing buffer is based on 1X ThermoPol Reaction buffer (10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$), with the addition of primers, and salt. Different concentrations of primers, from 0.5 μM to 3.5 μM, in 0.5 μM intervals, were tested. Primers with different lengths, 20mer, 37mer and 52mer, were compared under the same conditions. Annealing buffers with and without NaCl (0.5 M) were compared. Different annealing temperatures, 55°C, 60°C and 65°C, were tested. The comparison analysis was based on the fluorescent signal from single base extension with ddNTP-N_3-fluorophores after the hybridization. Generally, a 10 μl reaction mixture consisting of 0.2 μl solution A (3 μM 3'-*O*-N_3-dATP, 0.5 μM 3'-*O*-N_3-dGTP, 3 μM 3'-*O*-N_3-dCTP and 3 μM 3'-*O*-N_3-dTTP) and 7.3 μl solution B (200 nM ddATP-N_3-ROX, 100 nM ddGTP-N_3-Cy5, 50 nM ddCTP-N_3-Bodipy-FL-510 and 100 nM ddUTP-N_3-R6G), 1 U of 9°N DNA polymerase, 20 nmol of $MnCl_2$ and 1X Thermopol II reaction buffer was added to the spot at 65°C for 15 min. After washing the chip with SPSC buffer containing 0.1% Tween-20 for 1 min,

the surface was rinsed with dH_2O, dried briefly and then scanned to detect the fluorescence signal. The finalized hybridization condition used for the experiment was as follows: 10 μl of hybridization mixture consisting of 3.5 μM 37mer primer, 1X ThermoPol Buffer and 0.5 M NaCl were incubated with the template spots on the slide for 30 min at 65°C, and the hybridization process was repeated three times to achieve higher efficiency. For the comparison of the different slide surfaces, the same optimized hybridization condition was applied for both slides.

After a series of tests, the procedures for hybrid SBS on the linear template was established as follows. For the incorporation step, a 10 μl reaction mixture composed of solution A consisting of 3'-*O*-N_3-dCTP (1.4 μM), 3'-*O*-N_3-dTTP (1.5 μM), 3'-*O*-N_3-dATP (1.8 μM) and 3'-*O*-N_3-dGTP (1.3 μM), and solution B consisting of ddCTP-N_3-Bodipy-FL-510 (120 nM), ddUTP-N_3-R6G (160 nM), ddATP-N_3-ROX (40 nM) and ddGTP-N_3-Cy5 (80 nM), 1 unit of 9°N DNA polymerase, 20 nmol of $MnCl_2$ and 1X Thermopol II reaction buffer was incubated with linear template on the slide at 55°C for 15 min. A capping step was followed to synchronize any unincorporated templates. An extension solution consisting of 60 pmol each of 3'-*O*-N_3-dTTP, 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP, and 97.5 pmol of 3'-*O*-N_3-dCTP, 1 unit of 9°N DNA polymerase, 20 nmol of $MnCl_2$ and 1X Thermopol II reaction buffer was added to the same spot and incubated at 55°C for 15 min. After washing the chip with SPSC buffer (sodium phosphate and sodium chloride buffer with 0.2% Tween) for 1 min, the surface was rinsed with dH_2O, dried briefly and then analyzed under the 4-color scanner for the

fluorescent signal. To perform the cleavage, 10 µl of TCEP (100 mM, pH 9.0) was added to the spot and incubated at 55^{o}C for 10 min. After washing, the chip was scanned again to compare the intensity of fluorescence after cleavage with the original fluorescence intensity. This process was followed by the next cycle of polymerase extension using the 3'-*O*-N_3-dNTP/ddNTP-N_3-fluorophores, fluorescence detection, synchronization, washing, and cleavage processes performed as described above. To obtain DNA sequencing data, the SBS cycles were repeated multiple times using the various ratios of solution A and solution B. The volumes of solution A and B in each SBS cycle were adjusted to achieve relatively even fluorescence signals (Table 6.2).

Table 6. 2 Volumes of solution A and B in each SBS cycle during hybrid SBS

SBS cycle	Solution A (µl)	Solution B (µl)	SBS cycle	Solution A (µl)	Solution B (µl)
1st	7.3	0.2	14th	6.7	0.8
2nd	7.3	0.2	15th	6.6	0.9
3rd	7.3	0.2	16th	6.3	1.2
4th	7.3	0.2	17th	5.7	1.8
5th	7.3	0.2	18th	5.2	2.3
6th	7.3	0.2	19th	4.7	2.8
7th	7.2	0.3	20th	4.0	3.5
8th	7.2	0.3	21st	3.5	4.0
9th	7.1	0.4	22nd	3.0	4.5
10th	7.1	0.4	23rd	2.5	5.0
11th	6.9	0.6	24th	2	5.5
12th	6.8	0.7	25th	0.5	7.0
13th	6.8	0.7			

6.3.2 Primer resetting and walking

The final protocol used for denaturing the extended primer from the first round was

to incubate the chip in a solution consisting of 50 mM Tris-HCl buffer (pH 7.5), 80% formamide and 0.1% SDS at 65^{o}C for 30 min, twice. The efficiency of the denaturation was checked under the scanner, with complete denaturation expected to lead to complete loss of the fluorescent signal. Slightly different conditions were tested, such as varying the incubation temperature and time, before the final protocol was settled upon.

6.3.2.1 Primer walking using three natural nucleotides

Various commercially available DNA polymerases (9^{o}N, Pfu, Tgo and Thermo Sequenase) were tested for their fidelity and accuracy in the single base extension reaction using dTTP. Each extension reaction was carried out with a mixture consisting of a self-priming DNA template SP26T (5'-**GTCA**GCGCCGCGCCTTG-GCGCGGCGC-3', 40 pmol), 200 pmol of dTTP, 1X enzyme reaction buffer, and 1 unit of the DNA polymerase. For 9^{o}N polymerase, 20 nmol of $MnCl_2$ was used as the cofactor. The reaction mixture was incubated at 65^{o}C for 2 min. After ethanol precipitation and desalting with ZipTips (Millipore), the extension products were analyzed by MALDI-TOF MS.

Based on the single base extension results, Thermo Sequenase was chosen for primer walking with three natural nucleotides in solution. The three-base extension reaction consisted of 40 pmol SP26T, 200 pmol each of three natural nucleotides (dTTP, dGTP and dATP), 1X Thermo Sequenase reaction buffer and 1 unit of Thermo Sequenase with incubation at 65^{o}C for 15 min. The resulting products were analyzed by MALDI-TOF

MS as described above. Due to the presence of mis-incorporation products, the same extension was further tested with the addition of 2 μl PyroE solution (Pyrosequencing enzyme mixture, Roche), the products of which also underwent MALDI-TOF MS analysis.

To most mimic the primer walking, linear templates instead of self-priming templates were tested. 20 pmol of the linear template, Exon8, (5'-GAA GGA GAC ACG CGG CCA GAG AGG GTC CTG TCC GTG TTT GTG CGT GGA GTT CGA CAA GGC AGG GTC ATC TAA **TGG TGA TGA GTC CTA TC**C TTT TCT CTT CGT TCT CCG T-3', the letters in bold indicating the primer hybridization site), 60 pmol of a 17mer primer (P5163, 5'-GAT AGG ACT CAT CAC CA-3'), 200 pmol each of three natural nucleotides (dTTP, dCTP, dGTP), 1X Thermo Sequenase reaction buffer, 1 unit of Thermo Sequenase and various amount of PyroE solution were used for the extension reactions. The reactions were performed for 10 cycles of 94^{o}C for 20 s, 45^{o}C for 30 s and 68^{o}C for 40 s, and the products were analyzed using MALDI-TOF MS.

6.3.2.2 Primer walking using three natural nucleotides and one reversible nucleotide terminator

The MALDI-TOF based analysis was first performed to verify the feasibility of this strategy in solution. 20 pmol of Exon8, 40 pmol of P5163, 100 pmol each of three natural nucleotides (dTTP, dCTP, dATP), 500 pmol of 3'-O-N_3-dGTP, 20 nmol of $MnCl_2$, 1X Pol II buffer, and 1 unit of 9^{o}N enzyme, were combined and underwent 20 cycles of 94^{o}C for

20 s, 45°C for 30 s and 68°C for 40 s, and the products were analyzed using MALDI-TOF MS.

Primer walking was then carried out by adding 10 μl walking solution consisting of 200 pmol of each dNTP (dATP, dCTP, dTTP), 550 pmol 3'-O-N_3-dGTP, 1 unit of 9°N polymerase, 20 nmol of $MnCl_2$ and 1X ThermoPol II buffer and incubated at 55°C for 15 min. To synchronize the unextended templates, two capping steps were carried out. First, a 10 μl solution consisting of 100 pmol 3'-*O*-N_3-dGTP, 1 unit of 9°N polymerase, 20 nmol of $MnCl_2$ and 1X ThermoPolII buffer were added and incubation proceeded at 65°C for 20 min. Then the second capping step was carried out with a 10 μl solution consisting of 100 pmol of each ddNTP (ddATP, ddCTP, ddGTP, ddTTP), 2 unit of 9°N enzyme, 20 nmol of $MnCl_2$ and 1X ThermoPolII buffer at 55°C for 15 min. The cleavage reaction with 10 μl TCEP at 55°C for 15 min was then carried out to recover the 3'-OH group for the next walking cycles.

Direct hybrid SBS was first carried out after 3 walking cycles without first round sequencing. Following the same procedures as described in the hybrid SBS on linear template section, the volumes of solution A and B in each SBS cycle were adjusted to obtain relatively even fluorescence signals (Table 6.3).

To fully implement the primer walking strategy, in collaboration with Lin Yu, the complete process of SBS integrated primer walking was performed by using SBS chemistry, with dNTP-N_3-fluorophores/3'-*O*-N_3-dNTPs (CF-NRTs/NRTs). After the first round of sequencing, 30 bases were identified. The extended primer was denatured away,

and fresh primer re-annealed to the template. Four walking cycles were performed to extend the primer to the site where the first round sequencing stopped. A second round of sequencing was then carried out to double the original read-length.

Table 6.3 Volumes of solution A and B in each SBS cycle during hybrid SBS after three walking cycles

SBS cycle	Solution A (μl)	Solution B (μl)	SBS cycle	Solution A (μl)	Solution B (μl)
1st	7.3	0.2	14th	6.4	1.1
2nd	7.2	0.3	15th	6.2	1.3
3rd	7.2	0.3	16th	5.9	1.6
4th	7.2	0.3	17th	5.6	1.9
5th	7.1	0.4	18th	5.4	2.1
6th	7.1	0.4	19th	5.3	2.2
7th	7.1	0.4	20th	5.1	2.4
8th	7.1	0.4	21st	4.9	2.6
9th	7.1	0.4	22nd	4.6	2.9
10th	7.0	0.5	23rd	4.1	3.4
11th	6.9	0.6	24th	3.8	3.7
12th	6.8	0.7	25th	3.0	4.5
13th	6.6	0.9	26th	1.5	6.0

6.4 Conclusion

In this chapter, we developed a novel primer walking strategy to increase the read-length of DNA sequencing by synthesis. The combination of three natural nucleotides and one NRT effectively regulated the primer walking: the primer extension temporarily paused when the NRT was incorporated, and resumed after removing the 3' capping group to restore the 3'-OH group. The advantage of using this nucleotide mixture is that DNA polymerase could function with higher fidelity and accuracy,

templates were synchronized with the NRT serving as the last base in each walking cycle, and the primer resetting had the ability to restore all the templates after the denaturation step. We have successfully demonstrated the integration of this primer walking strategy into the sequencing by synthesis platform, and were able to obtain a total read-length of 53 bases, nearly doubling the read length of the previous sequencing. This primer walking strategy could not only be used in sequencing by synthesis with CF-NRTs or hybrid sequencing by synthesis, but also has universal application for other sequencing methods. Though much optimization is still needed, it is believed that the implementation of this primer strategy with the automated platforms could bring us closer to the $1000 genome goal.

References

1. Metzker ML. Emerging technologies in DNA sequencing. *Genome Research*, **2005**, *15*, 1767-1776.

2. Welch MB, Burgess K. Synthesis of fluorescent, photolabile 3'-O-protected nucleoside triphosphates for the base addition sequencing scheme. *Nucleosides Nucleotides*, **1999**, *18*, 197-201.

3. Lu G, Gurgess K. A diversity oriented synthesis of 3'-O-modified nucleotide triphosphates for DNA "sequencing by synthesis". *Bioorganic & Medicianl Chemistry Letters*, **2006**, *16*, 3902-3905.

4. Pelletier H, Sawaya MR, Kumar A, Wilson SH, Kraut J. Structures of ternary complexes of rat DNA polymerase β, a DNA template-primer, and ddCTP. *Science*, **1994**, *264*, 1891-1903.

5. Rosenblum BB, Lee LG, Spurgeon SL, Khan SH, Menchen SM, Heiner CR, Chen SM. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Research*, **1997**, *25*, 4500-4504.

6. Zhu Z, Chao J, Yu H, Waggoner AS. Directly labeled DNA probes using fluorescent nucleotides with different length linkers. *Nucleic Acids Research*, **1994**, *22*, 3418-3422.

7. Ju, J., Li, Z., Edwards, J., Itagaki, Y. **2003**, *US Patent* 6,664,079.

8. Seo TS, Bai X, Kim DH, Meng Q, Shi S, Ruparel H, Li Z, Turro NJ, Ju J. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proceedings of the National Academy of Science of the United States of the America*, **2005**, *102* (17), 5926-5933.

9. Bi L, Kim DH, 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. *Journal of the American Chemistry Society*, **2006**, *128(8)*, 2542-2543.

10. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Science of the United States of the America*, **2006**, *103(52)*, 19635-19640.

11. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano Marma M, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Science of the United States of the America*, **2008**, *105(27)*, 9145-9150.

