# Progress Towards Ultra-rapid DNA Sequencing with Protein Nanopores

A dissertation by

David Stoddart

Submitted to the University of Oxford for the degree of

Doctor of Philosophy

Major Subject: Chemical Biology

Wadham

2011

# TABLE OF CONTENTS

DECLARATION ..... 7

ACKNOWLEDGEMENTS ..... 8

ABBREVIATIONS ..... 9

ABSTRACT ..... 13

CHAPTER 1: INTRODUCTION ..... 15

1.1. Introduction ..... 15

1.2. Sanger sequencing ..... 16

1.3. Next generation DNA sequencing ..... 20

1.3.1. Sequencing by ligation ..... 21

1.3.2. Sequencing by synthesis ..... 22

1.3.2.1. Cyclic reversible termination ..... 24

1.3.2.2. Pyrosequencing ..... 25

1.4. Single-molecule next generation sequencing ..... 28

1.4.1. True single-molecule sequencing (tSMS™). ..... 30

1.4.2. Single-molecule real-time sequencing (SMRT™) ..... 30

1.5. Nanopore sequencing ..... 33

1.6. Challenges of nanopore sequencing ..... 39

References ..... 42

CHAPTER 2: MATERIALS AND METHODS ..... 46
2.1. Molecular biology – mutant generation, protein expression and purification ..... 46
2.1.1. Mutant generation ..... 46
2.1.1.1. Plasmids ..... 46
2.1.1.2. Mutagenesis by PCR ..... 46
2.1.1.2.1. Quik change ..... 46
2.1.1.2.2. Homologous recombination ..... 47
2.1.1.2.3. Sequencing ..... 52
2.1.1.2.4. Plasmid amplification ..... 52
2.1.2. Protein expression and purification - coupled *in vitro* transcription and translation (IVTT) ..... 52
2.1.3. Hemolysis assays ..... 53
2.1.4. Proteolysis of α-hemolysin proteins with proteinase K ..... 55
2.1.5. PAGE in the absence of SDS ..... 55
2.2. Electrical recordings ..... 56
2.2.1. Electrical recording equipment ..... 57
2.2.2. Faraday cage ..... 57
2.2.3. Electrodes ..... 57
2.2.4. Chambers ..... 59
2.2.5. Preparing bilayers ..... 60
2.2.6. Inserting pores and perfusion ..... 63
2.3. Data acquisition and analysis ..... 64
2.3.1. Data digitization ..... 64

2.3.2. Filtering and sampling ..... 65
2.3.3. Acquisition protocols ..... 66
2.3.4. Analysis of DNA-btn•streptavidin molecules ..... 67
References ..... 73

CHAPTER 3: NUCLEOBASE DISCRIMINATION IN IMMOBILIZED OLIGONUCLEOTIDES WITH THE α-HEMOLYSIN PROTEIN NANOPORE ..... 76
Abstract ..... 76
3.1. Introduction ..... 77
3.2. Results and discussion ..... 81
3.2.1. Discrimination of homopolymeric oligonucleotides with the α-hemolysin (αHL) pore ..... 81
3.2.2. Improved discrimination with an αHL mutant protein ..... 85
3.2.3. Defining the recognition element within the αHL pore. ..... 87
3.2.4. Probing the recognition sites of αHL for four-base discrimination ..... 99
3.2.5. Four-base discrimination in a heteropolymeric sequence. ..... 103
3.3. Conclusions ..... 106
References ..... 108

CHAPTER 4: NUCLEOBASE RECOGNITION IN ssDNA AT THE CENTRAL CONSTRICTION OF THE α-HEMOLYSIN PROTEIN NANOPORE ..... 111
Abstract ..... 111
4.1. Introduction ..... 112
4.2. Results and discussion ..... 114

4.2.1. Effects of amino acid substitutions at position 113 on nucleobase recognition at recognition site $R_1$ ..... 114
4.2.2. Effects of amino acid substitutions on base discrimination at distal recognition sites ..... 125
4.2.3. Utilising multiple base-recognition sites in the αHL nanopore. ..... 130
4.3. Conclusions ..... 136
References ..... 139

CHAPTER 5: TRUNCATION OF THE RECOGNITION ELEMENT IN α-HEMOLYSIN PROTEIN NANOPORES BY DELETING PORTIONS OF THE TRANSMEMBRANE β BARREL ..... 142
Abstract ..... 142
5.1. Introduction ..... 142
5.2. Results and discussion ..... 144
5.2.1. Site-directed mutagenesis at the *trans* entrance of the αHL pore . 144
5.2.2. Amino acid deletions to truncate the β barrel at the *trans* entrance of the αHL pore ..... 150
5.2.3. Defining the recognition element within the truncated barrel mutant (TBM) pores ..... 166
5.2.4. Probing recognition site $R_3$ of αHL TBM pores for four-base discrimination ..... 171
5.3. Conclusions ..... 172
References ..... 175

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS ..... 177

6.1. Conclusions ..... 177

6.2. Future directions ..... 178

References ..... 181

APPENDIX ..... 183

A.1 Gene sequences ..... 183

A.2. Primer sequences ..... 185

A.3. Protein sequences ..... 186

A.4. Oligonucleotide sequences ..... 188

# DECLARATION

The work described in this thesis was carried out between April 2008 and January 2011 in the laboratory of Professor Hagan Bayley, at the Chemistry Research Laboratory at the University of Oxford. All the work described is entirely my own, unless where otherwise indicated. This work has not been submitted previously for any other degree at the University of Oxford, or any other university.

# ACKNOWLEDGEMENTS

First and foremost I would like to thank Professor Hagan Bayley for all the support, advice and encouragement that he has given me during my time in his lab. I am also indebted to Drs. Giovanni Maglia and Andrew Heron for their assistance in experimental design and data analysis techniques. I also thank Ellina Mikhailova for her efforts regarding mutant generation and protein preparation and Dr. Jochen Klingelhoefer for providing informative molecular dynamics simulation data. All the other members of the Bayley group, both past and present, have contributed to making my time in the lab both stimulating and enjoyable – I shall fondly remember my heated discussions with Dvir. My thanks also go to Dr. Stu Reid, who put me in this lab in the first place and has done much for me over the years; he really is a true gentleman and has always had the best interests of others at heart.

I would also like to thank my examiners, Prof. Jim Naismith and Dr. Mark Howarth, for sacrificing their time to read this thesis.

Finally, I must congratulate and apologise to Michelle, who put up with me, unceasingly, throughout the whole ordeal...may her patience and sense of humour never wear thin.

# ABBREVIATIONS

| | |
|---|---|
| A | adenine |
| AMP | adenosine monophosphate |
| αHL | α-hemolysin |
| AFM | atomic force microscopy |
| Biotin-TEG | biotin-triethyleneglycol |
| BSA | bovine serum albumin |
| C | cytosine |
| CAE | capillary array electrophoresis |
| CCD | charge-coupled device |
| DNA | deoxyribonucleic acid |
| dNTP | deoxynucleoside triphosphate |
| ddNTP | dideoxynucleoside triphosphate |
| (d)ATP | (deoxy)adenosine triphosphate |
| (d)CTP | (deoxy)cytosine triphosphate |
| (d)GTP | (deoxy)guanosine triphosphate |
| (d)TTP | (deoxy)thymine triphosphate |
| dATPαS | deoxyadenosine-5′-[α-thio]triphosphate |
| DDM | n-dodecyl-β-D-maltoside |
| dsDNA | double-stranded DNA |
| DLPC | 1,2-dilauryl-*sn*-glycero-3-phosphocholine |
| DPhPC | 1,2-diphytanoyl-*sn*-glycero-3-phosphocholine |
| EDTA | ethylenediaminetetraacetic acid |

| | |
|---|---|
| G | guanine |
| HCl | hydrochloric acid |
| h | hour(s) |
| $I_O$ | open pore current value (pA) |
| $I_B$ | blocked pore current value (pA) |
| $I_{RES}$ | residual current level (%) = ($I_B/I_O$) X 100 |
| $\Delta I_{RES}$ | difference in residual current levels |
| $\Delta I_{RES}^{OVERALL}$ | the difference between the most widely dispersed residual current levels in a histogram of residual current blockades |
| δ | the product of the sequential differences between each of the residual current levels |
| IVTT | *in vitro* transcription and translation |
| KCl | potassium chloride |
| KOH | potassium hydroxide |
| LASER | light amplification by stimulated emission of radiation |
| MBSA | 150 mM NaCl, pH 7.4, containing 1 mg $mL^{-1}$ BSA |
| MOPS | 3-[*N*-morpholino]propane sulphonic acid |
| MspA | *Mycobacterium smegmatis* porin A |
| min | minute(s) |
| mF | mutagenic forward primer |
| mR | mutagenic reverse primer |
| nmF | non-mutagenic forward primer |
| nmR | non-mutagenic reverse primer |
| nt | nucleotide |

| | |
|---|---|
| NaCl | sodium chloride |
| NIH | National Institutes of Health |
| NN | E111N/K147N |
| NNX | E111N/K147N/M113X |
| PAGE | polyacrylamide gel electrophoresis |
| PCR | polymerase chain reaction |
| PMSF | phenylmethylsulfonyl fluoride |
| PPi | pyrophosphate |
| PTFE | polytetrafluoroethylene |
| RNA | ribonucleic acid |
| rRBC | rabbit red blood cells |
| rRBCm | rabbit red blood cell membrane |
| SBS | sequencing by synthesis |
| SCR | single-channel recording |
| SDS | sodium dodecyl sulfate |
| SMRT™ | single-molecule real-time sequencing |
| SMS | single-molecule sequencing |
| SOLiD™ | support oligonucleotide ligation detection |
| ssDNA | single-stranded DNA |
| ssNA | single-stranded nucleic acid |
| T | thymine |
| TAE | 40 mM Tris-acetate, 1 mM EDTA, pH8.3 |
| TBM | truncated barrel mutant |
| TE 8 | 10 mM Tris.HCl, 0.1 mM EDTA, pH8 |

| | |
|---|---|
| TEG | triethyleneglycol |
| TIRF | total internal reflection fluorescence |
| Tris | 2-amino-2-hydroxymethyl-propane-1,3-diol |
| tSMS™ | true SMS |
| WT | wild-type |
| ZMW | zero-mode waveguide |

Progress Towards Ultra-rapid DNA Sequencing with Protein Nanopores

David Stoddart

Wadham College

Doctor of Philosophy

Hilary Term, 2011

# ABSTRACT

The sequencing of individual DNA strands with nanopores is being developed as a rapid, low-cost platform in which bases are identified in order as a DNA strand is transported through a pore under an electrical potential. Several challenges remain and this thesis focuses on one major area, the base identification properties of the α-hemolysin (αHL) nanopore. Under the potentials required for threading, DNA translocates too fast for single bases to be identified. However, immobilization of the DNA within the pore increases the residence time and therefore improves the precision of the electrical current reading and allows for the small differences in current flow, associated with different sequences, to be observed. DNA molecules with a 3′-terminal biotin-tag were complexed with streptavidin. Streptavidin is too large to be transported through the αHL pore and therefore the DNA-btn•streptavidin complex is not fully translocated; thus, the DNA strand is immobilized within the pore. Using this approach the nucleobase recognition properties of the αHL pore were mapped. The data suggest that the

transmembrane β barrel domain of the pore contains at least three nucleobase recognition sites, termed $R_1$, $R_2$ and $R_3$.

Additional sequence information can be gained when multiple recognition sites are employed within a single αHL pore, as compared to the simple case of a single recognition site. Recognition site $R_1$, which is located near the central constriction, can be modified by site-directed mutagenesis of Met-113. It was observed that amino acids with related side chains produce similar patterns of nucleobase recognition. Amino acids that provide an energy barrier to ion flow (e.g. bulky or hydrophobic residues) strengthen base identification, while amino acids that lower the barrier, weaken identification. Deletion and site-directed mutagenesis were used to remove one recognition site and generate an αHL pore with a truncated β barrel domain that contains only two recognition sites.

# CHAPTER 1: INTRODUCTION

## 1.1. Introduction

DNA is the information storage molecule for all complex life and it contains data that is fundamental in a wide variety of biological processes at the cellular and whole-organism level, ranging through development, reproduction and death. The sequencing of DNA molecules therefore offers insight into the functions, malfunctions and evolutionary relationships of all organisms. Until recently, sequencing methods were based on the technique developed by Fred Sanger in the 1970's. Indeed, the first drafts of the human genome, published in 2001 (Lander, Linton et al. 2001; Venter, Adams et al. 2001), were completed by Sanger sequencing, after a decade of concerted effort and at an estimated cost of $3 billion. Given this cost, the idea of personal genome sequencing seemed like science fiction. However, in 2004, the National Institutes of Health (NIH) set a remarkable challenge: the $1000 genome (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html). New funding opportunities and the $10 million "Archon X Prize for genomics" offered by the X Prize Foundation (which will go to the first private effort to sequence 100 genomes in 10 days, for less than $10000/genome – http://genomics.xprize.org/genomics/archon-x-prize-for-genomics), have sparked rapid advances in sequencing technology. The cost of sequencing is falling faster than ever and is paving the way for a genome revolution (Venter 2010).

## 1.2. Sanger sequencing

In December 1977, Fred Sanger published his “chain-termination method” for sequencing DNA in the Proceedings of the National Academy of Sciences (Sanger, Nicklen et al. 1977). In this work, he outlines how the bacteriophage φX174 genome was sequenced using 2′, 3′ dideoxynucleoside triphosphate (ddNTP) chain terminators, in addition to the traditional 2′ deoxynucleoside triphosphates (dNTPs). Enzymatic DNA replication is based on the sequential addition of complementary dNTPs to the 3′ hydroxyl group at the end of a growing polynucleotide chain (which is non-covalently bound to the template through Watson-Crick base pairing). However, after the addition of a ddNTP there is no 3′ hydroxyl group at the end of the oligonucleotide and chain elongation cannot continue. So, if a primer and template are incubated with a DNA polymerase, a mixture of 4 dNTPs (labelled with $^{32}P$) and a single ddNTP type (e.g. ddTTP), then a mixture of radio-labelled fragments will be produced, all of which are terminated at the 3′ end with a T nucleobase. When the mixture is subject to PAGE, the different length fragments are separated and the pattern of bands on the gel (as detected by autoradiography) shows the distribution of ddTTPs in the newly synthesized DNA fragments. If three other similar reactions are separately performed for each different ddNTP (i.e. ddATP, ddGTP and ddCTP), and run in parallel on the same gel, a pattern of bands is obtained from which the sequence can be read (Figure 1.1).

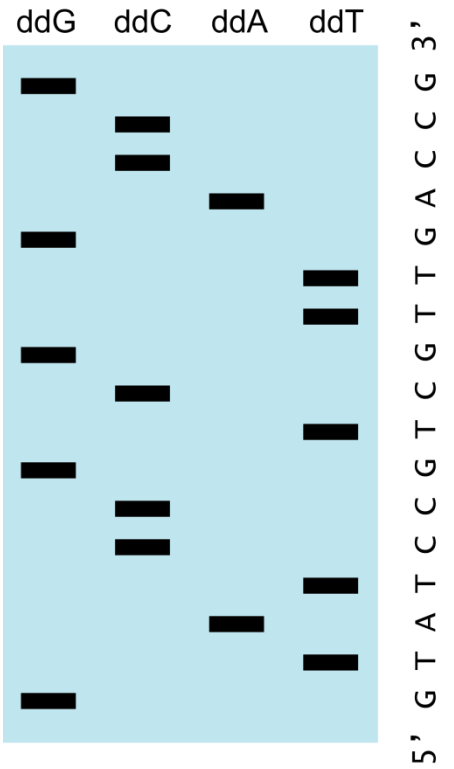



Figure 1.1. Sanger sequencing. In four separate reactions, the template DNA is incubated with primers, radio-labelled dNTPs and one type of ddNTP, which will terminate chain-elongation. Enzymatic replication of the template results in the synthesis of a mixture of radio-labelled fragments, all of which terminate with a particular ddNTP. PAGE purification of the fragments and subsequent autoradiography produces a pattern of bands, from which the DNA sequence can be inferred.

The advent of Sanger sequencing increased the speed at which sequence data was acquired and resulted in the establishment of the first DNA sequence repository by Walter Goad at Los Alamos National Laboratories in 1979 (now GenBank). However, the technique was far from optimal; it required unstable and hazardous radio-isotopes and the base-calling was done manually and was therefore laborious and expensive.

An advance on the traditional Sanger sequencing method was made in 1986 by Hood and colleagues, which involved the semi-automation of DNA

sequencing (Smith, Sanders et al. 1986). Firstly, the radio-isotopes were replaced. Instead, the primers were fluorescently labelled and in each of the four reactions, a differently labelled primer was used. For example, in the ddGTP reaction, a “red” primer was used and in the ddCTP reaction, a “green” primer was used etc. After the four separate reactions had been performed, the samples were pooled and the fragments subjected to PAGE, as in traditional Sanger sequencing, but only in one lane. Fragments produced in each of the four reactions produced different fluorescent signatures and provided information about which of the ddNTPs terminated a given fragment (in the example above, a “red” fragment is terminated with a G and a “green” fragment is terminated by C). The fluorescence data were acquired and stored on a computer in “real-time” (i.e. as the gel was running) and the base-calling also carried out by the computer. Hence, DNA sequencing had been automated. The new advances were so successful that the technology was commercialized by Applied Biosystems, in 1987. Subsequently, the fluorescently labelled primers were replaced and the ddNTPs themselves were labelled with different fluorescent tags, meaning that sequencing now only required a single reaction, with all four ddNTPs present (Prober, Trainor et al. 1987).

The throughput of this technique was significantly increased with the development of capillary array electrophoresis (CAE). A low viscosity separation matrix was pumped into an array of capillaries (eliminating the need for PAGE) and the samples run through at high voltages (as capillaries dissipate the heat associated with high voltages very efficiently). The new array format, coupled with the ability to operate at high voltages decreased the run-time and therefore

massively improved throughput (Figure 1.2). Such systems could deliver high quality reads of 500-1000 bases in hours. In order to obtain a bacterial genomic sequence, Fleischmann et al., randomly subdivided long genomic DNA fragments into shorter (more manageable) fragments which could be sequenced by the Sanger method. Several rounds of fragmentation and sequencing (so-called shotgun sequencing) provided multiple overlapping reads which were reconstructed to give the original sequence (Fleischmann, Adams et al. 1995). However, to sequence a human genome using this method was, and still is, an expensive business.

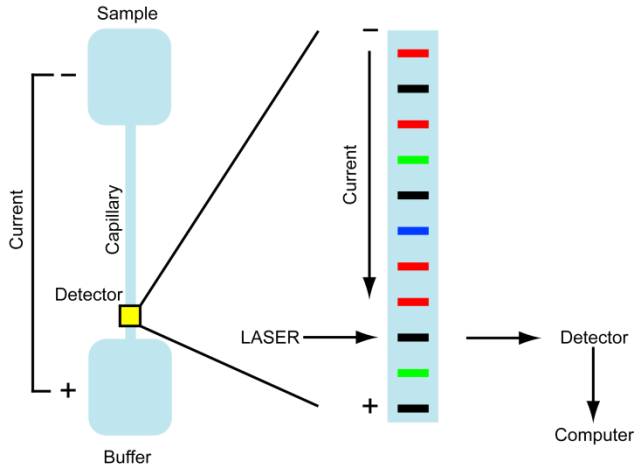



Figure 1.2. Capillary array electrophoresis Sanger sequencing. The radio-labelled dNTPs have been replaced with fluorescently labelled ddNTPs, meaning that all four ddNTPs can be mixed together and only a single sequencing reaction is required. The slab gel has been replaced with an array of capillaries. These advances allow for higher throughput, compared to the first iteration of Sanger sequencing. As the fluorescently tagged fragments are separated in a capillary, they pass a laser which excites the fluorophore of the terminated chain. The fluorescence signal is detected and the computer automatically makes the "base-call" in real time.

## 1.3. Next generation DNA sequencing

On the road to the $1000 genome, the NIH set an interim goal of the $100,000 genome. Despite vast improvements in Sanger sequencing technology, it would be surprising if such a target could be met using this method. Therefore, in order bring the cost of sequencing a complete human genome to below the $100,000 level, new highly parallel approaches to DNA sequencing were developed and since 2005 several next generation technologies have been made commercially available. A number of these technologies make use of so-called emulsion PCR, rather than the conventional bacterial sub-cloning and subsequent PCR amplification that is required in Sanger sequencing. Initially, in emulsion PCR, the target genome is randomly fragmented. Typically, streptavidin-coated beads are then incubated with biotinylated oligonucleotides (to generate beads which are studded with PCR primers). Template fragments and the primer-studded beads are stirred together with the other components required for PCR in an oil/detergent mix to create aqueous droplets, in oil, each of which contains an average of less than one template molecule and less than one bead. In the droplets that contain both a single template molecule and a single bead, the primers on the bead will act as an initiation point for PCR. A conventional PCR thermocycle is applied and many copies of the fragment are produced. The amplified products are tethered to the beads, through the bound primers. The beads are then purified and denaturation of dsDNA produces individual beads, each of which is studded with multiple

copies of the same ssDNA sequence (Dressman, Yan et al. 2003). The bead-bound ssDNA is then subjected to sequence determination.

### 1.3.1. Sequencing by ligation

One of the next generation DNA sequencing technologies is called sequencing by ligation and was pioneered by Church and co-workers (Shendure, Porreca et al. 2005). In this study, thousands to millions of beads were generated in which each bead was studded with multiple copies of the same ssDNA fragment, using emulsion PCR (different beads were studded with different fragments). The beads were then arranged on a chip to create an array of template DNA features. The array chip was then subjected to a series of flow-cycles. Initially, “anchor primers” were annealed at complementary adaptor sites within the amplified target DNA. Subsequently, a mixture of four fluorescently-labelled, nonomeric “query primers” were added (Figure 1.3). The query primers were fully degenerate, except at one position, the query position, where the identity of the base would correlate with the emission properties of the attached fluorophore. Only the complementary query primers would anneal to the target molecules on a bead and be ligated to the anchor primers. Excess, non-bound query primers were washed away and the fluorescence of the ligated query primers was measured, and hence the identity of the non-degenerate, query base detected (Figure 1.3). The array contained thousands to millions of template DNA features and the fluorescence detected at any given feature gave sequence information for that particular template fragment. As each feature/bead contained many copies of the same target sequence, many copies of the same

complementary query primer will have annealed and be ligated. Thus, the fluorescence signal detected from each feature has contributions from multiple fluorophores. The anchor•query primer oligonucleotides were stripped from the target strands, and the cycle repeated with a fresh set of anchor primers and a different set of query primers, in which the query position (the position of the non-degenerate base) was altered so that a different position in the template could be addressed. Additional repeat cycles allowed for the sequencing of many positions with a low error-rate, although the read-length was limited to 26 in the study by Shendure and colleagues. Due to the vast arrays that were possible the entire *E.coli* genome could be sequenced in two 60 hour instrument runs (Shendure, Porreca et al. 2005). This sequencing by ligation technique was commercialized by Applied Biosystems; the platform was called the Support Oligonucleotide Ligation Detection System, or The SOLiD$^{TM}$ System.

### 1.3.2. Sequencing by synthesis

This term describes an array of methods which involve the sequencing of target DNA polynucleotides, by identifying the incorporation of bases into complementary strands as a target strand is copied.