12. Lin Y. Novel strategies to increase read length and accuracy for DNA sequencing by synthesis. Dissertation (Ph.D.), Columbia University, **2010**

13. Kircher M, Kelso J. High throughput DNA sequencing - concepts and limitations. *Bioessays*, **2009**, *32*, 524-536.

Chapter 7 Exploration of an "Emulsion PCR-Bead-on-Chip" Approach to Improve the Throughput of Sequencing by Synthesis

7.1 Introduction

There is an ever-growing demand for DNA sequencing, which has spawned the second generation DNA sequencing industry, mostly represented by DNA sequencing by synthesis (SBS) technologies, including pyrosequencing, sequencing by ligation, and SBS with fluorescent nucleotide reversible terminators. The prospective goal of the $1000 genome has driven the optimization of these technologies from different perspectives, yet the typically short read-length remains a major challenge. To address this problem, in Chapter 6, we introduced a primer walking strategy to extend the read-length of SBS. Nevertheless, the capacity of SBS could also be improved by the number of templates sequenced at a time, since the full read-length of SBS is determined by the product of single fragment read-length and number of sequenced fragments. Current sequencing by synthesis technologies with cleavable fluorescent reversible terminators coupled with the use of a high density DNA array present tremendous potential for ultra-high throughput DNA sequencing. However, to realize massively-parallel sequencing and meet the challenge of the $1000 genome, further innovations are needed.

Sequencing on beads, a platform based on current sequencing methods and emulsion

PCR technology, appears very promising. The first advantage of this system is that it circumvents the need for generating a clone library, thus reducing the complexity of the sequencing process, time, need for tracking and storage requirements, and also avoiding inherent biases of vectors that might decrease the replication fidelity and sequencing accuracy.[1] This cell-free system was achieved by emulsion PCR, a method that uses water-in-oil emulsions to separately and individually amplify millions of DNA templates in miniaturized compartments each containing a single bead. After amplification, the DNA molecules bound to the beads provide excellent templates for high-throughput sequencing; since each bead bears thousands of copies of the same PCR product, the signal-to-noise ratio obtained by hybridization or enzymatic assay is extremely high.[2, 3] Because different beads contain the products of different compartmentalized PCR reactions, thousands of different templates could be sequenced at the same time after immobilizing ePCR beads in a single layer that allows enzymatic manipulation and imaging.

One successful example is 454-sequencing developed and commercialized by 454 Life Science (Brandford, CT), which integrated the pyrosequencing method, emulsion PCR and their PicoTiterPlate (PTP) platform. ePCR beads were deposited into wells of a fibre-optic slide (picotiter plate), in which most wells contain a single bead. Only picolitre-scale volumes are needed for each well. In one four-hour run, 25 million bases were sequenced at 99% or better accuracy.[4,5] It would seem that such an approach could reach the ultra-massively parallel sequencing goal; however, inherent limitations of

pyrosequencing, like ambiguity in homopolymer regions, still present bottlenecks for the further development of this technology.

Church's group[6] also integrated the "sequencing by ligation" method with sequencing on beads, through immobilizing ePCR beads in an acrylamide gel to form a disordered, monolayered bead array. As they claimed, 1 billion 1 μm beads can potentially be fit into the area of a standard microscope slide. This system was applied to resequence an evolved strain of *Escherichia coli* at less than one error per million consensus bases. At the time, the major disadvantage of this technique was the short read length: only six (seven) continuous accurate bases from the ligation junction in 5'to3' direction, with a total of 13bp per tag and 26bp per amplicon, though this has been improved somewhat in the SOLiD system commercialized by Life Technologies.

DNA sequencing on beads has provided a paradigm shift in sequencing capacity, so it is worthwhile to integrate sequencing by synthesis with cleavable fluorescent reversible terminators (CF-NRTs) into the bead array system, since SBS with CF-NRTs has already shown its great potential for sequencing. In addition, the ability to capture and utilize longer templates on beads than directly on surfaces for bridge PCR offers the opportunity to better take advantage of mate pair approaches without the need for cell-based cloning. This might achieve ultra-massively parallel sequencing with excellent accuracy and low cost. With the aim of developing an emulsion-PCR-bead-on-chip approach for SBS in mind, we tested different prerequisite conditions, including DNA attachment to beads, 4-color SBS on beads, and the formation of emulsions. The

resulting accurate sequencing of DNA templates on beads indicates that it has the potential to improve the throughput of sequencing by synthesis.

7.2 Experiment Rationale and Overview

In this chapter, we explore the applicability of the emulsion-PCR-bead-on-chip approach for improving the throughput of sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators (CF-NRTs)/nucleotide reversible terminators (NRTs). The fundamental rationale of this idea is to transfer arrayed DNA sequencing into bead arrayed DNA sequencing, considering the advantages and higher throughput of the bead-based approach. In general, as shown in Fig. 7.1, universal forward DNA primers are first immobilized onto the microbeads. Emulsion PCR is then carried out with a mixture of DNA templates, primer attached microbeads, an external reverse primer and deoxynucleotides in a water-in-oil droplet, resulting in beads bearing an amplified colony of the same DNA template. The microbeads carrying DNA templates were attached to the surface to form high density arrays for sequencing by synthesis analysis. Here, each microbead represents one template, while in the conventional microarray, a cluster of DNA molecules on the chip represents one template. Theoretically, a bead array provides many more templates available for sequencing. However, to implement this strategy, the first step is to attach DNA onto the beads efficiently and stably to undergo PCR and SBS cycles. The next key step is actually to perform SBS with CF-NRTs/NRTs on beads. As our lab has successfully utilized the

combination of 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) and NRTs (3'-*O*-N_3-dCTP, 3'-*O*-N_3-dTTP, 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP) (Fig. 7.2) to perform SBS on a chip, these two sets of nucleotides were used.[7] To thoroughly investigate the requisite conditions, a self-priming DNA template was tested both in the DNA conjugation and sequencing experiments, with the amount of incorporation on the bead surface as well as the amount of specific SBS signal used to test the stability of the DNA molecules on the bead surface.

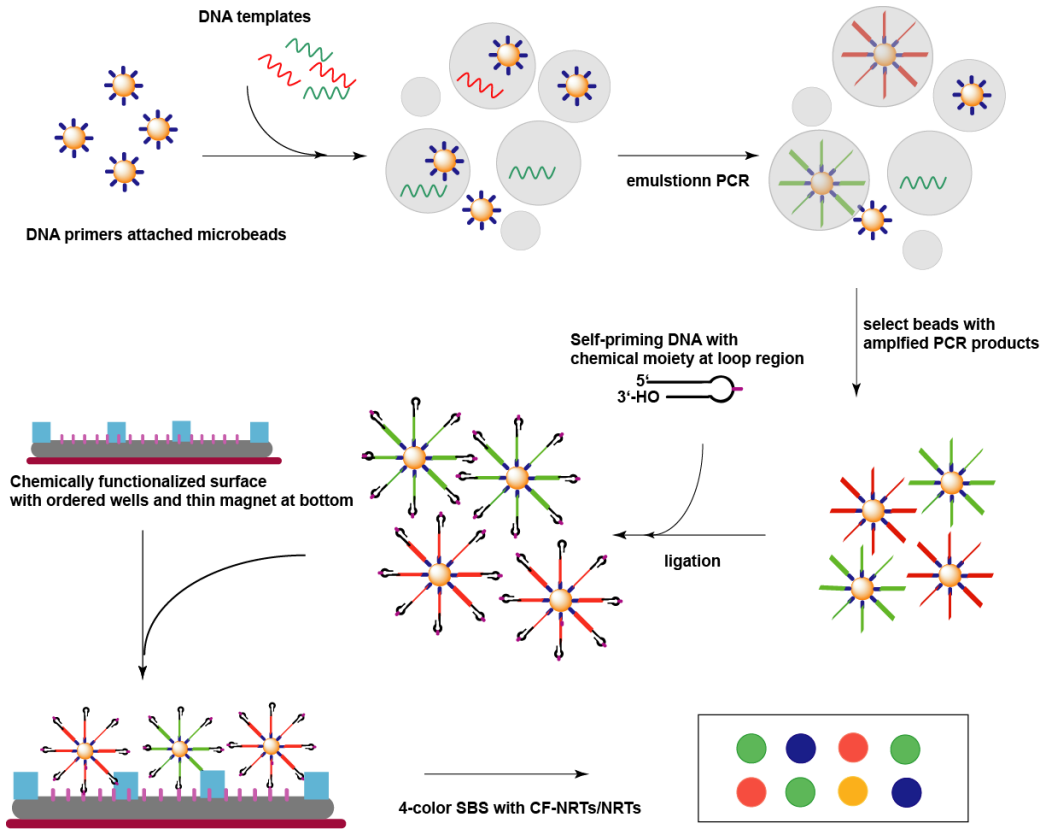


Fig. 7.1. Scheme of SBS integrated emulsion PCR-beads-on-chip for ultra-massively parallel sequencing

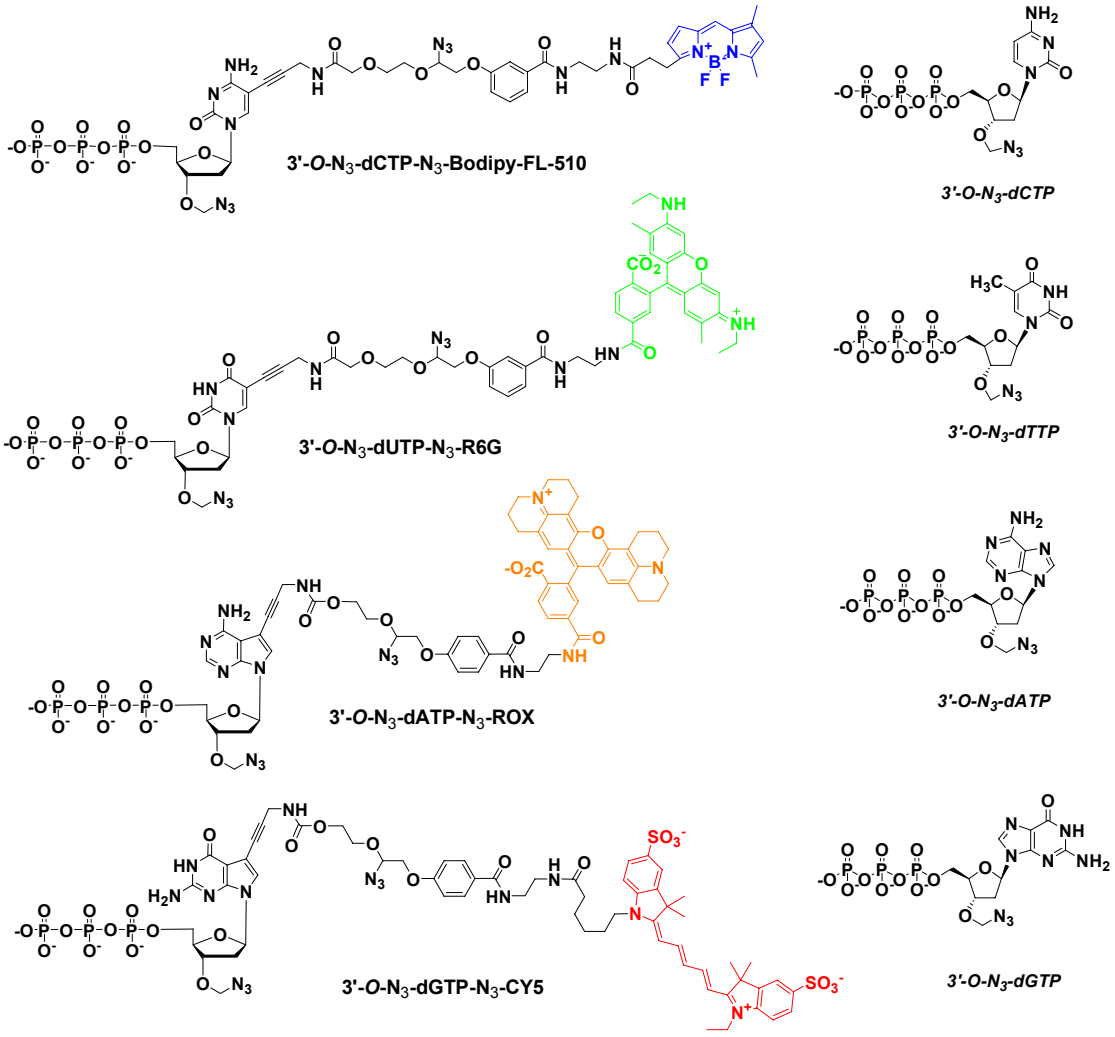


Fig. 7.2. Chemical structures of 3'-*O*-N_3-dCTP-N_3-Bodipy-FL-510, 3'-*O*-N_3-dCTP, 3'-*O*-N_3-dUTP-N_3-R6G, 3'-*O*-N_3-dTTP, 3'-*O*-N_3-dATP-N_3-ROX, 3'-*O*-N_3-dATP, 3'-*O*-N_3-dGTP-N_3-Cy5 and 3'-*O*-N_3-dGTP

7.3 Results and Discussion

7.3.1 Covalent attachment of DNA onto the beads

Magnetic beads were chosen based on their ease of separation and manipulation for future applications. To obtain a good sequencing signal, it is very important to maximize the amount of DNA template immobilized onto the beads and minimize the background

signal caused by nonspecific binding of dye-linked nucleotides. 3'-FAM labeled oligonucleotides were therefore used in the coupling test. After subtraction of signal produced by a negative control (absence of a chemical required for coupling), a higher fluorescent signal was taken to indicate more DNA attached to the beads. PEG (polyethylene glycol) grafting has been widely used to reduce non-specific binding and hence improve the signal-to-noise ratio.[8-11] Introduction of PEG chains onto the beads can be achieved in two ways: either the bead can be coated with PEG first and then coupled to DNA, or be directly coupled to PEG-DNA conjugates. Different classes of beads were tested, via related chemistry methods, as shown briefly in Table 7.1. Results indicated that one-step 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) coupling on Chemagen PVA beads produced the highest fluorescence signal and hence had the largest amount of coupled DNA (part of the fluorescence imaging data was shown in Fig. 7.3). The mechanism of the EDC coupling chemistry is illustrated in Fig. 7.4. EDC activates the carboxyl group on the bead surface to form an active intermediate, O-acylisourea, which then reacts with an amino group at the 3' end of the DNA to form coupled products. As the intermediate is very easy to degrade, the addition of fresh EDC improved the coupling efficiency.