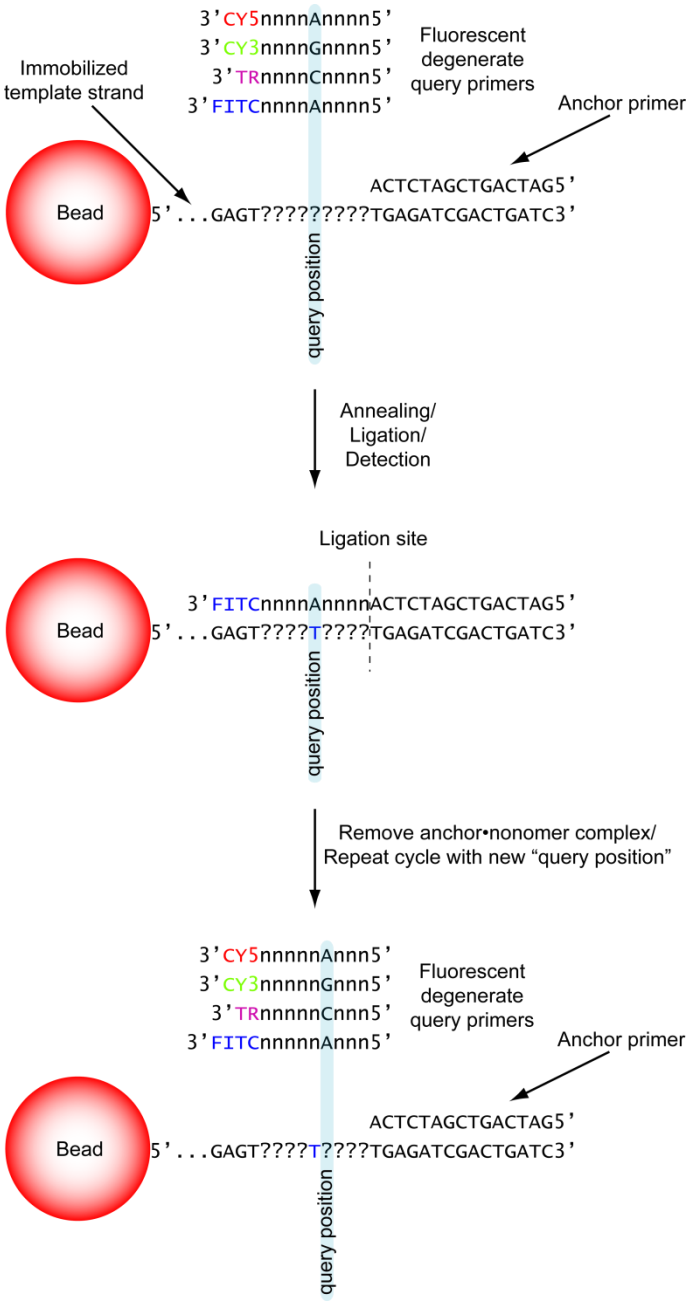



Figure 1.3. Sequencing by ligation. ssDNA, amplified by emulsion PCR is immobilized on beads (many copies of the same fragment per bead, although just a single copy is shown for illustration purposes) and an anchor primer is annealed to an adaptor sequence. Four fluorescent, degenerate, query primers are added and only the primer with the complementary base at the query position will anneal and be ligated. The non-bound excess query primers are washed away and the fluorescence signal of the ligated query primer is detected to give the identity of the query position. The anchor•query oligonucleotide is stripped off and the cycle repeated with a new query position. Figure adapted from (Shendure, Porreca et al. 2005).

#### 1.3.2.1. Cyclic reversible termination

This technique was initially developed by Solexa Inc., and they launched the commercial Genome Analyzer in 2006. Solexa Inc. has subsequently been acquired by Illumina Inc. in 2007 (www.illumina.com). Rather than using emulsion PCR to amplify template DNA, this technique employs so-called “solid-phase amplification” (Figure 1.4). Fragmented genomic DNA has adaptor sequences ligated to both ends. Subsequently, ssDNA (with adaptor sequences ligated to both ends) is added to a flow cell and randomly binds to the surface. One end of the DNA is bound to the surface, via one adaptor sequence, and the other end of the DNA has a free adaptor sequence. The surface of the flow cell is coated with primers, which are complementary to the adaptor sequences. The ssDNA which is immobilized to the surface of the flow cell forms a bridge so that the free adaptor sequence binds to an immobilized primer. The primer acts as the start point for a polymerase enzyme which synthesizes a complementary strand to form dsDNA. The two strands in the duplex are each immobilized to the surface at one end, and have a free terminus. Denaturation of this duplex creates two complementary ssDNA molecules. This cycle can be repeated to generate amplified clusters of each DNA that was deposited on the chip. This occurs at several million different places on the same chip, to generate several million clusters of each original genomic fragment. This is analogous to the array of bead-bound DNA features employed in the SOLiD$^{TM}$ system. After cluster generation, the so-called sequencing by synthesis begins; primers and the four dNTPs are added. Each dNTP is labelled with a different, reversible, chain-terminating fluorophore. At each cluster, the complementary dNTP is

incorporated into the growing strand by a DNA polymerase, and this terminates the chain. Every copy of the target DNA, within a single cluster, is simultaneously copied, so there are many single incorporation events, per cluster. The identity of the incorporated base at each cluster is determined by total internal reflection fluorescence (TIRF). The fluorescence signal that is detected from a single cluster is a summation of the signal emitted from each fluorophore on each incorporated base. The fluorophores are cleaved to allow the next incorporation and the cycle is repeated (Figure 1.4).

#### 1.3.2.2. Pyrosequencing

Another next generation sequencing technique is a variant of sequencing by synthesis, and is termed pyrosequencing. Using pyrosequencing, Margulies et al. were able to sequence the *Mycoplasma genitalium* genome (Margulies, Egholm et al. 2005). As in sequencing by ligation, emulsion PCR was used to generate the template DNA and after amplification, the beads, each of which is covered with immobilized DNA (many copies of the same sequence per bead) were deposited in wells on an array chip. The subsequent sequencing reaction is based on the controlled synthesis of strands which are complementary to the bead-immobilized target molecules; one of the four dNTPs is washed over the wells and if incorporated, would result in the release of pyrophosphate. The release of pyrophosphate is detected as photon release (by a CCD) due to the presence of the luciferin/luciferase system (Figure 1.5). PPi is converted to ATP, which is converted to AMP and light by luciferase.

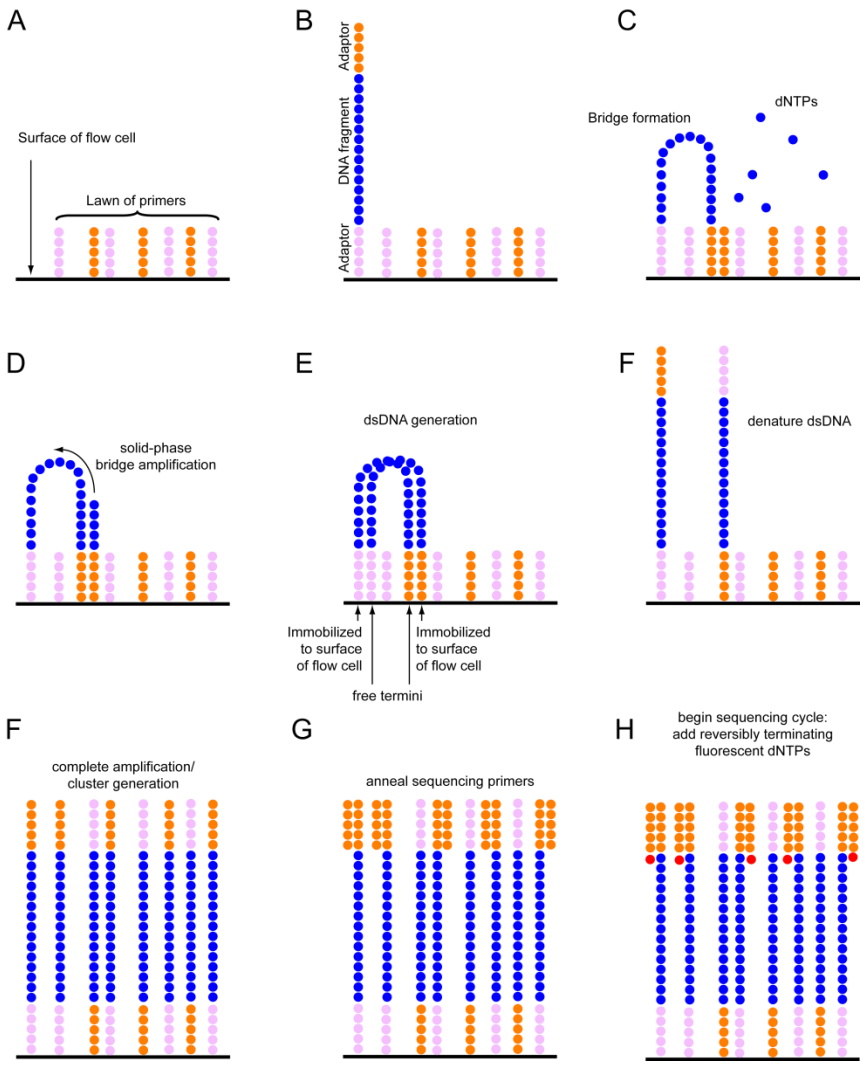



Figure 1.4. Sequencing by synthesis using cyclic, reversible chain-termination. A) A flow-cell, with a lawn of primers. B) Genomic ssDNA fragments (with ligated adaptor sequences) bind randomly to the flow cell surface. C) The DNA forms a bridge structure, as the adaptor sequence at the free end anneals with an immobilized primer. D) The ssDNA is amplified (with non-labelled nucleotides) to produce duplexed dsDNA. E) Each of the two strands in the dsDNA that is formed is tethered to the surface at one end, with the other end free. F) Denaturation disrupts the duplex and liberates the free ends, which are now free to form new bridges. F) Cycles of the solid-phase bridge amplification (panels C-E) are repeated to generate a cluster of identical template molecules. G) Sequencing primers are added, which anneal to the adaptor sequence. H) Fluorescently labelled, reversible, chain-terminating dNTPs are added and complementary dNTPs are incorporated into the growing chain. The fluorescence signal is detected from each of the incorporated bases and the terminating fluorophore is cleaved to allow the next incorporation. Figure adapted from www.illumina.com.

To minimize the background signal, dATPαS is used as the template dNTP in the sequencing reaction instead of dATP, as dATPαS is not as good a substrate for luciferase as dATP and will prevent the production of false positives. Unbound dNTP was washed away, the next nucleotide flowed in and base incorporation measured. Many photons are generated per correct base incorporation, as each bead contains multiple copies of the same sequence (although only one base is incorporated per strand). The signal detected is the sum of the photons released from each single-base incorporation. When there is a homopolymeric stretch in the target sequence, several bases will be incorporated, per strand when the correct dNTP is added. This will result in more photons being released, compared to a single incorporation. As more photons are released, the intensity of the signal will increase and is proportional to the number of nucleotides, providing information on the length of the homopolymer. However, once the stretch is over 8 nucleotides long, then the strength of the signal is no longer proportional to the number of nucleotides, and therefore cannot give a reliable measure of the length of the homopolymer. Repeated cycles produced read-lengths of ~100 per immobilized bead. As with other next generation sequencing techniques, pyrosequencing is highly parallelized and chips contain ~$10^6$ wells (each well containing one bead and each bead has many copies of the same sequence bound to it) (Margulies, Egholm et al. 2005). Originally, this technology was marketed by 454 Life Sciences Corp., which has subsequently been acquired by Roche.

Ion Torrent (www.iontorrent.com), a company recently acquired by Life Technologies, employs a similar sequencing chemistry; identifying the

incorporation of single nucleotides during the controlled synthesis of strands which are complementary to a template. Unlike pyrosequencing, which detects the release of PPi upon cognate dNTP incorporation, Ion Torrent's platform detects the release of the $H^+$ ion from the 3' OH at the terminus of the growing complement.

.

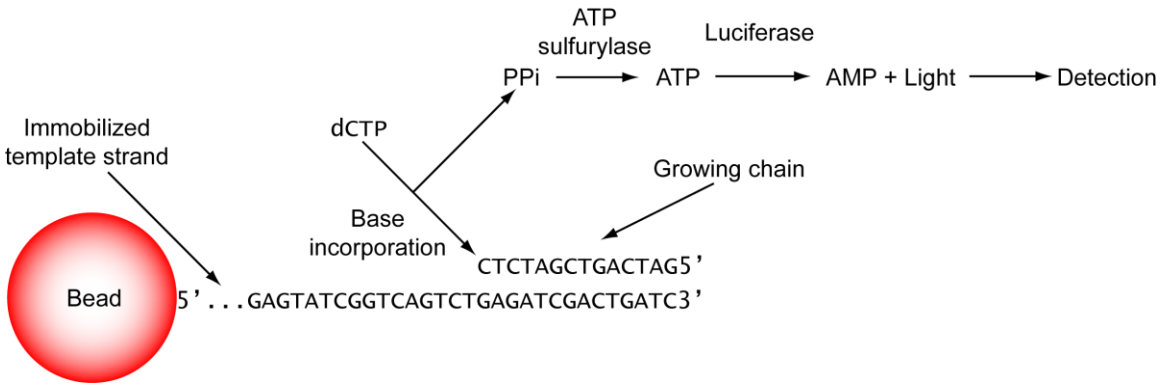



Figure 1.5. Sequencing by synthesis using pyrosequencing. ssDNA is amplified by emulsion PCR and immobilized on beads which are then deposited in wells on an array chip. Sequencing primers are added and subsequently, cycles of individual dNTPs are added. If the complementary dNTP has been added, it will be incorporated into the growing chain and PPi will be released. PPi is converted to ATP by ATP sulfurylase and ATP is converted into AMP and light by luciferase. The light, which is detected by a CCD, indicates the incorporation of a dNTP. In the example depicted, as dCTPs are flowed into the reaction, a dCTP is incorporated and light will be detected. Figure adapted from (Bayley 2006).

## 1.4. Single-molecule next generation sequencing

The massively parallel nature of the next generation sequencers, coupled with the replacement of the traditional bacterial cloning and PCR amplification allowed for a huge increase in the throughput of DNA sequencing. Although emulsion PCR and solid-phase amplification do offer advantages over bacterial cloning, the requirement for any sort of PCR is far from ideal; the techniques are cumbersome and require large amounts of genomic DNA material (3-20 µg –

(Metzker 2010). A single human cell contains ~3 pg of DNA and therefore, thousands of cells are required. Although this is not a problem when taking DNA from a patient (buccal swabs or blood samples would contain ample DNA), sequencing DNA where material is scarce, e.g. in fossils or in forensics would be problematic). Furthermore, PCR generates mutations that will be detected as sequence variants (i.e. it can create false positives). Also, there is a bias in the amplification yield of AT and GC-rich regions which may result in the under-representation of certain sequences, making it difficult to piece together overlapping fragments into a whole genome. It would therefore be advantageous to use single-molecule templates for sequencing, as sample preparation is more straightforward, less template is required, and no errors or bias from PCR are introduced. In addition to this, sequencing single molecules will eradicate the problem of "dephasing" which affects second generation techniques. For example, sequencing by synthesis relies on multiple copies of the same sequence (e.g. on a bead, or in a cluster) being synthesized in parallel, with each base addition occurring at the same point and the signals summing to allow for detection of the incorporated base. Over time and over multiple base additions, the sequencing reaction within a cluster/bead loses synchronisation due to mis-incorporations and becomes out of phase. As the sequencing reaction proceeds there is an accumulation of this dephasing until the point where sequence information can no longer be accurately determined. This can lead to errors in sequencing and is one of the main limitations to the read-length offered by such technologies. However, this limitation can be overcome by sequencing single

molecules. Such methods have been developed and are now being commercialized.

### 1.4.1. True single-molecule sequencing (tSMS™).

This technique has been commercialized by Helicos Biosciences and is based on the work of Stephen Quake and colleagues, use of fluorescently labelled nucleotides. Unamplified target DNA is labelled with a poly(dA) tail which pairs with poly(dT) oligonucleotides that are immobilized on a glass cover slip. The sequence of the individual target DNA molecules is then determined by the cyclic addition of single, fluorescently labelled dNTPs. The addition of nucleotides to the poly(dT) primer, depending on complementarity, is imaged by TIRF. The fluorescent tag is cleaved and is proceeded by the addition of the next nucleotide (Figure 1.6) (Pushkarev, Neff et al. 2009). Multiple four-base cycles results in generation of complementary strands and provides read-length of >25 per individual template (www.helicos.com). Although amplification of the template DNA is not required and single target molecules are sequenced (eliminating the dephasing problem), the multiple flow cycles required to introduce bases, cleave and wash away the tags are time consuming.

### 1.4.2. Single-molecule real-time sequencing (SMRT™)

The single-molecule, sequencing by synthesis technique being developed by Pacific Biosciences involves the sequencing of single molecules of DNA, by single DNA polymerase enzymes in real-time (www.pacificbiosciences.com). In this technique, DNA polymerase enzymes are immobilized at the bottom of

special wells called zero-mode waveguides (ZMW) on so-called “SMRT chips”. At the bottom of each ZMW chamber is a hole, tens of nm in diameter, fabricated in a metal film (which has been deposited on a glass-slip). The chamber is constantly excited by a laser below, but because of the small size of the hole, visible light (which has a much longer wavelength than the diameter of the hole) is prevented from passing all the way through the chamber. Rather, the light exponentially decays as it enters the ZMW (c.f. microwaves decaying as they pass through the metal mesh of the oven door). Therefore, only the bottom 30-nm of the well are illuminated (Levene, Korlach et al. 2003).

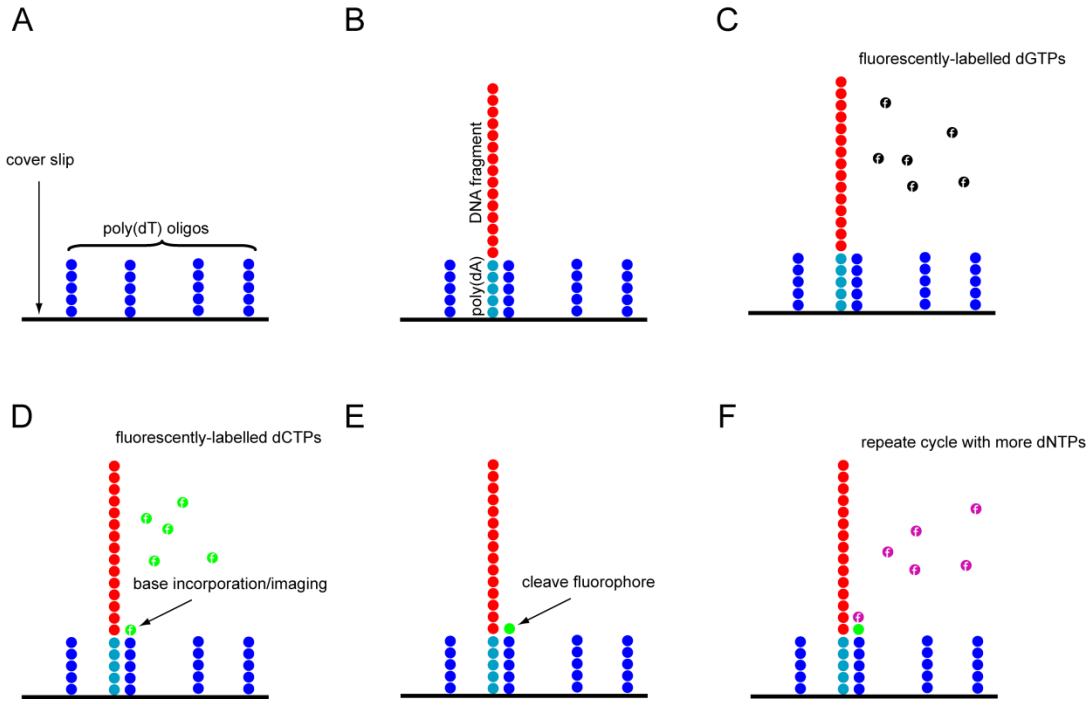



Figure 1.6. True single-molecule sequencing. A) poly(dT) oligonucleotides are immobilized on a cover slip. B) Template ssDNA fragments anneal to the poly(dT) sequences via their poly(dA) tail. C) Fluorescently-labelled dNTPs are added in cycles. D) Addition of a complementary dNTP leads to base incorporation and detection. E) The fluorescent group is cleaved. F) The cycle is repeated. Figure adapted from (Metzker 2010).

ssDNA templates are added and bind to the polymerase. The polymerase synthesizes copies of the template using four fluorescently-labelled dNTPs. The fluorescent-tag is attached to the phosphate group, and each dNTP has a different fluorophore attached. The binding of a cognate nucleotide to the DNA polymerase during synthesis brings the fluorescently-labelled dNTP into the ZMW chamber and the attached fluorophore is excited (by the laser) and this (and therefore the identity of the added base) is detected. The base is incorporated into the growing strand and the fluorophore on the terminal phosphate is released and diffuses away and out of the ZMW, and therefore no longer produces a fluorescence signal. The ZMW is crucial in preventing the illumination of non-bound nucleotides by the laser and preventing "background noise" although error rates are still predicted to be significant. The polymerase pulls the template along and the next cycle of cognate base binding, detection and incorporation can continue (Figure 1.7 (Eid, Fehr et al. 2009)). This repetitive process allows for single-molecule sequencing in real-time sequencing and is designed to exploit the high reaction rates and processivity of DNA polymerases to improve read-length and throughput, by eliminating the need for complex flow cycles to remove chain-terminating and/or fluorescent tags (c.f. tSMS$^{TM}$). However, as the synthesis of each DNA molecule is being performed by single enzymes that are constantly subjected to excitation by a laser (to excite the fluorophores on the cognate dNTPs), the enzyme is susceptible to photo-damage, which can retard the read-length.

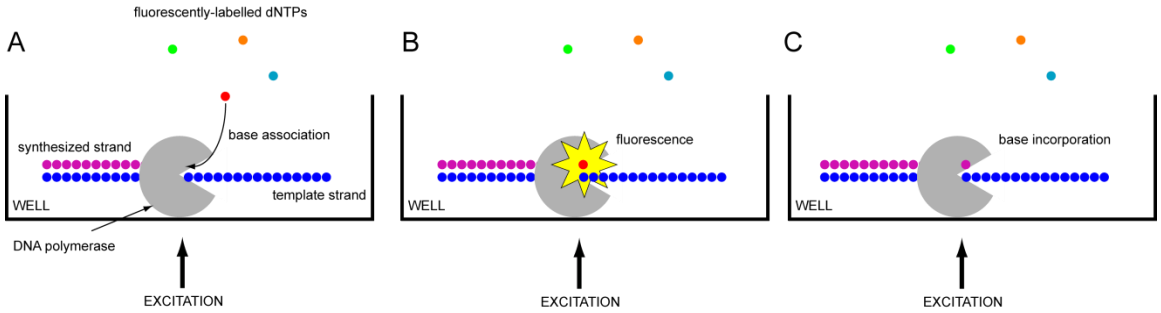



Figure 1.7. Single-molecule, real-time, sequencing by synthesis. A) A DNA polymerase is immobilized in a ZMW well (which is excited by a laser) and binds to the ssDNA template. dNTPs which are labelled on the terminal phosphate with a fluorescent tag, are added. Watson-Crick base-pairing ensures that only cognate bases will bind to the polymerase. B) As the cognate base associates with the polymerase, it is excited and will fluoresce. C) The base is incorporated and the fluorophore is cleaved and diffuses away and the cycle can continue.