7.3.2 Biological affinity attachment of DNA onto the beads

The interaction between streptavidin and biotin is known to be one of the strongest non-covalent bonds, exhibiting a dissociation constant of about 1.3×10^{-15}M. The

bio-specificity is similar to antibody-antigen or receptor-ligand recognition, but displays much higher affinity constants.[12] Streptavidin coated beads were chosen to take advantage of this strong affinity and the simple coupling procedure. It was found that coupling biotinylated oligonucleotides with just a single 5'-biotin group resulted in dissociation from the beads during temperature cycling in PCR, whereas oligonucleotides labeled with dual biotin groups at their 5'end were stable to cycling.[2] Hence dual-biotin labeled DNA was used to enhance the stability. Again, to reduce non-specific binding, dual-biotin-PEG-DNA was used.

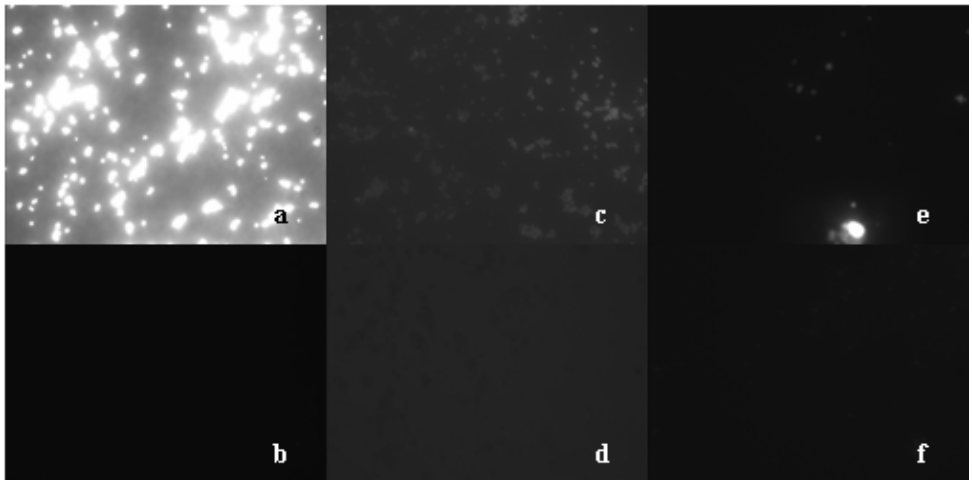


Fig. 7.3. Fluorescence imaging data on Chemagen PVA beads: a, one-step EDC coupling; c, two-step EDC coupling with sulfo-NHS; e, PEG coated first and then coupled to DNA; b, d, f are the respective negative controls.

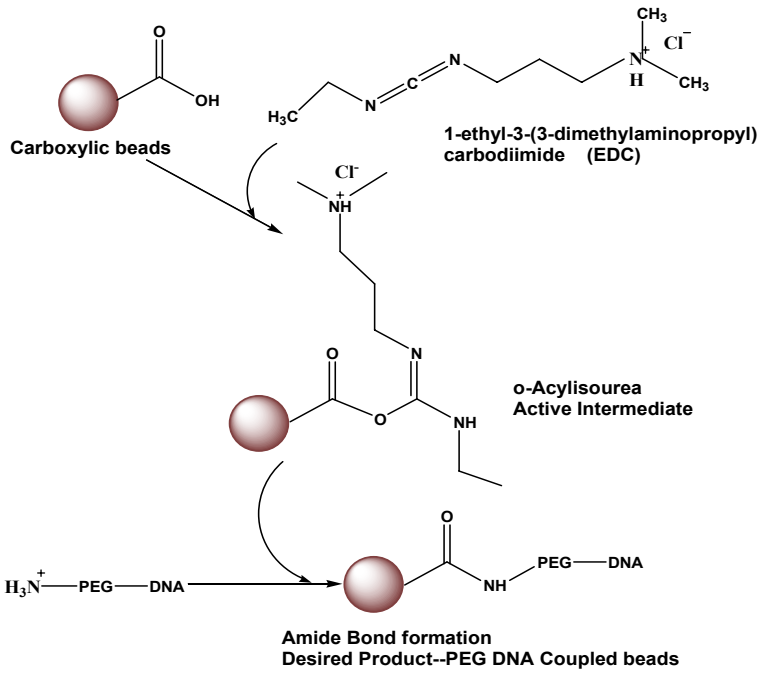


Fig. 7.4. Mechanism of EDC coupling

7.3.3 Nucleotide incorporation studies on bead surfaces

7.3.3.1 Pyrosequencing

To ensure that self-priming templates on beads could incorporate nucleotides and that enough templates were immobilized on the beads, the coupled PVA beads were first tested by pyrosequencing. For test purposes, the set of nucleotides and enzymes was added only once, which meant that only the first few bases could be sequenced. As shown in Fig. 7.5, the first 7 bases were correctly identified, which demonstrated the feasibility of using these coupled beads for SBS.

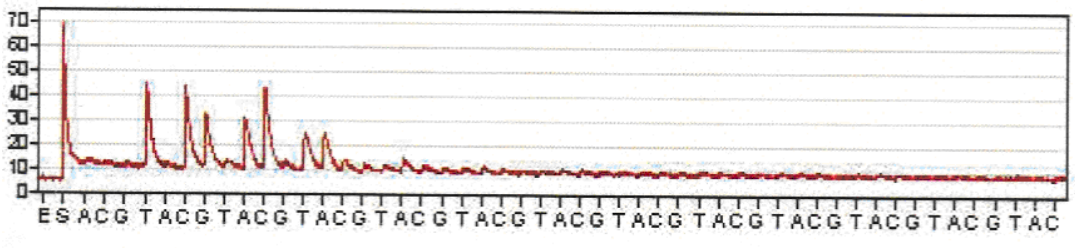


Fig. 7. 5. Pyrosequencing data

7.3.3.2 Incorporation of fluorescence labeled ddNTPs and natural dNTPs

To further confirm the incorporation capability of immobilized DNA templates and the quality of the fluorescence signal, a set of commercially available fluorescently labeled terminators (Cyanine5-ddGTP, R6G-ddUTP, ROX-ddATP and Fluorescein-12-ddCTP) were used in another incorporation test. As shown in the scheme (Fig. 7.6A), natural nucleotides were first used to extend the self-priming primer to one base upstream of the interrogated base, then single base extension with fluorescently labeled terminators was carried out to identify the base. With the exception of the second base position, DNA templates on the PVA beads could incorporate modified bases accurately, resulting in a high fluorescent signal with low background, as shown in Fig. 7.6 B. The inaccuracy of the second base was probably caused by mis-incorporation by polymerase: since only dUTP was added, without competition by other nucleotides to move to the second base position, natural dUTPs could fill the second position as well leading to a signal due to the third base.

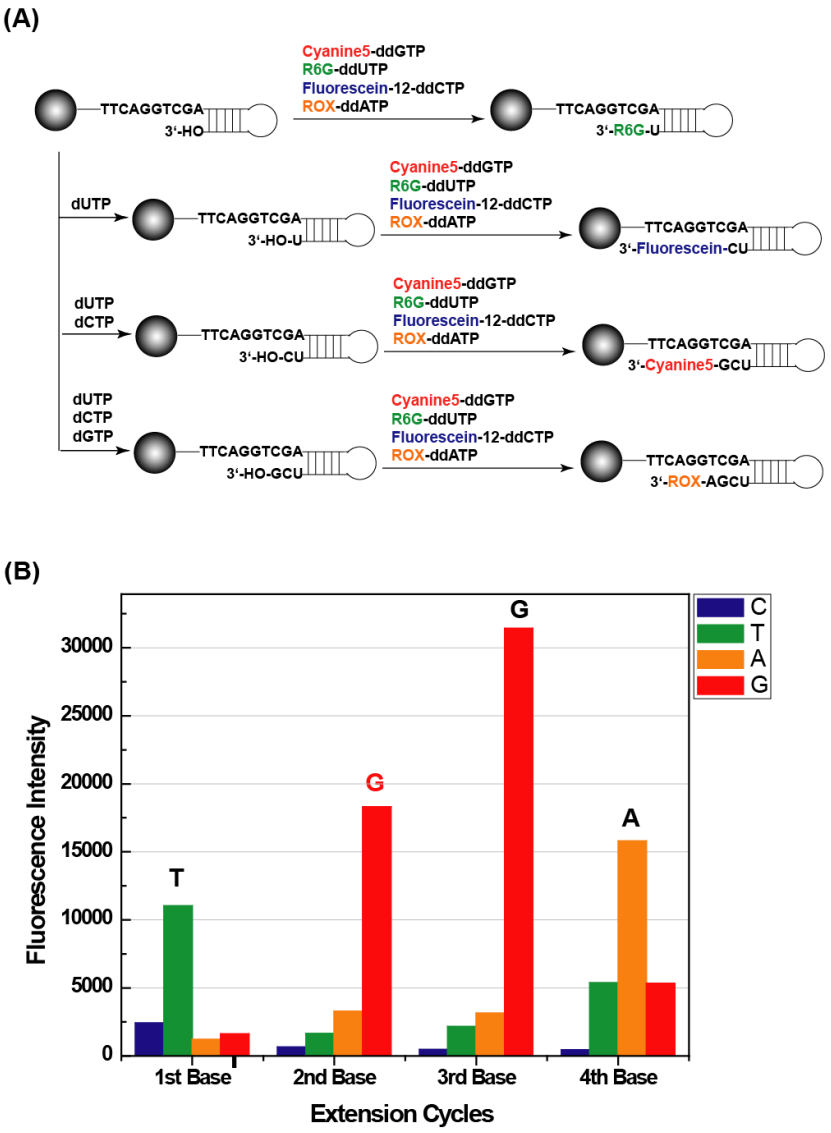


Fig. 7.6. (A) Scheme of incorporation test with ddNTP-dyes and dNTPs; (B) Sequence data achieved by using ddNTP-dyes/dNTPs

7.3.3.3 Sequencing by synthesis on beads using CF-NRTs/NRTs

As discussed above, the key step to implement the ePCR-bead-on-chip approach for sequencing by synthesis is to demonstrate the feasibility of performing SBS sequencing with CF-NRTs/NRTs on beads, as the amount of DNA templates and the surface

conditions would be different from that on a conventional DNA chip. SBS with CR-NRTs/NRTs was carried out on bead surfaces. It was demonstrated that this bead based sequencing strategy was able to identify continuous sequence with high accuracy, both on PVA carboxylic beads (Fig. 7.7) and streptavidin beads (Fig. 7.8).

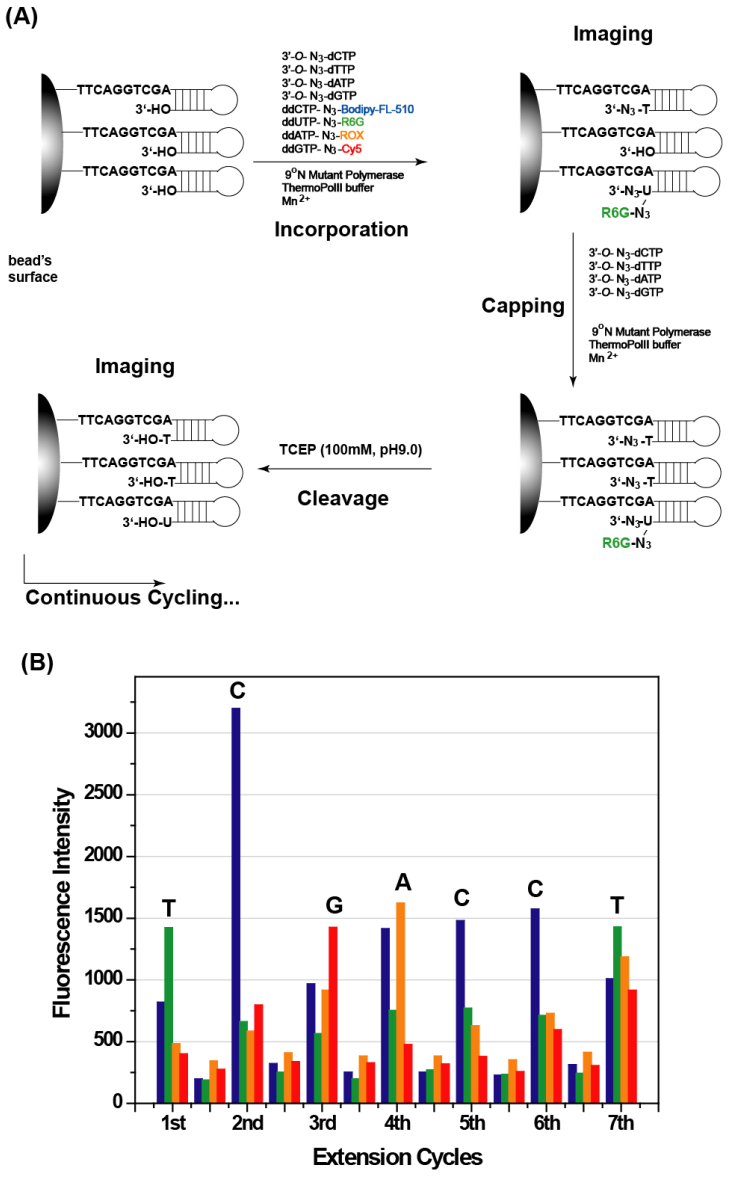


Fig. 7.7. (A) Scheme of SBS using 3'-*O*-N_3-dNTPs and 3'-*O*- N_3-dNTP-N_3-fluorophores; (B) 4-color SBS data on PVA carboxylic beads.