## 1.5. Nanopore sequencing

Conducting pores that have nanometer dimensions are referred to as nanopores, and it has been proposed that such nanopores could be used to sequence single molecules of DNA. In one proposed means of nanopore sequencing, a potential is applied across a nanopore, which separates two chambers containing electrolyte solution, and this will electrophoretically drive both an ionic current and a charged ssDNA molecule through the pore. As bases pass a recognition site within the pore, the ionic current flow is modulated in a base-dependent manner; in other words, a cheap electrical read-out of the DNA sequence would be obtained (Kasianowicz, Brandin et al. 1996). Such a nanopore-based system would be advantageous over existing sequencing technologies for several reasons. Firstly, sample preparation is expected to be minimal (no need to attach adaptor sequences), there is no requirement for expensive reagents (e.g. fluorescent tags), or state-of-the-art imaging and data

processing equipment. Therefore, it is likely that the cost of nanopore sequencing would be far lower than any of the approaches discussed above. Furthermore, there is theoretically no limit to the read-length that can be achieved; the only practical limitation would be on the length of a DNA molecule that could be threaded through the nanopore (Branton, Deamer et al. 2008) and if an array of nanopores could be produced, then the throughput of nanopore sequencing would rival the high-throughput next generation techniques. Another advantage of nanopore sequencing is that it should be capable of direct determination of modified bases, notably 5-methylcytosine (Clarke, Wu et al. 2009; Wallace, Stoddart et al. 2010) and 5-hydroxy-methylcytosine (Wallace, Stoddart et al. 2010; Wanunu, Cohen-Karni et al. 2010). Such epigenetic sequence information is often lost in other sequencing technologies when the target is amplified or during sequencing by synthesis.

The nanopores employed could be protein pores, such as the *Staphylococcus aureus* toxin, α-hemolysin (αHL) (Song, Hobaugh et al. 1996), or the *Mycobacterium smegmatis* porin A (MspA) (Faller, Niederweis et al. 2004). Both these pores have dimensions that would allow the passage of ssDNA through the pore, where bases would pass through the pore in strictly single-file, but would not permit dsDNA to translocate through. The αHL nanopore has been studied extensively as a stochastic sensor (Bayley and Cremer 2001; Bayley, Luchian et al. 2008; Hammerstein, Shin et al. 2010; Lu, Li et al. 2010) and has been shown to be extremely robust; the pore is stable and functional at high temperature (Kang, Gu et al. 2005), high pH (Maglia, Henricus et al. 2009) and high concentrations of urea (Japrung, Henricus et al. 2010). Furthermore, the

pore is able to withstand extensive genetic modifications (Cheley, Braha et al. 1999; Maglia, Rincon Restrepo et al. 2008; Maglia, Heron et al. 2009). The αHL protein is synthesized and secreted as water-soluble monomers, but oligomerizes on target membranes to form homoheptameric pores (Figure 1.8). The protein consists of two domains; a globular cap which contains an approximately spherical, water-filled vestibule and a fourteen-stranded, antiparallel β barrel transmembrane domain. The opening of the pore at the cap domain is referred to as the *cis* entrance, whilst the opening at the end of the β barrel is referred to as the *trans* entrance. The narrowest part of the pore is at the "top" of the β barrel, and is called the central constriction (Song, Hobaugh et al. 1996).

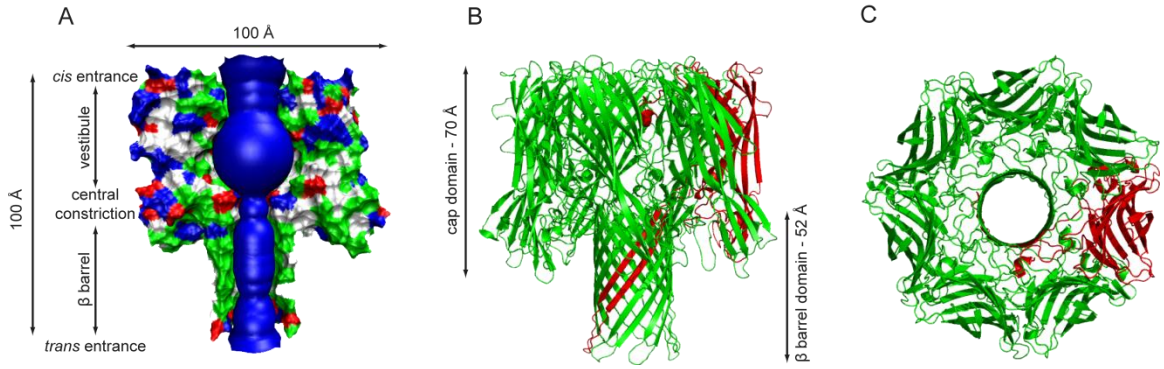



Figure 1.8. The α-hemolysin nanopore. A) A water-filled, cut-through representation of the homoheptamer is shown (J. Klingelhoefer – personal communication). Negatively charged amino acids are coloured red, while positively charged amino acids are coloured blue. The central constriction divides the protein into two different domains, an upper cap, which contains a spherical vestibule, and the lower transmembrane β barrel domain. The pore is ~100 Å tall and ~100 Å wide (Song, Hobaugh et al. 1996). B and C) Cartoon representations of the αHL heptamer are shown (with one protomer highlighted in red) from the side (B) and from above/the *cis* side (C). The side view (panel B) shows that the β barrel domain is partly recessed within the cap domain.

In 1996, Deamer and colleagues showed the ssDNA would traverse the pore, but dsDNA would not (Kasianowicz, Brandin et al. 1996). It has been proposed that base identification would be likely to occur within the narrow confines of the β barrel, rather than in the comparatively large vestibule (Figure 1.9A). However, the β barrel is approximately 5 nm long and studies by Meller et al. have suggested that around 12 nucleotides are likely to be contained within the β barrel (this is an internucleotide distance of ~0.4 nm) (Meller, Nivon et al. 2001) and this presents a problem; how does one identify a single nucleotide in the β barrel over the others? With this in mind, a second method of nanopore sequencing with αHL nanopores was proposed, whereby an exonuclease enzyme would clip off individual nucleobases from a template strand, and each base would be captured, in turn, by the pore and would translocate through the pore and be sequenced (Figure 1.9B). Indeed, it has been shown that all four dNMPs can be discriminated, based on differing current blocks, by genetically engineered αHL nanopores that have been fitted with a molecular adaptor (Astier, Braha et al. 2006; Clarke, Wu et al. 2009).

The structure of the MspA pore was recently deduced by X-ray crystallography (Figure 1.10), and like αHL, is reported to have suitable dimensions for nanopore sequencing (Faller, Niederweis et al. 2004). Indeed, it has been proposed that MspA may have superior resolution compared to αHL for “strand sequencing” (Figure 1.9A).

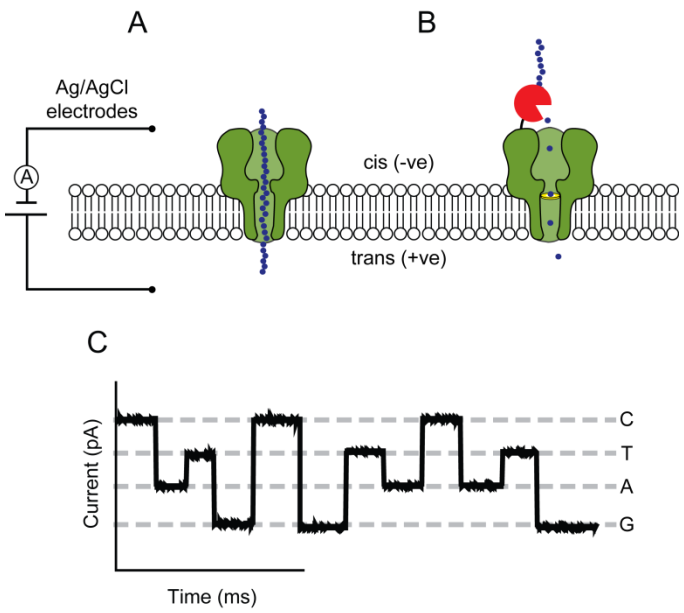



Figure 1.9. Nanopore sequencing with the α-hemolysin (αHL) nanopore. A) Whole ssDNA (blue circles) strands are driven through the αHL protein (green) under an applied potential, and as the bases pass a recognition point, the current is modulated to give a current read-out of the DNA sequence (panel C). B) Alternatively, an enzyme (red) is attached to the pore, and clips off bases one-by-one, which are driven through the pore, which has been fitted with an adaptor (yellow). C) An example of a hypothetical electrical read-out of a target DNA sequence.

The 5 nm-long β barrel in αHL is unlikely to be able to resolve single nucleobases (as multiple bases will reside within the β barrel) and this perceived drawback may be overcome by MspA which has a much shorter "constriction region" (~0.5 nm) and therefore may be able to resolve single bases in a translocating strand (Butler, Pavlenok et al. 2008; Derrington, Butler et al. 2010). However, it has been proposed that even an "infinitely short" channel would not achieve such single-nucleobase resolution as the high electric field region that determines the electrical "read" region of the channel will extend for approximately one channel diameter to each side of the channel (Aksimentiev, Heng et al. 2004; Aksimentiev and Schulten 2005; Muthukumar and Kong 2006;

Liu, Qian et al. 2007; Branton, Deamer et al. 2008); i.e. the read region of the MspA pore = length + (2 x pore diameter) = 0.5 nm + 2.4 nm = ~3 nm. Given an approximate internucleotide distance of ~0.4 nm, then the so-called “read region” of the MspA pore may also address multiple bases (~6-7 bases).

Alternatively, the nanopores employed could be fabricated in insulating materials, such as silicon nitride, silicon oxide, metal oxides or graphene, using ion beam sculpting (Li, Gershow et al. 2003), e-beam drilling (Storm, Chen et al. 2003) or atomic layer deposition (Chen, Mitsui et al. 2004). The main advantage of these solid-state nanopores is their superior stability over the delicate bilayer support required for protein nanopores. However, as yet, reproducible fabrication of solid-state nanopores with an appropriate diameter (1.5–2.0 nm) has not been achieved (Dekker 2007).

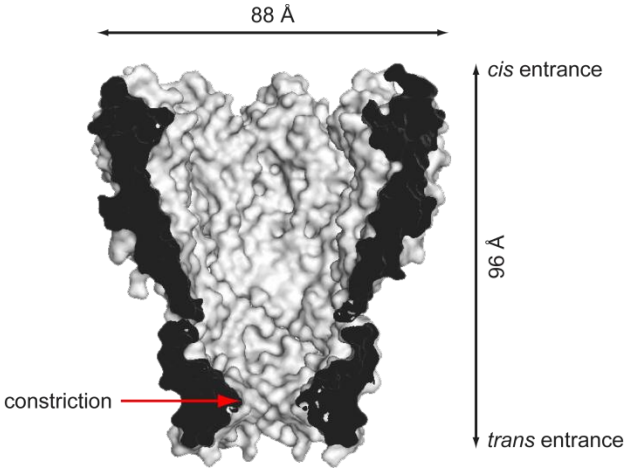



Figure 1.10. The MspA nanopore. The pore forms a wide opening at the *cis* entrance and tapers to a ~0.5 nm-long, narrow constriction of ~1.2 nm diameter at the *trans* entrance.

## 1.6. Challenges of nanopore sequencing

Although initial experiments proved promising (Kasianowicz, Brandin et al. 1996), several fundamental challenges need to be met before αHL nanopores can be used for sequencing molecules of genomic DNA (Branton, Deamer et al. 2008). For example, genomic DNA is double stranded, and dsDNA does not translocate through the αHL nanopore. Therefore, the genomic DNA must be converted from dsDNA to ssDNA. At high pH, dsDNA will become ssDNA and using a genetically engineered pore, under suitable pH conditions, dsDNA can be denatured and the ssDNA produced can be captured and translocated through the pore (Maglia, Henricus et al. 2009). Exonuclease enzymes could also be used as an alternative means to generate ssDNA. For example, Lambda exonuclease (λexo) is an enzyme that binds and processes dsDNA and releases 5'-mononucleotides and the non-hydrolyzed, complementary ssDNA strand as products (Little 1967; Little, Lehman et al. 1967). Therefore, it can be envisaged that genomic dsDNA fragments could be incubated with λexo to generate the necessary ssDNA template.

Whatever mechanism is used to generate ssDNA from duplexed genomic DNA, the resulting ssDNA needs to be captured and threaded through the pore. Single oligonucleotide chains are prone to form base-pair dependent intramolecular interactions (secondary structure), such as terminal hairpins and internal stem-loops. Such secondary structural motifs will not be translocated, and must be eradicated in order to thread the DNA molecule through the pore. Furthermore, the length of ssDNA that can be threaded through the pore will determine the read-length that can be achieved. Unpublished reports have

suggested that up to 25 kb of ssDNA can be threaded through protein nanopores (Branton, Deamer et al. 2008). If possible, this could provide unprecedented read-lengths for sequencing. However, the rate at which molecules of this size are captured by a protein nanopore is expected to be very low in practise, if evidence from RNA experiments is comparable (D. Japrung *in preparation*). Therefore, it is likely that the ideal fragment length will be a compromise between the requirement for high read-length and the requirement for efficient threading. It has been shown that the threading efficiency can be improved by genetic engineering (Maglia, Rincon Restrepo et al. 2008) and by altering the experimental conditions (Wanunu, Morrison et al. 2010).

The sequencing of strands of DNA requires that as an individual ssDNA molecule traverses the pore, it does so in such a way that the bases pass through in single-file, and that the speed and direction of translocation are constant. Stochastic DNA motion is reflected in the broad distribution of translocation times; transit times for DNA molecules of the same sequence can differ by an order of magnitude (Kasianowicz, Brandin et al. 1996; Meller, Nivon et al. 2000). This suggests that DNA translocation speed is not constant, and that interactions between the DNA and protein pore lead to “stick-slip” phenomena. This could lead to uncertainties in the number of bases that have translocated through the pore in a given time, and if there are periods where the DNA slips through the pore very quickly, then some nucleotides could be missed. Furthermore, it is likely that whilst within the pore, the DNA will temporarily move against the direction of the electrophoretic driving force through Brownian diffusion, although the net direction of translocation will be in the direction of the

driving force. Again, if diffusion does occur, then bases may pass recognition sites more than once, which could be detected as genetic insertions. Therefore, it will be necessary to control the kinetics of DNA translocation, such that it is constant, and preferably unidirectional. The use of DNA processing enzymes to feed, or pull DNA through the pore would presumably bring the kinetics of DNA translocation in line with the kinetics of the enzyme. In other words, provided that enzyme processed the DNA at a constant rate, then the DNA would also translocate at a constant rate. However, this may not resolve the problem of DNA diffusion within the nanopore. It is therefore likely that multiple scans of the same sequence would be necessary in order to obtain a consensus sequence. A consequence of the use of enzymes, would be that translocation would be rate-limited by the turnover rate of the enzyme (which could be expected to be ~10-100 base $s^{-1}$). One the one hand, this could be a disadvantage, as it limits the rate at which DNA sequencing can be achieved, but on the other hand, it will prevent the DNA translocating through the pore too fast to be sequenced.

Once the fundamental challenges of DNA manipulation (getting genomic dsDNA into ssDNA templates, and prevention of secondary structure, within the ssDNA templates), capture, threading and control of DNA translocation have been met, finally, the bases need to be identified. It is necessary that as the bases translocate through the pore, the ionic current flow is modulated in a way that correlates with the identity of the base (or bases) that are blocking the pore. The work described in this thesis addresses the issue of base identification in αHL pores.

References


Aksimentiev, A., J. B. Heng, et al. (2004). "Microscopic Kinetics of DNA Translocation through synthetic nanopores." Biophys J **87**(3): 2086-2097.

Aksimentiev, A. and K. Schulten (2005). "Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map." Biophys J **88**(6): 3745-3761.

Astier, Y., O. Braha, et al. (2006). "Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter." J Am Chem Soc **128**(5): 1705-1710.

Bayley, H. (2006). "Sequencing single molecules of DNA." Curr Opin Chem Biol **10**: 628-637.

Bayley, H. and P. S. Cremer (2001). "Stochastic sensors inspired by biology." Nature **413**: 226-230.

Bayley, H., T. Luchian, et al. (2008). Single-molecule covalent chemistry in a protein nanoreactor. Single Molecules and Nanotechnology. R. Rigler and H. Vogel. Heidelberg, Springer**:** 251-277.

Branton, D., D. W. Deamer, et al. (2008). "The potential and challenges of nanopore sequencing." Nature Biotechnology **26**: 1146-1153.

Butler, T. Z., M. Pavlenok, et al. (2008). "Single-molecule DNA detection with an engineered MspA protein nanopore." Proc Natl Acad Sci U S A **105**(52): 20647-20652.

Cheley, S., O. Braha, et al. (1999). "A functional protein pore with a "retro" transmembrane domain." Protein Sci. **8**: 1257-1267.

Chen, P., T. Mitsui, et al. (2004). "Atomic layer deposition to fine-tune the surface properties and diameters of fabricated nanopores." Nano Letters **4**(7): 1333-1337.

Clarke, J., H. Wu, et al. (2009). "Continuous base identification for single-molecule nanopore DNA sequencing." Nature Nanotechnology **4**: 265-270.

Dekker, C. (2007). "Solid-state nanopores." Nat Nanotechnol **2**(4): 209-215.

Derrington, I. M., T. Z. Butler, et al. (2010). "Nanopore DNA sequencing with MspA." Proc Natl Acad Sci U S A **107**(37): 16060-16065.

Dressman, D., H. Yan, et al. (2003). "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations." Proc Natl Acad Sci U S A **100**(15): 8817-8822.

Eid, J., A. Fehr, et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Faller, M., M. Niederweis, et al. (2004). "The structure of a mycobacterial outer-membrane channel." Science **303**(5661): 1189-1192.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Hammerstein, A. F., S. H. Shin, et al. (2010). "Single-molecule kinetics of two-step divalent cation chelation." Angew Chem Int Ed Engl **49**(30): 5085-5090.

Japrung, D., M. Henricus, et al. (2010). "Urea facilitates the translocation of single-stranded DNA and RNA through the alpha-hemolysin nanopore." Biophys J **98**(9): 1856-1863.

Kang, X., L.-Q. Gu, et al. (2005). "Single protein pores containing molecular adapters at high temperatures." Angew.Chem.Int.Ed.Engl. **44**: 1495-1499.

Kasianowicz, J. J., E. Brandin, et al. (1996). "Characterization of individual polynucleotide molecules using a membrane channel." Proc.Natl.Acad.Sci.USA **93**: 13770-13773.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Levene, M. J., J. Korlach, et al. (2003). "Zero-mode waveguides for single-molecule analysis at high concentrations." Science **299**(5607): 682-686.

Li, J., M. Gershow, et al. (2003). "DNA molecules and configurations in a solid-state nanopore microscope." Nat Mater **2**(9): 611-615.

Little, J. W. (1967). "An exonuclease induced by bacteriophage lambda. II. Nature of the enzymatic reaction." J Biol Chem **242**(4): 679-686.

Little, J. W., I. R. Lehman, et al. (1967). "An exonuclease induced by bacteriophage lambda. I. Preparation of the crystalline enzyme." J Biol Chem **242**(4): 672-678.

Liu, H., S. Qian, et al. (2007). "The effect of translocating cylindrical particles on the ionic current through a nanopore." Biophys J **92**(4): 1164-1177.

Lu, S., W. W. Li, et al. (2010). "A primary hydrogen-deuterium isotope effect observed at the single-molecule level." Nat Chem **2**(11): 921-928.

Maglia, G., A. J. Heron, et al. (2009). "Droplet networks with incorporated protein diodes show collective properties." Nat Nanotechnol **4**(7): 437-440.

Maglia, G., M. Rincon Restrepo, et al. (2008). "Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge." Proc Natl Acad Sci U S A **105**: 19720-19725.

Maglia, M., M. Henricus, et al. (2009). "DNA strands from denatured duplexes are translocated through engineered protein nanopores at alkaline pH." Nano Letters **9**: 3831-3836.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Meller, A., L. Nivon, et al. (2000). "Rapid nanopore discrimination between single polynucleotide molecules." Proc.Natl.Acad.Sci.USA **97**: 1079-1084.

Meller, A., L. Nivon, et al. (2001). "Voltage-driven DNA translocations through a nanopore." Phys.Rev.Lett. **86**: 3435-3438.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Muthukumar, M. and C. Y. Kong (2006). "Simulation of polymer translocation through protein channels." Proc Natl Acad Sci U S A **103**(14): 5273-5278.

Prober, J. M., G. L. Trainor, et al. (1987). "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides." Science **238**(4825): 336-341.

Pushkarev, D., N. F. Neff, et al. (2009). "Single-molecule sequencing of an individual human genome." Nat Biotechnol **27**: 847-852.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Smith, L. M., J. Z. Sanders, et al. (1986). "Fluorescence detection in automated DNA sequence analysis." Nature **321**(6071): 674-679.

Song, L., M. R. Hobaugh, et al. (1996). "Structure of staphylococcal a-hemolysin, a heptameric transmembrane pore." Science **274**: 1859-1865.

Storm, A. J., J. H. Chen, et al. (2003). "Fabrication of solid-state nanopores with single-nanometre precision." Nat Mater **2**(8): 537-540.

Venter, J. C. (2010). "Multiple personal genomes await." Nature **464**(7289): 676-677.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Wallace, E. V., D. Stoddart, et al. (2010). "Identification of epigenetic DNA modifications with a protein nanopore." Chem Commun (Camb).

Wanunu, M., D. Cohen-Karni, et al. (2010). "Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules." J Am Chem Soc.

Wanunu, M., W. Morrison, et al. (2010). "Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient." Nat Nanotechnol **5**(2): 160-165.

# CHAPTER 2: MATERIALS AND METHODS

## 2.1. Molecular biology – mutant generation, protein expression and purification

The following section outlines how the α-hemolysin (αHL) mutant genes were generated and the protein nanopores prepared. Also described are the various biochemical assays used to determine pore structure and function.

### 2.1.1. Mutant generation

#### 2.1.1.1. Plasmids

All αHL genes and constructs were assembled in the pT7 expression vector. See appendix for all gene and protein sequences.

#### 2.1.1.2. Mutagenesis by PCR

##### 2.1.1.2.1. Quik change

The αHL E111N/K147N (NN) gene was prepared by E. Mikhailova using a kit for site-directed mutagenesis (QuikChange II XL, Cat. #200522-5, Stratagene), using the WT αHL gene in pT7 as a template. Single point (or multiple point) αHL mutants, based on the NN mutant were also generated by E. Mikhailova, using the same kit and the NN αHL gene in pT7 as a template.

2.1.1.2.2. Homologous recombination

The αHL truncated barrel mutants (TBM) were generated from the NN gene in the pT7 vector by PCR mutagenesis and ligation-free *in vivo* recombination (Howorka and Bayley 1998). Briefly, two overlapping fragments of DNA are separately synthesized by PCR. The first piece of DNA (forward fragment) is generated by mixing template DNA with a mutagenic forward primer (mF) and a non-mutagenic reverse primer (nmR), and amplifying by PCR. The second piece of DNA (reverse fragment) is generated by mixing the template DNA with a mutagenic reverse primer (mR) and a non-mutagenic forward primer (nmF), and amplifying by PCR. The mF and mR prime from within the NN gene and are complementary in sequence to the antisense and sense strands, respectively; both contain the desired mutation sequence and have sections at their 5′ ends that are homologous to each other. The nmF and nmR (originally designed by S. Cheley) prime from outside of the NN gene and are complementary in sequence to the opposite strand from their partner (i.e. nmR, which is partnered with mF, is complementary to the sense strand whereas nmF, which is complementary to mR, is complementary to the antisense strand) and are also homologous to each other. The forward fragment and the reverse fragment are mixed together with competent cells, which when heat shocked, take up and recombine the DNA fragments (which have overlapping, homologous sequences at their termini) *in vivo* to generate a circular plasmid encoding the mutated αHL gene. (See Figure 2.1).