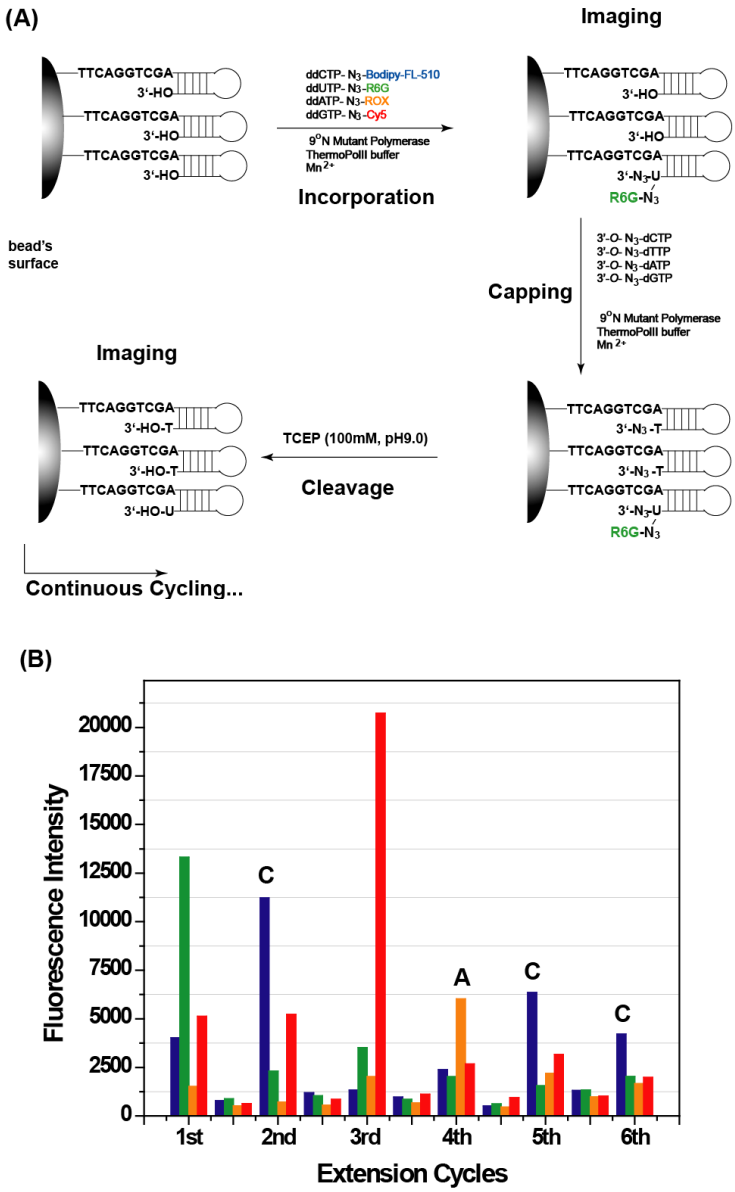


Fig. 7.8. (A) Scheme of SBS on Streptavidin beads with an alternative strategy; (B) 4-color SBS data on Streptavidin beads.

7.3.4 PCR on beads in solution and beads in emulsion

As an initial exploration of ePCR, PCR on beads without emulsion generation was first tested. Taking PVA carboxylic beads as the example, the forward primer attached beads were mixed into PCR solutions consisting of DNA template, reverse primer,

nucleotides and enzyme to carry out direct PCR on beads. The PCR result was verified by performing direct Sanger sequencing on the beads. If there are templates growing on the beads, Sanger sequencing should be able to retrieve this sequence information. As shown in Fig. 7.9, the Sanger sequencing was able to obtain information from DNA molecules on beads; the negative result on the supernatant after the last wash essentially excluded the possibility of non-specific binding of solution generated DNA template (data not shown). This demonstrated the possibility of performing ePCR on these particular beads.

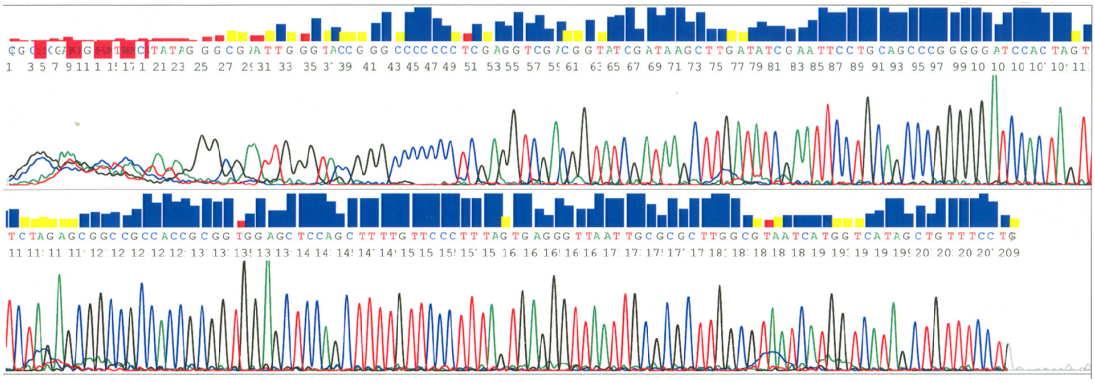


Fig. 7.9. Sanger sequencing result for PCR products on beads. The associated bar graphs indicate the quality of the sequence: good sequence is indicated in blue.

The conditions for generating water-in-oil droplets (emulsions) were also tested, including the ratio of water and oil phase, the concentration of bead suspension, and the mixing speed. As shown in Fig. 7.10, for both 5 μm and 8 μm beads, we have successfully generated stable single bead resident droplets.

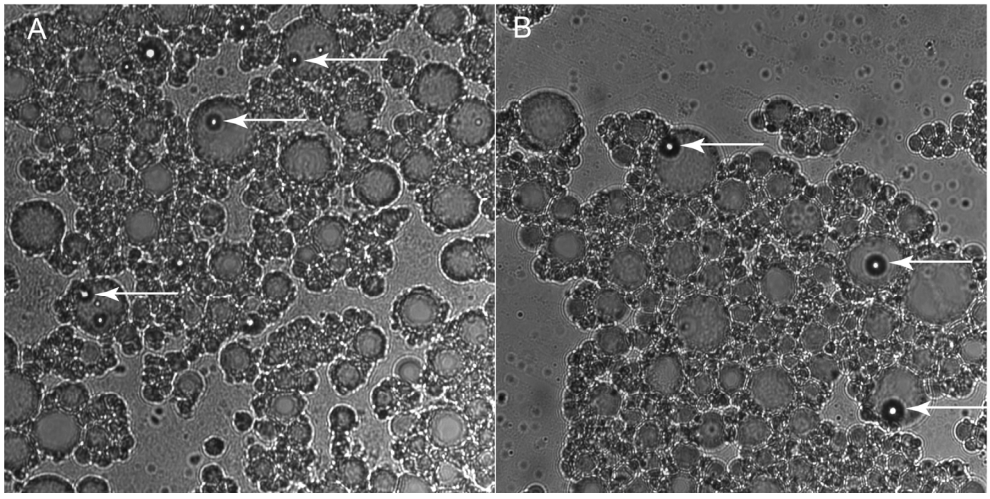


Fig. 7.10. Emulsion formation containing 5 μm beads (A) and 8 μm beads (B).

7.4 Materials and Methods

General information. All the chemicals were obtained from Sigma-Aldrich unless specified. PVA Carboxylic beads (M-PVA C22) were purchase from Chemagen Biopolymer-Technologie AG (Germany), and Streptavidin Coated COMPEL^{TM} Magnetic beads (8.03 μm, 5 μm) were from Bangs Laboratories, Inc. (Fishers, IN). Fluorescence detection was carried out either under a four-color ScanArray Express Scanner (PerkinElmer Life Sciences, Boston, MA) or Canon Fluorescence Microscope.

7.4.1 DNA attachment to beads

Before finalizing the chemical coupling protocol, different coupling methods were tested, the details of which are shown in Table 7.1. 3' FAM labeled oligonucleotides were used for coupling, and the products were checked under the fluorescence microscope. It turned out that one-step EDC coupling on PVA beads gave the best results. Chemagen PVA beads consist of a matrix of polyvinyl alcohol, highly carboxyl modified, with a size

distribution from 1.0-3.0 μm. The protocol was described as follows. 40 μl of PVA bead suspension (50 mg/ml) was washed twice with 0.01N NaOH by mixing using a vortex at room temperature. The beads were then washed three times with deionized water and 2-(N-morpholino)ethane sulfonic acid (MES) buffer (0.1 M, pH 4.5) in the same manner and the supernatant was discarded. 20 μl of EDC/MES (~1 M) solution containing 750 μM 5'NH_2-PEG-DNA (Self-priming template ready for sequencing: 5'-NH_2-PEG-TTC AGGTCGACTTAAGGCGCTTGCGCCTTAAG-3') was first added to the beads. Then the solution was mixed using a rotator for 2.5 hours at room temperature, with addition of 10 μl EDC/MES (1M) every 30 min. The beads were washed three times with SPSC buffer (sodium phosphate, sodium chloride buffer with 0.2% Tween) three times with deionized water, and finally suspended in 200 μl of water and stored at 4^oC. Therefore the concentration of the coupled bead solution was ~10 mg/ml. In the case of PCR on beads, the M13 forward primer (5'-CACTCGCATGGG-TAAAACGACGGCCAG-3') was attached to the beads.

For Streptavidin bead coupling, 20 μl of streptavidin coated magnetic beads were washed with Binding/Wash Buffer (20 mM Tris pH 7.5, 1 M NaCl, 1 mM EDTA, 0.0005% Triton-X100) three times and suspended in 20 μl Binding/Wash Buffer, after which 4 μl of dual-biotin-PEG-DNA (50 μM in H_2O) (5'-dual-biotin-PEG-TTC AGG TCG ACT TAA GGC GCT TGC GCC TTA AG-3') was added. The solution was mixed using a rotator for 30min at room temperature. After binding, the beads were washed twice with Binding/Wash Buffer, three times with SPSC, three times with H_2O, and

finally suspended in 20 µl 1X ThermoPol II (Mg-free) reaction buffer (10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl, 0.1% Triton X-100, pH 8.8, Biolabs Inc.) and stored at 4°C.

Table 7.1 Coupling experiments on different classes of beads

Class of Beads	Dynabeads® M-270 Carboxyl	Chemagen M-PVA C22 Carboxyl	Chemagen M-PVA C22 Carboxyl	Chemagen M-PVA C22 Carboxyl	Dynabeads® M-270 Amine	Dynabeads® M-270 Epoxy
Step1 ligand	5'-NH_2-DNA-FAM	5'-NH_2-PEG-DNA-FAM	NH_2-PEG-DNA-FAM	NH_2-PEG-COOH-t-butyl	HOOC-PEG-N_3	NH_2-PEG-COOH
Step1 coupling reagent	EDC (one-step coupling)	EDC (one-step coupling)	EDC & sulfo-NHS (two-steps: activation, coupling	PyBop & DIPEA	EDC (one-step) or PyBop & DIPEA	.1M Sodium phosphate buffer (pH7.4) & Ammonium sulfate
Step2 ligand				5'-NH_2-DNA-FAM	5'-Alkyne-DNA-FAM	5'NH_2-DNA-FAM
Step2 coupling reagent				EDC (one-step coupling)	Vitc, TBTA & Cu^+ ("Click Chemistry")	EDC (one-step coupling)
Negative control	w/o EDC	w/o EDC	w/o EDC	w/o EDC	w/o Cu^+	w/o EDC
Relative fluoresc-ence intensity	weak	**very strong**	weak	weak	no signal	no signal

Note: When NH_2-PEG-COOH-t-butyl was used, a deprotection step with TFA was applied to remove the t-butyl group and generate the carboxyl group before the next coupling step. EDC: 1-Ethyl-3-(3-dimethylaminopropyl)-carbodiimide; sulfo-NHS: N-Hydroxy sulfo-succinimide; PyBop: benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluoro phosphate; DIPEA: N, N'-Diisopropylethylamine; TFA: Trifluoroacetic acid; Vc: ascorbic acid; TBTA: tris[(1-benzyl-1*H*-1,2,3-triazol-4-yl)methyl]amine;

7.4.2 Nucleotide incorporation test on beads

To confirm the incorporation capability of immobilized DNA templates, one step pyrosequencing was performed. 5 μl of beads were suspended in the pyrosequencing plate with reaction buffer, and detected in a 96PSQ Pyrosequencer (Biotage, Uppsala, Sweden) under standard operating conditions. For test purposes, the set of reagents was added only once, which meant that only the first few bases could be sequenced.

To further confirm the incorporation capability of immobilized DNA templates and the quality of the fluorescence signal, a set of commercially available dye labeled terminators, Cyanine5-ddGTP, R6G-ddUTP, Fluorescein-12-ddCTP and ROX-ddATP (Perkin Elmer, Boston, MA), were used in another test. As shown in Figure 7.6, this test was designed to detect the first four bases. 4 μl of coupled PVA beads were washed three times with 1X Thermo Sequenase buffer and divided into four parts, each aliquot of which was used for the detection of one base. For the first base, the beads were added to 10 μl "detection solution" consisting of Cyanine 5-ddGTP (1 μM), R6G-ddUTP (1 μM), Fluorescein-12-ddCTP (1 μM), ROX-ddATP (1 μM), 1.5 unit of Thermo Sequenase DNA^{TM} Polymerase (GE Healthcare, UK) and 1X Thermo Sequenase Buffer, and incubated for 15 min at 65^{o}C. For the detection of the second base, the beads were added to 10 μl "extension solution" consisting of dUTP (2 μM), 1.5 unit of Thermo Sequenase DNA Polymerase and 1X Thermo Sequenase Buffer, and incubated for 5 min at 65^{o}C to fill the first base position, and then 10 μl "detection solution" was added, reacting for 15 min at 65^{o}C. The third and fourth bases were detected in a similar manner:

"extension solution" first to fill in previous bases and then the "detection solution". The scheme of this method is shown in Fig.7.6 A. Before scanning, the beads were washed with SPSC and water several times and then suspended in water, 0.2 μl of which was spotted on a glass slide. After the spot dried, it was analyzed in a four-color scanner.

7.4.3 Sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators on beads

After washing PVA beads with 1X ThermoPolII buffer three times, 12.5 μl of a solution consisting of 8 pmol 3'-*O*-N_3-dATP-N_3-ROX, 1 pmol 3'-*O*-N_3-dGTP- N_3-Cy5, 2 pmol 3'-*O*-N_3-dUTP-N_3-R6G, 2 pmol 3'-*O*-N_3-dCTP-N_3-bodipy-FL-510, 8 pmol 3'-*O*-N_3-dATP, 16 pmol 3'-*O*-N_3-dGTP, 16 pmol 3'-*O*-N_3-dCTP, 16 pmol 3'-*O*-N_3-dTTP, 1 unit Therminator^{TM} II DNA polymerase (also known as 9^oN mutant Polymerase enzyme, New England Biolabs, Ipswich, MA), 1X ThermoPolII buffer, and 2 mM of $MnCl_2$ was added to 2 μl of coupled PVA beads (10 mg/ml), then incubated at 65^oC for 30 min, mixed using a rotator. After incorporation, the beads were washed three times with SPSC buffer and once with water by heating at 65^oC for 5 min each time. Then the beads were re-suspended in 20 μl water, 0.2 μl of which was spotted on the slide. After the spot dried, it was scanned. To extend any remaining priming strand, a capping step was performed with 12.5 μl solution containing 3'-*O*-N_3-dNTPs (30 pmol 3'-*O*-N_3-dATP, 15 pmol 3'-*O*-N_3-dGTP, 40 pmol 3'-*O*-N_3-dCTP and 20 pmol 3'-*O*-N_3-dTTP), 1 unit of 9^oN mutant Polymerase enzyme, 1X ThermoPolII buffer, and 2 mM of $MnCl_2$, 65^oC for

25 min, twice. After the capping step, the beads were incubated in TCEP cleavage solution (100 mM, pH 9) for 15 min at 65^{o}C twice to cleave both the fluorophore and 3'-*O*-azido groups. Upon confirmation of cleavage by scanning, the cycles of incorporation, detection, capping, cleavage, and detection were repeated to sequence the subsequent bases.

The extension strategy was changed for performing 4-color SBS on streptavidin beads. Instead of using a mixture of 3'-*O*-N_3-dNTP-N_3-fluorophores and 3'-*O*-N_3-dNTPs, only 3'-*O*-N_3-dNTP-N_3-fluorophores were used for base identification. Both these dye-labeled nucleotides' concentration and the incorporation time were reduced due to the reduced competition. 10 µl of a solution consisting of 3 pmol 3'-*O*-N_3-dATP-N_3-ROX, 0.3 pmol 3'-*O*-N_3-dGTP-N_3-Cy5, 1 pmol 3'-*O*-N_3-dUTP-N_3-R6G, 0.5 pmol 3'-*O*- N_3-dCTP-N_3-bodipy-FL-510, 1 unit of 9^{o}N mutant Polymerase enzyme, 1X ThermoPolII buffer, and 2 mM $MnCl_2$ was added to 5 µl of coupled beads (10mg/ml), then incubated at 65^{o}C for 15 min, mixed using a rotator. After incorporation, the beads were washed twice with SPSC buffer and once with water by incubating in a 65^{o}C water bath for 5 min each time. Then the beads were re-suspended in 100 µl water, 0.2 µl of which was spotted on the slide. After the spot dried, the slide was checked under the microscope to guarantee the beads were dispersed well, and then scanned. To synchronize any unincorporated templates, an extension solution consisting of 37.5 pmol of each 3'-*O*-N_3-dNTP, 1 unit of 9^{o}N mutant Polymerase enzyme, 1X ThermoPolII buffer, and 2 mM of $MnCl_2$,was added to the beads and

incubated at 65°C for 25 min, twice. After capping, the beads were incubated in TCEP cleavage solution for 15min at 65°C twice to cleave both the fluorophores and 3'-*O*-azido groups. Upon confirmation of cleavage by scanning, the cycles of incorporation, detection, capping, cleavage, and detection were repeated to sequence the ensuing bases, as shown in Fig.7.8 A.

7.4.4 PCR on beads and emulsion-beads in droplet

The bead PCR reaction was carried out in 50 μl of a PCR mixture consisting of 20 ng pBluescript II SK+, 10 nmol dNTPs, 15 nmol M13 reverse primer (5'-CAGGAAACAGCTATGAC-3'), 3 μl forward primer coupled bead suspension, 1 U JumpStart™ REDTaq® DNA Polymerase and 1X corresponding reaction buffer. Amplification was performed under the following conditions: incubation at 94°C for 2 min, followed by 30 cycles of 94°C for 20 s, 55°C for 30 s, 72°C for 30 s and final extension at 72°C for 7 min, with occasionally stops at the annealing step to vortex the solution, preventing beads from settling down. After PCR, amplified beads were washed extensively to get rid of any non-specific binding of amplified template in solution, and the supernatant of the last wash was kept for Sanger sequencing analysis. The Sanger sequencing reaction mixture including 0.4 μl of BigDye Terminator 3.1 cycle sequencing mix (Applied Biosystems), 1X buffer 3.1, 4 pmol forward primer and the bead suspension (or the supernatant from the last wash) was carried out as follows: 25 cycles of 94°C for 20 s, 55°C for 45 s, 68°C for 90 s and final extension at 68°C for 3 min.

An emulsion was created by mixing 500 µl of oil phase with 240 µl H_2O at 25 rpm/sec for 5 min, followed by gradual addition of 160 µl mix with bead-containing aqueous phase at a speed of 18 rps for 5 min. The oil phase was composed of 490 µl of mineral oil, 450 µl of 10% Span-80, 40 µl of 10% Tween-80 and 20 µl of 10% TritonX-100. The aqueous phase consisted of 58 µl of 1X Taq polymerase buffer, 12 µl of 25 % glycerol and 5.0 µl of beads.

7.5 Conclusion

Aiming at massively parallel sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators, we have explored the potential of employing a sequencing-on-beads platform to increase the throughput, to overcome current short sequencing read-length, and to achieve maximum coverage across the genome. Various bio-conjugation chemistries with different types of microbeads have been tested to obtain better DNA attachment and less non-specific binding. After incorporation tests on beads using conventional pyrosequencing and fluorescently labeled terminators, SBS with CF-NRTs (3'-*O*-N_3-dNTP-N_3-fluorophores) and NRTs (3'-*O*-N_3-dNTPs) were applied to on-bead sequencing and resulted in several contiguous correct base calls. Although the approach still needs optimization, it appears very promising for realizing ultra-massively parallel sequencing and the goals of whole genome sequencing.

References

1. Hall N. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, **2007**, *209*, 1518-1525.

2. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, **2003**, *100*, 8817-8822.

3. Diehl F, Li M, He Y, Kinzler KW, Vogelstein B, Dressman D. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nature Methods*, **2006**, *3*, 551-559.

4. Leamon JH, Lee Wl, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD, Berka J, Weiner M, Rothberg JM, Lohman KL. A massively parallel PicoTiterPlate^{TM} based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **2003**, *24*, 3769-3777.

5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **2005**, *437*, 376-380.

6. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang M, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **2005**, *9*, 1728-1732.

7. Lin Y. Novel strategies to increase read length and accuracy for DNA sequencing by synthesis. Dissertation (Ph.D.), Columbia University, **2010**.

8. Bentzen EL, Tomlinson ID, Mason J, Gresch P, Warnement MR, Wright D, Sanders-Bush E, Blakely R, Rosenthal SJ. Surface modification to reduce nonspecific binding of quantum dots in living cell assays. *Bioconjugate Chemistry,* **2005***, 16*, 1488-1494.

9. Byun J, Kim J, Chung W, Lee Y. Surface-grafted polystyrene beads with comb-like poly(ethylene glycol) chains: preparation and biological application. *Macromolecular Bioscience,* **2004***, 4*, 512-519.

10. Ajikumar PK, Ng JK, Tang YC, Lee JY, Stephanopoulos G, Too H-P. Carboxyl-terminated dendrimer-coated bioactive interface for protein microarray: high-sensitivity detection of antigen in complex biological samples. *Langmuir*, **2007**, *23,* 5670-5677.

11. Nagasaki Y, Kobayashi H, Katsuyama Y, Jomura T, Sakura T. Enhanced immunoresponse of

antibody/mixed-PEG co-immobilized surface construction of high-performance immunomagnetic ELISA system. *Journal of Colloid and Interface Science*, **2007,** *309*, 524-530.

12. Hermanson GT, Bioconjugate Techniques, **1996**, Academic. Press.

Part III Detection of Nucleic Acids with Molecular Probes

The improving DNA sequencing and SNP genotyping technologies provide essential information on DNA, as well as its encoded RNA and proteins, driving the further exploration of biological functions of these molecules. In this chapter, aimed at tracking mRNA's (e.g., sensorin mRNA) and their involvement in long term memory storage, we describe our quantum-dot based binary probes for *in vitro* detection of a DNA target similar to a portion of the sensorin mRNA molecule, as a step in their eventual use to trace actual mRNAs in living cells.

Chapter 8 Quantum Dot based FRET Binary Probes for Detection of Nucleic Acids

8.1 Introduction

The study of the flow of biomolecules from their origins in the cell body to their targets within highly localized cell compartments can suggest their possible roles in physiological and pathological states. Specifically, selective detection and tracking of nucleic acids *in vivo* is of great significance for studying particular biological functions and mechanisms. For example, monitoring mRNA transport in living neurons would provide important information about the dynamics of gene expression and regional changes in post-transcriptional regulation. Over the past years, our group has been interested in studying the role of sensorin mRNA in synaptic plasticity for long term memory studies in a model organism, the sea slug *Aplysia californica*, due to its simple nervous system, the relatively large size of the neurons, and the easy identification of individual neuronal types.[1] Long-term memory requires the transcription of specific genes in neurons, and the localization of the protein encoded by these transcripts to regions in the communicative neuronal connections, the synapses.[2] Sensorin is a sensory neuron-specific peptide neurotransmitter, whose mRNA (sensorin mRNA) localizes to distal neuronal processes and concentrates at synapses in sensory neurons (SNs) paired with motor neurons (MNs). The translation of sensorin mRNA is crucial to the synapse stabilization between SNs and MNs, synatic plasticity and hence long term memory

storage.[3]

However, state-of-the-art molecular probes are required to study the specific mRNA both *in vitro* and *in vivo* with highly selectivity and sensitivity. A variety of molecular beacons (MBs) and binary probes (BPs) have been explored as tools to detect oligonucleotide sequences or RNA in different environments.[4] Nevertheless, non-specific opening of MBs often occurs in cells, causing false positive signals. Binary probes with multiple dyes displaying Förster resonance energy transfer (FRET) have advantages in terms of high specificity and the ease of detection at the injection site, which is very important to gain accurate signal reporting.

Typically, FRET-based BP systems are composed of two single-stranded DNAs (ssDNAs) which are complementary to adjacent regions of a common target sequence where a donor and acceptor are attached at opposite ends. In the absence of target, the fluorescence from the donor is mainly observed. When the pair of BPs selectively hybridizes to their target, the two oligonucleotide strands are drawn into close proximity and generate a FRET-based signal, the fluorescence of the acceptor. This requires that for ideal binary probes, the emission band of the donor should overlap as much as possible with the excitation band of the acceptor to maximize FRET efficiency, but the excitation band of the donor should be far from that of the acceptor to reduce direct excitation of the latter leading to a background signal. This poses challenges to current organic fluorophore systems, since no available dye pairs fulfill this criterion perfectly, and hence most of the previous BP systems were constructed at the sacrifice of energy transfer

efficiency. To significantly enhance its sensitivity and specificity, a binary probe system with the choice of a novel fluorescence donor is required.

Quantum dots (QDs), consisting of CdSe, CdS or CdTe, and ZnS, are a collection of semiconducting nanocrystals in the 2-8 nm size range (about twice the size of fluorescent proteins) that have large absorption cross sections, size-tunable emission bandwidths, high quantum yields, long emission lifetimes and excellent photostability,[5] and have attracted great interest for medical diagnostics, drug delivery, DNA-based nanosensors and solar energy conversion.[6-8] Specifically, CdSe/ZnS core shell QDs are advantageous over traditional organic fluorophores thanks to their large quantum yields and absorption coefficients, photostability, large Stoke's shifts, relatively long fluorescence lifetimes, and narrow and symmetric emissions from the ultraviolet to the infrared allowing multicolor multiplex assays. Their broad absorption and narrow emission spectra make them excellent FRET donors.[9] The QD fluorescence characteristics also allow for the selection of a wide range of excitation wavelengths that reduce the direct excitation of the acceptor; this also enables the use of a narrow bandpass filter for the effective separation of donor and acceptor fluorescence. Therefore, the use of the QD as the donor is a promising option for fundamental revision of the current binary probe system.

In our study, using a sensorin mRNA-like DNA sequence as the target, we describe the design and synthesis of a new QD-based BP in which the donor, QD 605, and the acceptor, Cy5, are attached to two different oligonucleotide strands with a range of 6, 9, and 16 base pair distances (around 2.0 nm, 3.0 nm, and 5.4 nm, respectively) between the

fluorophores when they are bound to their target. A simple and convenient method to synthesize a compact carboxylic QD-DNA conjugate through carbodiimide coupling is described, as an improvement to streptavidin QD-DNA conjugates in terms of FRET efficiency. The distance dependence of FRET efficiency was studied to optimize this BP system, with the evaluation of FRET using the S/B ratio, in which S/B = (Em_{667} with target/Em_{667} without target)/(Em_{605} with target/Em_{605} without target).

8.2 Experimental Rationale and Overview

It is believed that the introduction of quantum dots (QD) will bring a paradigm shift to the development of FRET-based binary probe systems due to the many advantages of QDs. In our binary probe system, the 605 nm carboxyl QD was chosen as the donor and Cy5 as the acceptor, since there is good spectral overlap between the emission of this QD at 605 nm and the absorption by Cy5, yet little overlap exists between the excitation wavelengths of the QD and Cy5, which avoids direct absorption by Cy5, as shown in Fig. 8.1. This exactly met the criteria described above for an ideal binary probe.