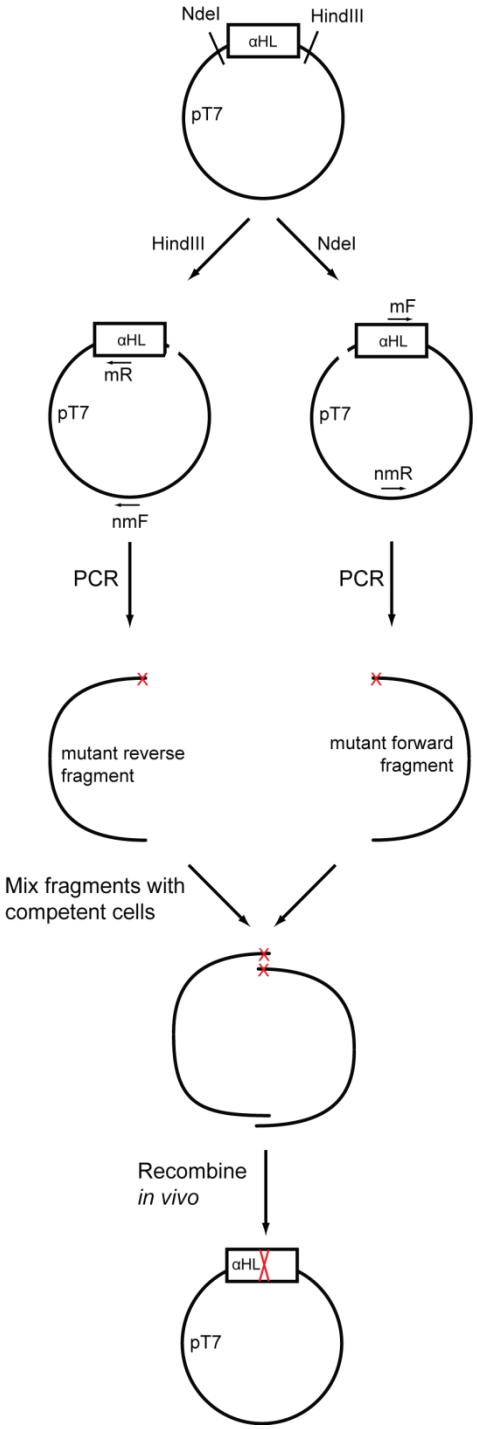



Figure 2.1. Scheme to show the principle of PCR mutagenesis and ligation-free *in vivo* recombination. Initially, two DNA fragments are generated by PCR amplification; mutagenic forward (mF) and non-mutagenic reverse (nmR) primers are mixed with template to yield a mutant forward fragment, whils mutagenic reverse (mR) and non-mutagenic forward (nmF) primers are also mixed to yield a mutant reverse fragment. The mutant fragments are mixed and are recombined *in vivo* by competent cells to re-generate a circular plasmid containing the αHL mutant gene.

This method of PCR mutagenesis requires a linear DNA template (rather than circular plasmid DNA. The carry-through of circular plasmids would generate false positive transformants). The pT7 plasmid, containing the NN gene was linearized with either NdeI (cuts the plasmid, upstream of the NN gene) or HindIII (cuts the plasmid, downstream of the NN gene) restriction endonucleases (NEB, UK). The reagents were mixed (see Table 2.1) and incubated in a water bath at 37°C for 1 hour (h). Each of the reaction mixtures was subsequently made up to 100 µL with the addition of water and stored at -20°C.

Table 2.1. Generation of linear DNA templates for mutagenesis by PCR and ligation-free *in vivo* recombination.

| Reagent | Volume (µL) |
|---|---|
| Water | 5.5 |
| NN in pT7 plasmid (400 ng $\mu L^{-1}$) | 2.5 |
| 10x buffer[1] | 1.0 |
| Enzyme | 1.0 |
| Total | 10.0 |

[1.] When the enzyme is NdeI, use NEB buffer #4 and when the enzyme is HindIII, use NEB buffer #2.

The mutant "forward" and mutant "reverse" linear fragments were generated, separately, by PCR amplification (the reaction conditions are shown in Tables 2.2 and 2.3, and the thermocycle employed is described in Table 2.4).

Table 2.2. PCR reaction to generate the mutant “forward” fragment for ligation-free *in vivo* recombination.

| Reagent | Volume (µL) |
|---|---|
| Water | 22.0 |
| NN in pT7 (linearized with NdeI)[1] | 1.0 |
| mF (50 µM) | 1.0 |
| nmR (50 µM) | 1.0 |
| 2x Phusion™ Flash PCR Master Mix[2] | 25.0 |
| Total | 50.0 |

1. The linear template was produced as described in Table 2.1.
2. Available from (NEB, UK)

Table 2.3. PCR reaction to generate the mutant “reverse” fragment for ligation-free *in vivo* recombination.

| Reagent | Volume (µL) |
|---|---|
| Water | 22.0 |
| NN in pT7 (linearized with HindIII)[1] | 1.0 |
| mR (50 µM) | 1.0 |
| nmF (50 µM) | 1.0 |
| 2x Phusion™ Flash PCR Master Mix[2] | 25.0 |
| Total | 50.0 |

1. The linear template was produced as described in Table 2.1.
2. Available from (NEB, UK)

The sequence of all primers used is shown in the Appendix (section A.2).

Table 2.4. Thermocycle applied for PCR using the Phusion™ Flash PCR system (NEB, UK).

| Cycle Step | Temperature (°C) | Time (s) | # of cycles |
|---|---|---|---|
| Initial denaturation | 98 | 10 | 1 |
| Denaturation | 98 | 1 | 30 |
| Annealing | 50 | 5 | |
| Extension | 72 | 15 s $kb^{-1}$ | |
| Final Extension | 72 | 60 | 1 |
| Hold | 4 | ∞ | N/A |

The success of PCR amplification was established by agarose gel electrophoresis of the amplified fragments. Fragments were run on a 1% agarose gel for 1 h, in TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH 8.3), at 100 V. Typically 5 µL of each of the mutant "forward" and mutant "reverse" fragments were mixed together with 100 µL of XL-10 Gold® Ultracompetent cells (Stratagene, USA, Cat # 200314), and incubated on ice for 30 minutes (min). The cell•DNA mix was heat-shocked by placing in a water bath at 42°C for 30 seconds (sec) and subsequently returned to ice. The heat-shocked cells were spread onto agar plates containing Luria broth medium and ampicillin (50 µg $mL^{-1}$) and incubated overnight at 37°C. Transformed colonies were picked and grown overnight, at 37°C (with shaking), in Luria broth containing ampicillin (50 µg $mL^{-1}$) and the DNA extracted using QIAGEN Plasmid Mini Kit (QIAGEN Ltd., UK, Cat # 27106).

The αHL TBM NN/M113G/Δ120-125/Δ133-138 (3G mutant) gene was also prepared by PCR mutagenesis and ligation-free *in vivo* recombination, using the NN/Δ120-125/Δ133-138 (TBM Δ3) gene, in pT7, as a template.

2.1.1.2.3. Sequencing

DNA samples obtained from QIAGEN Plasmid Mini Kit preparations of all αHL mutant genes were sequenced at the Source BioScience LifeSciences centre, located in the Department of Biochemistry, at the University of Oxford.

2.1.1.2.4. Plasmid amplification

After sequence verification, mutant constructs were amplified and subsequently purified using the QIAGEN Plasmid Maxi Kit (QIAGEN Ltd., UK, Cat # 12162). The concentration of the purified DNA was quantified using a NanoDrop® ND-1000 Spectrophotometer, and subsequently diluted to a final concentration of 400 ng $\mu L^{-1}$ in TE 8 (25 mM Tris.HCl, 0.1 mM EDTA, pH 8).

2.1.2. Protein expression and purification - coupled *in vitro* transcription and translation (IVTT)

αHL monomer was expressed in the presence of [$^{35}$S]methionine in a cell free *E. coli in vitro* transcription and translation (IVTT) system (*E. coli* T7 S30 Extract System for Circular DNA, Cat. #L1130, Promega) for 1 h at 37°C. The IVTT mix was centrifuged at 25,000 X g and the pellet discarded. To form heptamers, the supernatant from IVTT (containing the protein product) was incubated with rabbit red blood cell membranes (rRBCm) for 1 h at 37°C (Table

2.5). The membrane suspension was centrifuged at 25,000 X g and the pellet containing the heptamers was solubilized in loading buffer and run on a 5% SDS-polyacrylamide gel overnight at 40 V. The gel was vacuum-dried for 3 to 4 h onto Whatman 3M filter paper, without heating. The dried gel was exposed to photographic film for 2 h and the developed film was used to locate the position of the radio-labelled, heptameric protein in the gel. The protein-containing region of the gel was excised, rehydrated and crushed in 400 µL of TE 8. The polyacrylamide was removed by centrifuging the suspension at 25,000 X g for 10 min at room temperature through a cellulose micro spin column (Microfilterfuge tubes, Cat. #7016-024, Rainin) (Cheley, Braha et al. 1999; Stoddart, Heron et al. 2009). The concentration of heptamer, after purification, is typically ~ 1 ng $\mu l^{-1}$. Aliquots of the purified protein were stored at -80°C. Generally, WT, NN and other single (or multiple) point mutants were expressed and purified by E. Mikhailova.

### 2.1.3. Hemolysis assays

Rabbit red blood cells (rRBC) were washed repeatedly in MBSA (10 mM 3-[*N*-morpholino]propane sulphonic acid (MOPS), 150 mM NaCl, pH 7.4, containing 1 mg $mL^{-1}$ bovine serum albumin) and made up to ~1% in MBSA. Protein was translated by IVTT (without incubating the protein with rRBCm - see section 2.1.2.1) and the product centrifuged at 25,000 X g and the supernatant was retained.

Table 2.5. Conditions used to express αHL proteins by coupled *in vitro* transcription and translation.

| Reagent | Volume (μL) |
|---|---|
| Template plasmid (400 ng $\mu L^{-1}$) | 4.0 |
| S30 Premix | 10.0 |
| Amino acid mix[1] (1 mM) | 2.5 |
| [$^{35}$S]methionine[2] | 1.0 |
| T7 S30 Extract[3] | 7.5 |
| Total | 25.0 |

1. Amino acid mix (1 mM) was made by mixing equal volumes of amino acid mixture minus methionine (1 mM) with amino acid mixture minus cysteine (1 mM).
2. [$^{35}$S]methionine (1175 Ci $mmol^{-1}$, 10 mCi $mL^{-1}$, MP Biomedicals).
3. T7 S30 Extract was incubated with rifampicin (5 μg $mL^{-1}$) for 5 minutes on ice before usage.

The hemolytic activity of αHL proteins was then assayed in microtiter wells. Typically 5 μL of the retained supernatant was diluted 20-fold in MBSA. The protein mix was subsequently serially diluted two-fold with MBSA such that each well contained a volume of 50 μL. An equal volume of rabbit red blood cells (~1% in MBSA) was then added to each well, starting with the well which contained the most dilute protein sample. Lysis was recorded by monitoring the decrease in light scattered at 595 nm with a microplate reader (Bio-Rad Laboratories Ltd, UK) (Cheley, Malghani et al. 1997).