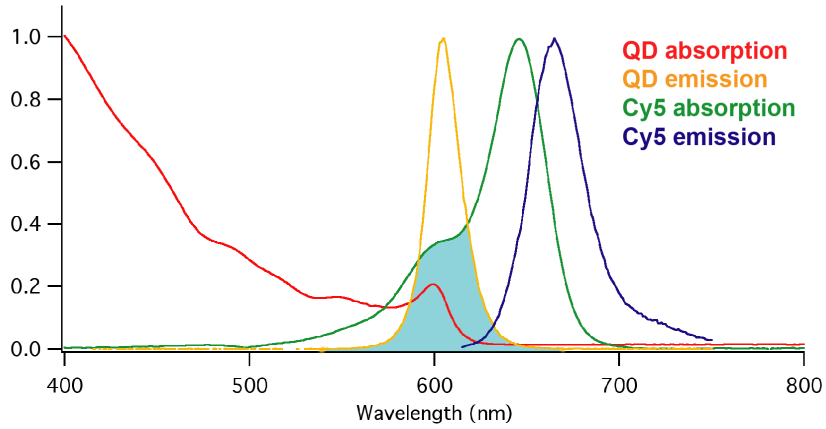


Fig. 8.1. The emission and excitation spectra of QD 605 and Cy5.

The existing methods for the construction of QD-DNA conjugates include streptavidin-biotin interaction,[10] glycosidic bonding,[11] electrostatic interaction,[12] metal-thiol bonding,[13] and carbodiimide reaction.[14, 15] Due to the molecules/spacers/ linkers that coat the QD surface, the size of the QDs is larger than the Förster distance, 4-7 nm, so the energy transfer efficiency between a core of QD and a single dye is sufficiently low.[16] The smaller moiety coated QD, carboxylic QD, was chosen to form a compact carboxylic QD-DNA conjugate, in which carboxylic groups on the QD surface are directly attached to the 3'-end of an oligonucleotide strand through carbodiimide coupling.[17] This enabled us to reduce the size of the QD probe and allow a separation distance between the QD and Cy5 of 4-7 nm.

Using sensorin mRNA similar DNA sequence as the target, as shown in Fig. 8.2, our binary probe system was designed in such a way that QD-based binary probes could hybridize to the target DNA with different separation distances between the QD and Cy5, by varying the base number between the two hybridization sites of the target DNA: 6, 9 and 16 bases, respectively. The most favorable separation distance for FRET between the QD donor and Cy5 acceptor was studied by measuring the FRET efficiency as a function of varying the base distances between two probes.

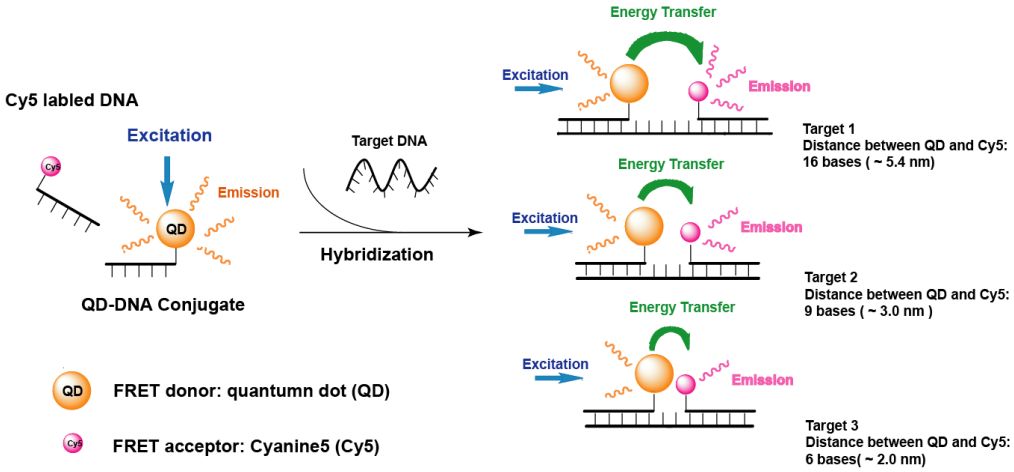


Fig. 8.2. Scheme of QD-based FRET binary oligonucleotide probes for DNA detection.

8.3 Results and Discussion

8.3.1 The synthesis of QD-DNA conjugates

3'-end amino modified oligonucleotides were attached to the carboxylic groups on the QD surface to form a carboxyl-QD-DNA conjugate through the carbodiimide coupling chemistry. The product of the carboxylic QD-DNA reaction was assessed using a gel mobility shift assay with the Agilent Bioanalyzer. As shown in Fig. 8.3, a new band appeared between the QD and amino-DNA band, which exhibited decreased mobility compared to free DNA, and increased mobility compared to free QD. It was assumed that the new band corresponds to the QD-DNA conjugates.

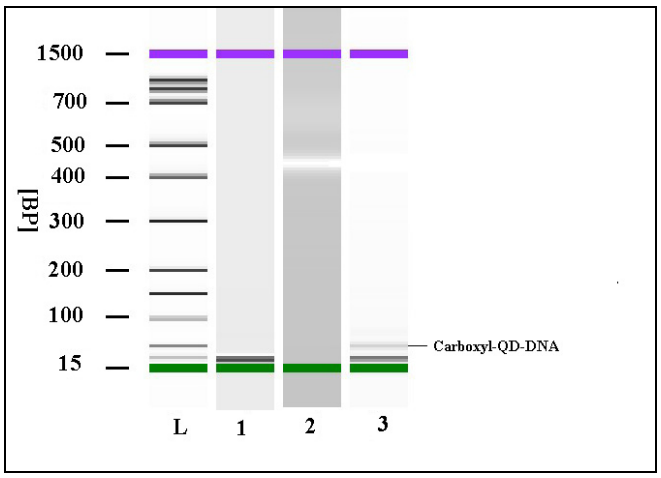


Fig. 8.3. The distribution of carboxyl-QD (1), DNA (2) and carboxyl-QD-DNA conjugates (3) after gel electrophoresis.

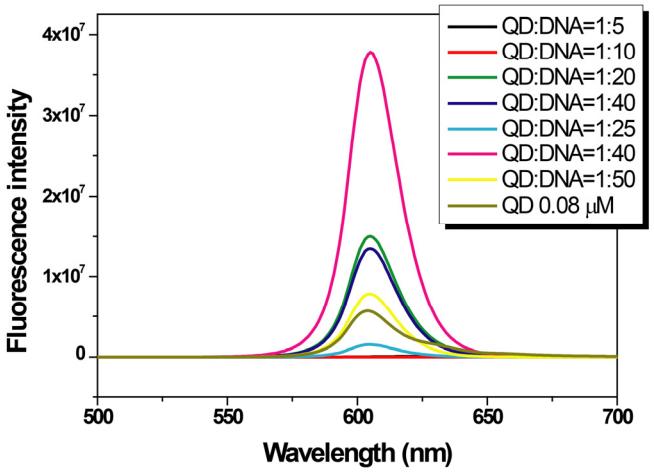


Fig. 8.4. Fluorescence intensity of conjugation products generated with different ratios of carboxylic QD to DNA.

It was noticed that the ratio of acceptor to donor is a very important factor for FRET efficiency. The ratio of acceptor to donor was determined by the ratio of amino-modified DNA to carboxylic QD. Hence, different ratios of amino-DNA to carboxylic QDs, 1:1, 5:1, 10:1, 20:1, 30:1, 40:1 and 50:1, respectively, were tested for the coupling reaction. It

was found that the fluorescence intensity was highest at the ratio of 40:1, as shown in Fig. 8.4. Therefore, the ratio of 40:1 was used for coupling synthesis in this work. To compare the carboxyl QD-DNA conjugates with streptavidin QD, the ratio of biotin DNA to streptavidin QD was also set to 40:1 for their interaction.

8.3.2 Hybridization kinetics studies

The feasibility of using this set of QD based binary probes was first assessed by measuring their hybridization kinetics. Taking target 3 (T3) as the example, steady-state fluorescence analysis of the hybridization kinetic reaction was carried out at an excitation wavelength of 460 nm in annealing buffer. Steady-state fluorescence spectra were recorded at different reaction intervals. As shown in Fig. 8.5, when the probes are randomly distributed in solution, emission primarily from the QD is observed. With the introduction of the target DNA, the fluorescence emission intensity of the QD-DNA conjugate gradually decreased over time, while the fluorescence intensity of Cy5 increased. The hybridization took around 5 hours to reach equilibrium. The long hybridization time may be because the high density of DNA conjugated on the QD surface hindered the hybridization process.[18, 19]

This study demonstrated that the binary probes (BPs) can be prepared using carboxylic QD as a donor and Cy5 as an acceptor for the detection of DNA. Unlike Wang's method,[20] in which spacers were used to coat the QD surface to form QD-DNA conjugate nanosensors for DNA detection, our modification for construction of QD-DNA

conjugates in which the QD and the Cy5 are on the inner portion of the probes rather than the outer portion reduces the distance between the dyes and the QD surface, and hence increases the FRET efficiency. This assumption was further verified by the relative effectiveness between the two binding methods used for constructing QD-based BPs, as discussed in the following section.

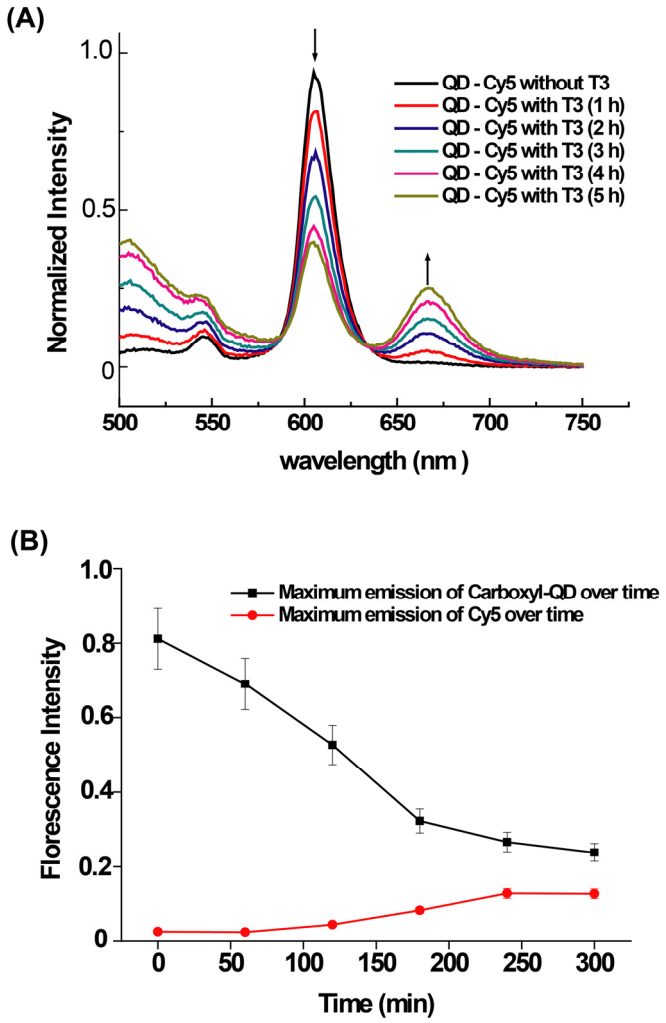


Fig. 8.5. The time-dependent spectral evolution of the FRET between carboxylic QD-DNA and Cy5-DNA with hybridization time, presented (A) as fluorescence spectra and (B) in line graph form.

8.3.3 Comparison of FRET efficiency between carboxyl-QD-DNA conjugate and streptavidin-QD-DNA conjugate based binary probes

In this study, two different methods were used to attach DNA to QDs: (1) covalent linkage of amino-labeled DNA to the carboxylic QD surface by carbodiimide coupling chemistry and (2) linkage of biotinylated DNA to streptavidin-modified QDs via streptavidin-biotin interaction. The two linkage methods have not been quantitatively compared in BPs, yet we observed that the linkage method has a measurable effect on the properties of BPs. As shown in Fig. 8.6, when the probes are

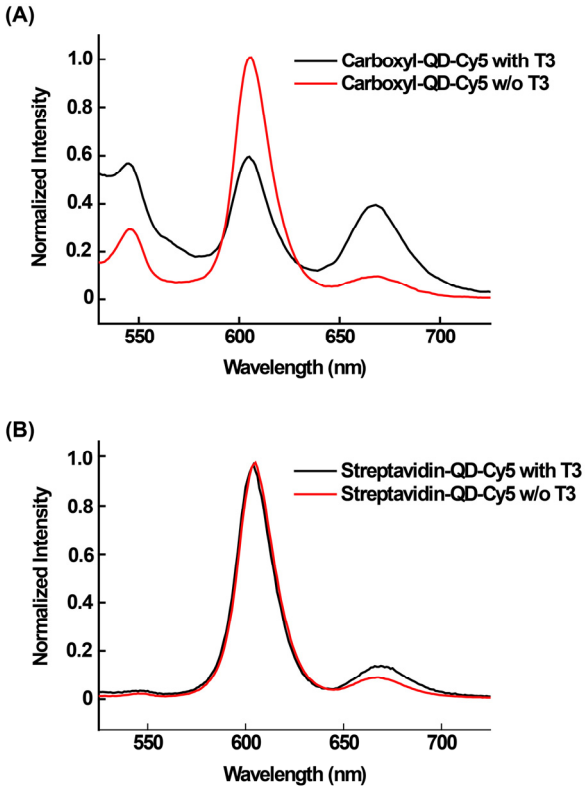


Fig. 8.6. Comparison of FRET efficiency between Carboxylic QD based BPs and Streptavidin QD based BPs: the fluorescence spectra before and after hybridization with target DNA 3. (A) Carboxylic QD-Cy5 system; (B) Streptavidin QD-Cy5 system.

randomly distributed in solution, emission primarily from the QD is observed. In the presence of target (T3), covalent QD-based BP exhibited a more significant increase in Cy5 emission at 667 nm and more significant decease in the QD emission at 605 nm than that of the streptavidin QD, at equivalent concentrations. The FRET efficiency between the QD donor and Cy5 acceptor was evaluated by the S/B ratio. The S/B ratio depends on the extent of the increase in the Cy5 signal and the decrease in the QD signal. Based on the steady-state fluorescence spectra shown in Fig. 8.6 A , the covalent QD-based BP shows a S/B ratio of 7.5 while the streptavidin QD has a S/B ratio of only 2.1 (Fig. 8.6 B). The different S/B ratios can be attributed to the different linkage strategies which are related to the different distances between the QD core and Cy5. For the streptavidin QD, large protein spacers on the surface of the QD greatly increase the distance between the QD core and Cy5, a parameter that is critical to FRET based quenching.

8.3.4 Distance dependent FRET studies with carboxyl-QD binary probes

In addition to studying the different linkage methods, we also investigated the effect of separation distance for FRET between the QD donor and Cy5 acceptor by varying the DNA base distance between the donor and acceptor (Fig. 8.2). The steady-state fluorescence spectra for all three DNA targets, T1, T2, and T3, all of which hybridize with the carboxylic QD-DNA and Cy5-DNA, are similar (Fig. 8.7). The S/B ratio increased as the number of bases between the QD donor and Cy5 acceptor increased, with T1 having a S/B ratio of 9.0 and the longest distance of 5.4 nm (16 bases). Since

there is a high density of DNA on the QD surface, the longest base distance might be favorable for QD hybridization with Cy5 due to less of a hindrance effect.

Time-resolved lifetime fluorescence spectra (Fig. 8.8) of the covalent QD-DNA and Cy5-DNA with different targets were consistent with the steady state fluorescence results with excitation of QD at 460 nm. The lifetime of free QD in buffer shows biexponential decays of 3.15 ± 0.3 ns and 10.28 ± 1.0 ns. Upon the addition of Cy5-DNA, the lifetime of the QD remained nearly the same with a lifetime of 4.40 ± 0.4 ns and 13.03 ±1.0 ns. In the presence of the target DNAs, T1, T2, and T3, the lifetime of the QD decreased as shown in Table 8.1, was and were fitted to a biexponential decay function. For the three sets (see Table 8.1), the lifetimes are 8.39 ± 0.8 ns, 10.17 ± 1.0 ns, and 12.16 ± 1.2 ns. The lifetime decreased as the separation distance between the QD and Cy5 decreased. It was assumed that with longer distance between QD and Cy5, faster hybridization would occur, and less delay would be observed due to decreased hindrance of hybridization.

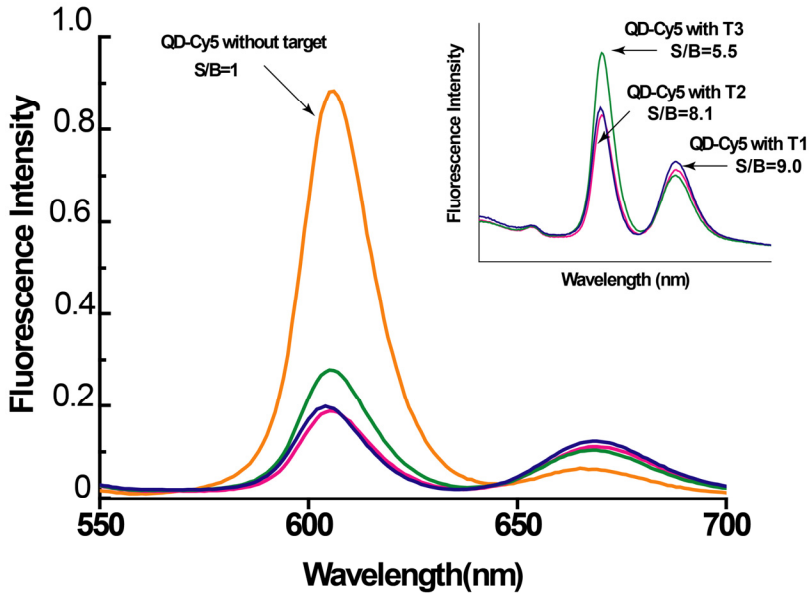


Fig. 8.7. Steady-state fluorescence spectra of carboxyl-QD and Cy5-DNA hybridization with different targets and without targets. Inset is an enlargement of just the target-containing hybridization reactions with the abscissa contracted for clarity.

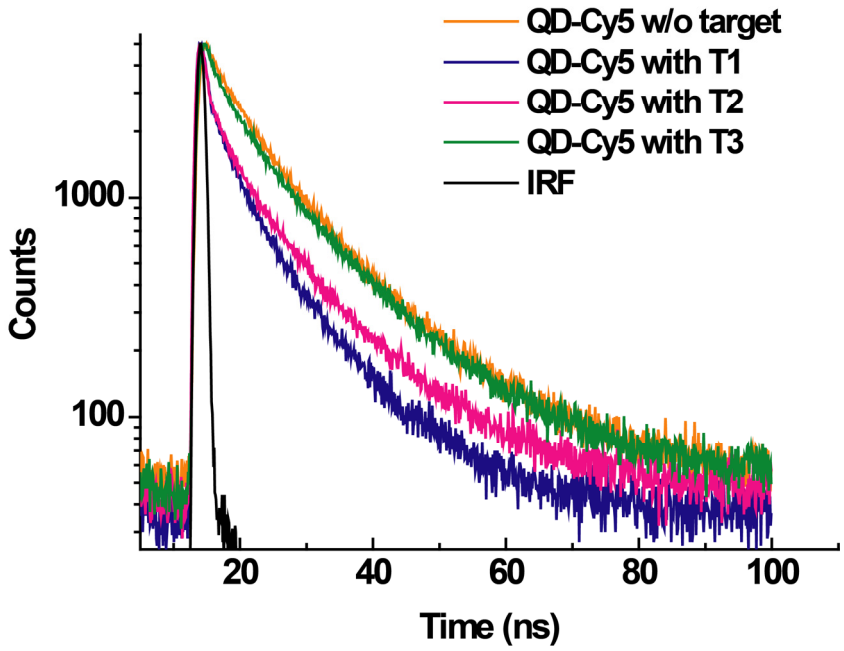


Fig. 8.8. Time-resolved lifetime fluorescence spectra of carboxyl-QD-DNA and Cy5-DNA with different targets

Table 8.1 Lifetime data for QD-DNA, Cy5-DNA hybridization with different targets

Entry	τ_{exc} 460 nm /em 605 nm, ns (abundance,%)	τ_{exc} 460 nm / em 667 nm, ns (abundance,%)	τ_{exc} 659 nm / em 667 nm, ns (abundance,%)
QD	10.28(61.11) 3.15(38.89)		
Cy5			1.39
QD+Cy5	13.03(72.24) 4.40(27.76)	1.36(86.92) 5.06(13.08)	1.45
QD+Cy5-T3	12.16(79.57) 2.83(20.43)	1.41(77.28) 3.38(22.72)	1.66
QD+Cy5-T2	10.17(76.25) 1.12(23.75)	1.41(80.32) 3.46(34.86)	1.67
QD+Cy5-T1	8.39(73.12) 1.06(26.88)	1.41(87.25) 4.40(12.75)	1.54

8.4 Materials and Methods

General information. Qdot® ITK™ Carboxyl Quantum Dots (605 nm) and Qdot® 605 ITK ™ Streptavidin were from Invitrogen by Life Technologies (San Diego, CA). 1-ethyl-3-(3-dimethyl-aminopropyl) carbodiimide hydrodrochloride (EDC) was obtained from Sigma-Aldrich. All the chemicals were purchased from Sigma-Aldrich unless specified. The oligonucleotides (Cy5-DNA, amino-DNA and target DNA), whose sequences were based on *Aplysia* sensorin mRNA, were synthesized on a DNA Synthesizer (Expedite 8909, Applied Biosystems by Life Technologies) using standard solid-phase phosphoramidite chemistry (Glen Research) or obtained from Integrated DNA Technologies (Coralville, IA). The sequences of the oligonucleotide probes and targets are shown in Table 8.2.

Table 8.2 Sequences of the probes and targets

DNA code	Sequence (5'-3')
QD-DNA	GUUGAUCACGGCUCAGGCGAAGGGA-QD
Cy5-DNA	Cy5-CAA AAG ACUUGGACCUGUCU
Target DNA 1(T2) (~5.4nm) *	AGACAGGUCCAAGUCUUUUG**AGUCUUCUGGACGGCU**UCCCUUCGCCUGAGCCGUGAUCAAC
Target DNA 2 (T4) (~3.0 nm) *	AGACAGGUCCAAGUCUUUUG**AGGACGGCU**UCCCUUCGCCUGAGCCGUGAUCAAC
Target DNA 3 (T6) (~2.0 nm) *	AGACAGGUCCAAGUCUUUUG**AGGGCU**UCCCUUCGCCUGAGCCGUGAUCAAC

Note: The underscored sequences indicate the positions to which the probes hybridize. * The approximate distance between QD and Cy5 after hybridization.

8.4.1 Synthesis of QD-DNA conjugates

The carboxylic groups on the QD surface were activated with EDC and allowed to react with 3' amino labeled DNA. 6.4 pmol of QD was mixed with 2.08 nmol of EDC and 0.256 nmol of 3' amino-labeled DNA in 85 µL of H_2O, with EDC added every 30 min to a total amount of 2.08 nmol. The mixture was allowed to react for 3 h at room temperature. Excess DNA and EDC were immediately removed from the carboxylic QD-DNA conjugates by spin filtration using Amicon Ultra 30,000 MWCO spin filters (Millipore, Billerica, MA). To maximize the recovery efficiency, the reaction mixture was first diluted with dH_2O to a volume of 1400 µl and evenly distributed into 4 spin filters for purification. The mixture was spun at 14,000 rpm for 1 min and the retentate was suspended in 400 µL water after the flow-through was discarded. The washing was repeated at least 3 times before the final products were collected by spinning down the filters. The purified QD-DNA conjugates were suspended in water and assessed with an

Agilent 2100 Bioanalyzer using the DNA 1000 Kit.

For the streptavidin QD-DNA conjugate, biotinylated DNA was attached to streptavidin QD by the streptavidin-biotin interaction. 6.4 pmol of streptavidin QD was mixed with 0.256 nmol of biotinylated DNA and 0.4 mg BSA in 400 μL of phosphate buffered saline (PBS, pH 7.4). The mixture was incubated at room temperature for 2h, purified and analyzed as described above for the carboxylic QD-DNA conjugate.

8.4.2 Hybridization of QD-DNA and Cy5-DNA with different targets

The hybridization of target DNA with the carboxylic QD-DNA conjugates and Cy5-DNA was carried out in 1X annealing buffer (10 mM Tris-HCl, 100 mM NaCl, 1mM EDTA, pH 7.4). For each hybridization event, 2.4 pmol of QD as the starting material for QD-DNA conjugate preparation was used and 0.28 nmol of Cy5-DNA was added. A fluorescence spectrum was obtained with a Fluorolog-3 spectrometer before adding target to generate the background emission spectrum. 40 pmol of target DNA was used for hybridization. The total volume of the hybridization reaction solution was kept constant at 300 μL. Therefore, 40 pmol of three different targets with different base distances were studied under identical conditions. The hybridization reaction was carried out at room temperature for ~5 h before the final fluorescence spectra were obtained.

For kinetic studies, the required carboxylic QD-DNA conjugate and Cy5-DNA solution in annealing buffer were prepared and transferred to a fluorescence cuvette, and a fluorescence spectrum was obtained. 0.04 nmol target DNA was then added and quickly

mixed with a micropipette. Steady-state fluorescence spectra were then recorded at different reaction intervals.

To quantitatively compare the energy transfer efficiency of the various QD-based BPs, the fluorescence spectra of the hybridization between the streptavidin QD-DNA conjugate, Cy5-DNA and target (T3) were also measured by steady-state fluorescence. The concentration of streptavidin QD-DNA conjugate was fixed at 2.4 pmol, Cy5-DNA at 0.28 nmol and targets at 0.04 nmol, respectively.

8.4.3 Steady-state fluorescence and time-resolved fluorescence measurement

Steady-state fluorescence was recorded at room temperature on a Fluorolog-3 spectrometer FL3-22 (J. Y. Horiba, Edison, NJ) using quartz cuvettes with a 4 mm path length. Time-resolved fluorescence measurements were performed on an OB920 single-photon counting spectrometer (Edinburgh Analytical Instruments) with a picoquant 460 nm pulsed LED or 659 nm diode laser as excitation source. Exponential fittings were obtained by a program included in the instrument (F900).

8.5 Conclusion

We have successfully constructed a carboxylic QD-based binary probe for detection of a DNA sequence using FRET. We have also demonstrated that QD carboxylic acid linkage to DNA reduced the overall particle size compared to the biotin-streptavidin linkage. We demonstrated the distance dependence in FRET, with the T1 distance of ~5.4 nm showing the most efficient FRET between a QD donor and Cy5 acceptor. Further

studies are in progress to evaluate the effectiveness of this QD-based probe inside a cell extract and in living cells.

References

1. Martí AA, Li X, Jockusch S, Li Z, Raveendra B, Kalachikov S, Russo JJ, Morozova I, Puthanveettil SV, Ju J, Turro NJ. Pyrene binary probes for unambiguous detection of mRNA using time-resolved fluorescence spectroscopy. *Nucleic Acids Research,* **2006**, *34(10)*, 3161-3168.

2. Korte M. Bridging the gap and staying local. *Science*, **2009**, *324*, 1527-1528.

3. Wang DO, Kimm SM, Zhao Y, Hwang H, Miura SK, Sossin WS, Martin KC. Synapse- and stimulus-specific local translation during long-term neuronal plasticity. *Science*, **2009**, *324*, 1536-1540.

4. Martí AA, Jockusch S, Stevens N, Ju J, Turro NJ. Fluorescent hybridization probes for sensitive and selective DNA and RNA detection. *Accounts of Chemical Research*, **2007**, *40*, 402-409.

5. Nathaniel CC, Strickland AD, Batt CA. Optimized linkage and quenching strategies for quantum dot molecular beacon. *Molecular and Cellular Probes*, **2007**, *21*, 116-124.

6. Goldman E, Medintz I, Whitley J, Hayhurst A, Clapp A, Uyeda H, Deschamps J, Lassman M and Mattoussi H. A hybrid quantum dot-antibody fragment fluorescence resonance energy transfer-based TNT sensor. *Journal of the American Chemical Society,* **2005***, **127***, 6744-6751.

7. Zhang CY, Yen HC, Kuroki MT, Wang TH. Single-quantum-dot-based DNA nanosensor, *Nature Materials,* **2005**, *4*, 826-831.

8. Shi L, Paoli VD, Rosenzweig N, Rosenzweig Z. Synthesis and application of quantum dots FRET-based protease sensors. *Journal of the American Chemical Society,* **2006**, *128*, 10378-10379.

9. Ute RG, Markus G, Sara CJ, Roland N, Thomas N. Quantum dots versus organic dyes as fluorescent labels, *Nature Methods*, **2008**, *5*, 763-775.

10. Xiao Y, Barke PE. Semiconductor nanocrystal probes for human metaphase chromosomes, *Nucleic Acids Research,* **2004**, *32*, e28.

11. Pathak S, Choi SK, Arnheim N, Thompson ME. Hydroxylated quantum dots as luminescent

probes for in situ hybridization. *Journal of the American Chemical Society,* **2001,** *123***,** 4103-4104.

12. Artemyev M, Kisiel D, Abmiotko S, Antipina MN, Khomutov GB, Kislov VV, Rakhnyanskaya AA. Self-organized, highly luminescent CdSe nanorod-DNA Complexes. *Journal of the American Chemical Society,* **2004**, *126*, 10594-10597.

13. Medintz IL, Berti L, Pons T, Grimes AF, English DS, Alessandrini A, Facci P, Mattoussi H. A reactive peptidic linker for self-assembling hybrid quantum dot-DNA bioconjugates, *Nano Letters*, **2007**, *7*, 1741-1748.

14. Wang X, Lou X, Wang Y, Guo Q, Fang Z, Zhong X, Mao H, Jin Q, Wu L, Zhao H, Zhao J. QDs-DNA nanosensor for the detection of hepatitis B virus DNA and the single-base mutants, *Biosensors &Bioelectronics*, **2010**, *25* ,1934-1940.

15. Zhou D, Ying L, Hong X, Hall EA, Abell C, Klenerman D. A compact functional quantum dot-DNA conjugate: preparation, hybridization, and specific label-free DNA Detection. *Langmuir,* **2008**, *24*, 1659-1664.

16. Roller RS, Winnik MA. The determination of the Förster distance for phenanthrene and anthracene derivatives in poly(methyl methacrylate) films. *The Journal of Physical Chemistry B,* **2005,** *109,* 12261-12269.

17. Shen H, Jawaid AM, Snee PT. Poly(ethylene glycol) carbodiimide coupling reagents for the biological and chemical functionalization of water-soluble nanoparticles. *ACS Nano*, **2009**, *3*, 915-923.

18. Peterson AW, Heaton RJ, Georgiadis RM, The effect of surface probe density on DNA hybridization, *Nucleic Acids Research,* **2001**, *29*, 5163-5168.

19. Dubertret B, Skourides P, Norris DJ, Noireaux V, Brivanlou AH, Libchaber A. *In vivo* imaging of quantum dots encapsulated in phospholipid micelles, *Science*, **2002**, *298*, 1759 -1762.

20. Wang X, Lou X, Wang Y, Guo Q, Fang Z, Zhong X, Mao H, Jin Q, Wu L, Zhao H and Zhao J. QDs-DNA nanosensor for the detection of hepatitis B virus DNA and the single-base mutants. *Biosensors and Bioelectronics*, **2010**, *25*, 1934-1940.

Chapter 9 Summary and Future Outlook

The ultimate goal of my thesis research was to contribute to the development and improvement of genomic technologies encompassing DNA sequencing and single nucleotide polymorphism (SNP) genotyping via different molecular approaches, and to the development of novel molecular probes for further understanding of the genetic flow and biological functions of nucleic acid molecules as a whole. The following specific goals, which were set to test and implement key components of DNA sequencing, SNP genotyping as well as the molecular probe development, were achieved: (1) Design, synthesis and analysis of a complete set of novel cleavable biotinylated dideoxynucleotides for mass spectrometry based DNA sequencing and SNP genotyping with high accuracy and sensitivity, and the initial exploration of a microfluidic lab-on-chip device with the potential for high throughput, miniaturization and automation; (2) Design and integration of a novel primer walking strategy for extending the read-length of sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators (CF-NRTs), and the exploration of a beads-on-chip approach for increasing the throughput as well as the coverage for SBS; (3) Design and construction of quantum dot based binary probes for detection of nucleic acids.

9.1 Mass spectrometric DNA sequencing and SNP genotyping with cleavable biotinylated dideoxynucleotides[1, 2]

Aiming to improve the current solid phase capturable mass spectrometric DNA

sequencing and SNP genotyping method, a complete set (A, C, G, and U) of mass tagged, cleavable biotinylated dideoxynucleotides (ddNTP-N_3-biotins) was synthesized, and successfully used for both DNA sequencing and SNP genotyping with significantly improved performance. A sequencing read-length of over 30 bases was achieved for both synthetic DNA template (37 bases) and biological samples (32 bases). In SNP genotyping, we have been able to detect as low as 2.5% heteroplasmy in mitochondrial DNA samples, with interrogation of human mitochondrial genome position 8344 which is associated with mitochondrial disease (myoclonic epilepsy with ragged red fibers, MERRF); we have been able to quantify the heteroplasmy level by generating a calibration curve; we have also determined several mitochondrial MERRF mutations in a multiplex approach. We further designed and constructed a SNP genotyping microfluidic lab-on-a-chip device, which is also applicable for mass spectrometric sequencing, and confirmed the feasibility of this microdevice by performing individual functional steps on-chip. We have demonstrated 100% on-chip single base incorporation, sufficient capture and release of the biotin terminated single base extension products, and high sample recovery from the C18 reversed-phase microchannel with as little as 0.5 pmol DNA molecules.

9.2 Strategies to improve sequencing by synthesis with cleavable fluorescent nucleotide reversible terminators[3]

To overcome the current short sequencing read-length, we have developed a novel primer walking strategy to increase read-length for DNA sequencing by synthesis with

CF-NRTs. The primer walking method was conducted using three natural nucleotides (dNTPs) and one nucleotide reversible terminator (NRT), so that the primer would extend with natural nucleotides until the NRT is incorporated, and would resume after the cleavage reaction to regenerate the 3'-OH. Repeated cycles of such extension, pause, and cleavage enabled initiation of the next round of SBS sequencing anew at a base downstream to the first round sequenced region, allowing one to achieve a combined read-length from two rounds of SBS. We have successfully demonstrated the integration of this primer walking strategy into the sequencing by synthesis platform, and were able to obtain a total read-length of 53 bases after performing one round of sequencing, four walking cycles and then a second round of SBS.

The alternative approach we explored is to develop the sequencing bead-on-chip method to increase the throughput and hence the overall coverage of SBS with CF-NRTs. We have investigated various requisite conditions, including DNA attachment to microbeads and the separation of single beads in emulsion droplets, and demonstrated the feasibility of performing SBS on microbeads by accurately identifying several bases.

9.3 Detection of nucleic acids with molecular probes[4]

Taking advantage of key characteristics of quantum dots, we designed and tested carboxylic QD-based FRET binary probes, with QD as the donor and Cy5 as the acceptor. We have demonstrated the feasibility of this BP system in the detection of a sensorin mRNA related DNA sequence, and by conducting parallel experiments with slightly

different DNA targets, we were able to discover the best probe distance for efficient FRET, which will be further evaluated in a cell extract and in living cells.

9.4 Future outlook

The molecular tools for genomic analysis are undergoing dramatic evolution with the increasing demand for genetic analysis (e.g., whole genome sequencing, SNP genotyping) and requirements of further understanding the genetic transmission leading to biological functions. It is believed that the advancements in mass spectrometric DNA sequencing and SNP genotyping, sequencing by synthesis with CF-NRTs and oligonucleotide binary probe systems will continue, and these technologies will continue to play important roles in specific applications geared toward personalized medicine.

9.4.1 Mass spectrometric DNA sequencing and SNP genotyping

MALDI-TOF mass spectrometry will continue to offer an attractive option for DNA analysis due to its high accuracy, sensitivity and speed. With our introduction of mass-tagged, cleavable biotinylated dideoxynucleotides, these mass spectrometry based reversible solid-phase-capture sequencing and single base extension SNP genotyping technologies would be an ideal choice for sequencing short fragments (e.g., miRNA sequencing, transcriptome sequencing), mid-level sample handling and multiplexing, for any projects demanding extremely high accuracy and sensitivity at low cost and in a time-efficient manner, where either conventional sequencing technologies or next-generation sequencing fail to meet these requirements. They will also have

applications in genetic linkage and association studies for SNPs in pooled samples and regions containing mini-indels. The development of MALDI-TOF mass spectrometry technology in terms of higher available mass range, higher sensitivity and higher throughput will contribute to these technologies, allowing longer sequencing read-length and higher multiplexing level. In particular, the completion of our highly integrated, miniaturized, automatic, mass parallel microfluidic platform will push these mass spectrometric genotyping technologies to the next level.

9.4.2 DNA sequencing by synthesis

As the core technology in second generation sequencing, sequencing by synthesis with CF-NRTs will continue to be the method of choice for genome-wide projects. It is believed that the sequencing read-length can be further improved by optimization of nucleotide chemistry, advances in detection sensitivity and increasing the number of sequenceable targets, as shown in Fig. 9.1. Theoretically, even with the current detection system and the number of starting molecules, the SBS approach has the potential to yield comparable read-length with Sanger sequencing (~800 bp) if only the cycle efficiency is 99%. And this will be doubled, tripled or more by our primer walking strategy and further improved by the introduction of the bead-on-chip system allowing greater amounts of starting DNA molecules. The simple calculation of 300 million DNA amplicon carrying microbeads which could be easily immobilized on the microscope slide with read-lengths of 150-300 bp with paired-end reads will give more than 24-fold

coverage of the human genome in a single instrument run.

As the development of third generation sequencing, especially single molecule sequencing methods, is becoming the trend, the SBS chemistry and our primer walking strategy can be still integrated into single molecule detection platforms, with the promise of *de novo* genome sequencing and even wider applications, including RNA expression, epigenetic analysis, haplotype analysis, and other important areas of genome biology and biomedical sciences.

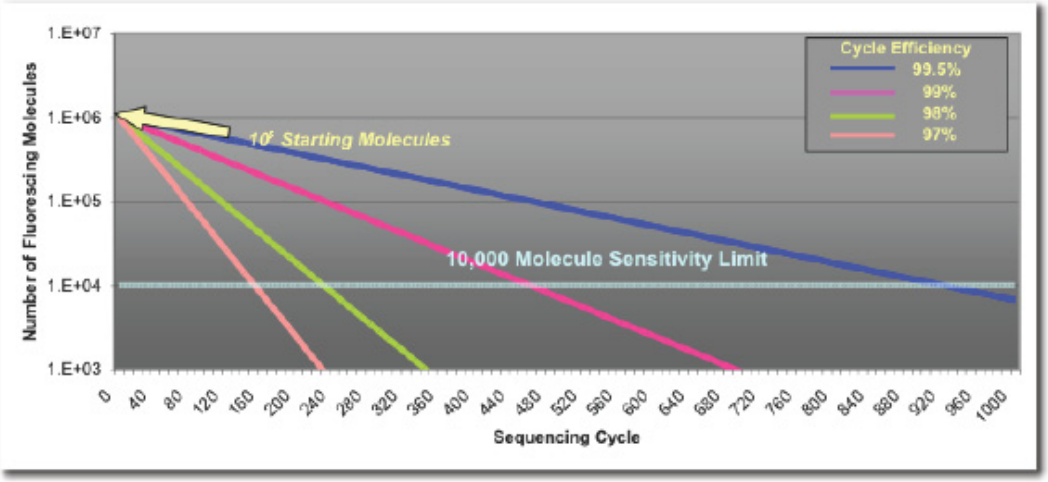


Fig. 9.1. Theoretical SBS read-length based on sequencing cycle efficiency.

9.4.3 *In vivo* visualization of nucleic acids

Real-time visualization of the nucleic acids *in vivo* is essential to decipher the fundamental biological functions of genes and transcripts and hence the whole mechanisms. The development of next generation molecule probes with long-lived fluorescence or luminescence, higher sensitivity and specificity for tracking low abundant transcripts and the capability to overcome the cell autofluorescence will be an

on-going project. Specifically, for the particular interest in long term memory study using *Alysia* as the model, continuing with the successfully example of using quantum dot as the donor in FRET based probe, there are more sets of quantum dot based molecular probes that will be developed and tested for *in vivo* studies, targeting different types of mRNA.

References

1. Qiu C, Kumar S, Guo J, Yu L, Guo W, Shi S, Russo JJ, Ju J. DNA sequencing by MALDI-TOF mass spectrometry using cleavable biotinylated dideoxynucleotides. 2011, to be submitted.

2. Qiu C, Kumar S, Guo J, Guo W, Lu J, Shi S, Kalachikov S, Russo JJ, Naini A, Schon E, Ju J. SNP genotyping by MALDI-TOF mass spectrometry using cleavable biotinylated dideoxynucleotides: application in mitochondrial disease. 2011, to be submitted.

3. Yu L, Qiu C, Guo J, Kalachikov S, Li Z, Xu N, Li X, Shi S, Russo JJ, Turro NJ, Ju J. Novel primer walking strategy for read length increment in DNA sequencing by synthesis. 2011, in preparation.

4. Peng Y, Qiu C, Jochusch S, Scott A, Li Z, Turro NJ, Ju J. Carboxyl quantum dots based FRET binary oligonucleotide probes for detection of nucleic acids. 2011, submitted to Photochemical & Photobiological Sciences.