### 2.1.4. Proteolysis of α-hemolysin proteins with proteinase K

Fresh proteinase K (Sigma Aldrich – a broad spectrum serine protease) solution of 5.0 mg $mL^{-1}$ was prepared (from a 10 mg $mL^{-1}$ stock) (Cheley, Malghani et al. 1997). αHL proteins were prepared by IVTT, as described in section 2.1.2.1. The IVTT product was centrifuged at 25,000 X g, and the pellet discarded. For expression in the presence of rRBCm, the IVTT supernatant was mixed with rRBCm for 1 h at 37°C (as described). The membrane suspension was centrifuged at 25,000 X g and the pellet containing the heptamers was re-suspended in 10 µL MBSA. 1 µL of the supernatant (or membrane suspension, in the case of expression in the presence of rRBCm) was diluted 20-fold with MBSA and divided into two tubes (each containing 9 µL). To one of the tubes, 1 µL of the proteinase K solution was added and 1 µL of water added to the second tube, as control. Proteolysis was allowed to occur for 5 min at room temperature before being terminated by the addition of PMSF (phenylmethylsulfonyl fluoride – which is a serine protease inhibitor) to 2 mM (in excess of proteinase K). Gel loading buffer was added and the solutions were heated at 95°C for 5 min, to denature any oligomeric species, and the solutions were analyzed by electrophoresis on 12% SDS-polyacrylamide gels (Bio-Rad Laboratories Ltd, UK. Cat # 345-0118) (Cheley, Malghani et al. 1997).

### 2.1.5. PAGE in the absence of SDS

Polyacrylamide gel electrophoresis in the absence of SDS was carried out in order to determine the oligomeric structure and sensitivity to SDS of some αHL proteins, using the NativePAGE™ gel system (Invitrogen Ltd, UK). In this gel

system, Coomassie G-250 is used as the charge-shift molecule in place of SDS (Schagger and von Jagow 1991). Monomeric αHL proteins were prepared by IVTT in the absence of rRBCm (see section 2.1.2.1), and MBSA and loading buffer were added. When making heptameric protein, the soluble product from IVTT was mixed with rRBCm for 1 h at 37°C. The membrane suspension was centrifuged at 25,000 X g. The membrane pellet was re-suspended in MBSA, solubilized in 2.5% (final concentration) n-dodecyl-β-D-maltoside (DDM) and loading buffer was added. The samples were subjected to electrophoresis in 4-16% Bis-Tris NativePAGE™ gels (Invitrogen Ltd, UK. Cat # BN1002). Gels were "fixed" by shaking in 40% ethanol, 10% acetic acid for 1 hour and destained overnight in 8% acetic acid. The gel was vacuum-dried for 1 to 2 h onto Whatman 3M filter paper, with heating at 90°C. The dried gel was exposed to photographic film for 3 h and the developed film was used to locate the position of the radio-labelled protein in the gel.

## 2.2. Electrical recordings

The properties of the purified protein nanopores were principally analyzed by single-channel recording (SCR) experiments: single pores were reconstituted into synthetic lipid bilayers and the small currents flowing through the pores were recorded in the absence and presence of blocking analyte molecules. The following section describes the apparatus and techniques required to acquire, and subsequently analyze, these current measurements.

### 2.2.1. Electrical recording equipment

SCR measurements were acquired using the commercial Axopatch 200B patch-clamp amplifier, with a 1440A 16-bit digitizer (Molecular Devices). Experiments were conducted in the voltage-clamp mode, where the applied potential is fixed, and the resulting transmembrane current is measured. It was imperative to control the applied potential as interactions between analytes (DNA) and the nanopore are voltage-dependent. The transmembrane current that is measured is proportional to the conductance of the protein nanopore.

### 2.2.2. Faraday cage

Experiments were performed in a Faraday cage to shield the sensitive acquisition electronics from external electrical noise sources. The enclosure, a steel box with rigid side walls, was grounded via the common ground in the patch-clamp amplifier.

### 2.2.3. Electrodes

Two electrodes were used; a “working electrode”, or “active electrode” which connects the *trans* compartment (see below) to the recording equipment (see above) and a “ground electrode” which connects the *cis* compartment (see below) to the recording equipment (see above). Ag/AgCl electrodes were used and are advantageous due to their long-term chemical stability, reversible electrochemical behaviour, predictable junction potentials and their low-noise electrical performance (Figure 2.2A) (Janz 1961; Maglia, Heron et al. 2010). The electrodes were prepared from Ag wire, 1.5 mm in diameter (Sigma-Aldrich),

which had been cleaned in water and then ethanol and subsequently scrubbed with glass paper (to increase the surface area of the electrode). A thick coating of AgCl was generated by immersing the cleaned and roughened electrodes overnight in ~30% hypochlorite (NaClO) solution (Fisher Scientific) (Purves 1981). Before, and in between experiments, electrodes were rinsed in water and when not in use, were stored in the dark to prevent the UV conversion of AgCl, back to Ag.

The use of KCl/agarose bridges between the Ag/AgCl electrode and the electrolyte solution reduces junction potential offsets (Figure 2.2B). To create the KCl/agarose bridge, Ag/AgCl electrodes, with a rubber stopper to act as a seal, were incorporated into 200 µl pipette tips (Gilson, USA) which had been filled with molten, low-melt agarose at 3% (w/v) in 3 M KCl (Sigma-Aldrich, UK). The molten agarose was allowed to set before using the electrodes. A second advantage of the KCl/agarose bridge is the fact that it prevents $Ag^+$ ions from the electrode, leaching into the electrolyte buffer solution. The $Ag^+$ ions can influence the behaviour of some mutant pores However, no difference was observed between the KCl/agarose bridged electrodes and the “bare” Ag/AgCl electrodes, in the experiments described, and experiments were predominantly carried out with “bare” Ag/AgCl electrodes.

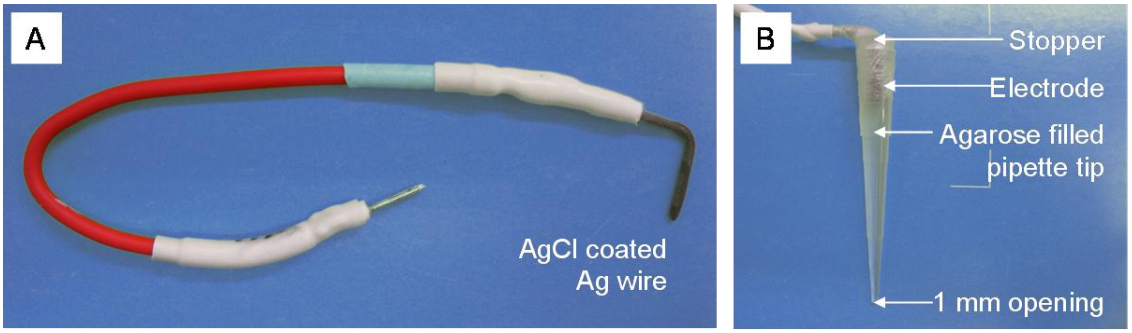



Figure 2.2. Ag/AgCl electrodes. A) Photograph of a Ag/AgCl electrode, prepared by hypochlorite treatment of a piece of 1.5 mm-diameter Ag wire (soldered onto standard electrical wire). B) To make agarose bridged electrodes, the bare electrodes were incorporated into 200 µL pipette tips filled with agarose (3% w/v, 3 M KCl), sealing the top with a rubber stopper and parafilm. Figure adapted from (Maglia, Heron et al. 2010).

### 2.2.4. Chambers

Open planar-bilayer chambers, machined from Delrin (Figure 2.3-left), were used for all the SCR measurements acquired and described in this thesis. Specifically, custom-built, vertical-bilayer devices where two compartments (*cis* and *trans*), each with a volume of 500 µL and separated by a 25 µm-thick polytetrafluoroethylene film (PTFE, Teflon) (Goodfellow, UK Cat. # FP301200/10) were used (Figure 2.3-right). PTFE film is advantageous, as it produces lower noise than some other plastics (Wonderlin, Finkel et al. 1990; Mayer, Kriebel et al. 2003). In this arrangement, both compartments were easily accessible for the purposes of adding reagents or solution perfusion (see below). The *cis* and *trans* compartments were connected via a small aperture in the PTFE film, across which bilayers were formed. To form the aperture, the PTFE film was weakened at the position where the aperture was required, by gentle indentation with a needle. The film was then “zapped” with a custom-built, high-voltage spark generator, which produces a spark ~ 15 mm long at a frequency of ~ 1 Hz. The

high temperatures associated with the spark passing through the weakened part of the PTFE film cause localized melting to produce a round aperture. The diameter of the aperture was measured under a microscope with a graticule and was typically ~100 µm (Mayer, Kriebel et al. 2003). Chambers were cleaned in between experiments by repeated rinsing firstly with water, then distilled water and finally 100% ethanol. The chamber was dried thoroughly under a gentle stream of nitrogen gas.

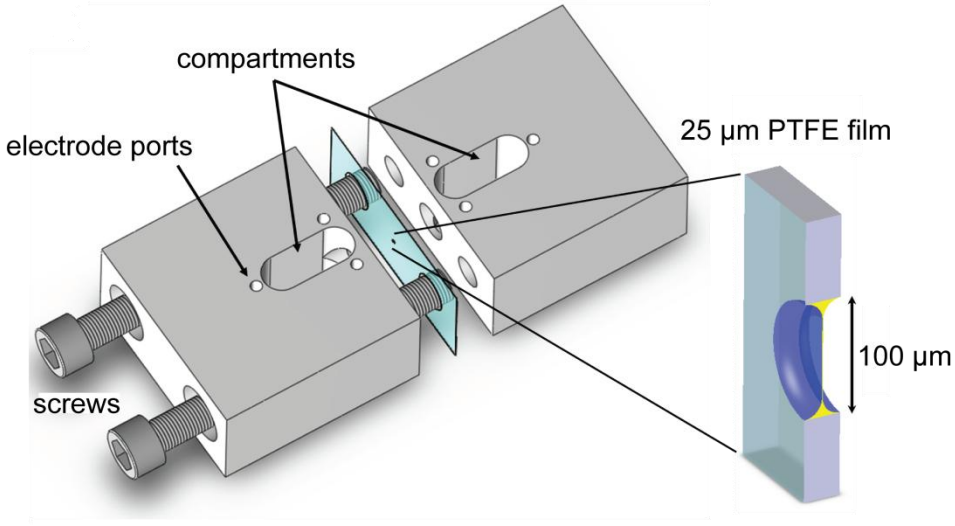



Figure 2.3. Custom-built, vertical planar-bilayer chamber (left), with two 500 µL compartments separated by a 25 µm PTFE film (shown in an expanded view - right). The two halves are fastened together with screws, clamping the PTFE film between them and sealed with silicone glue to ensure water-tightness. The PTFE film contained a central aperture ~100 µm diameter, across which the bilayers were formed. Figure adapted from (Maglia, Heron et al. 2010).

### 2.2.5. Preparing bilayers

To form planar lipid bilayers (Montal and Mueller 1972), the aperture was initially “pre-treated” with ~5 µL hexadecane (10% (v/v) in pentane). The hexadecane oil creates an annulus around the rim of the aperture, which allows the bilayer to form an interface with the PTFE film (Plateau-Gibbs border). The

pentane was allowed to evaporate (~20 s) and both the *cis* and *trans* compartments were filled with 500 µL buffered, aqueous electrolyte solution (typically 1 M KCl, 25 mM Tris.HCl, 100 µM EDTA, pH8), and the Ag/AgCl electrodes (Figure 2.2) were connected via the electrode ports, such that the *cis* compartment was connected to ground. At this point, the *cis* and *trans* compartments were not electrically isolated, as no bilayer existed to seal the aperture and therefore a current could be measured between the *cis* and *trans* compartments (even when no potential was applied by the patch-clamp amplifier, a small potential difference existed between the two compartments resulting in a flow of current). If no current was measured it meant that the aperture was blocked with excess oil, which prevented bilayer formation. Gentle pipetting of the solution near the aperture was usually sufficient to remove the excess oil, such that a current could be observed. Subsequently, a ~5 µL drop of 1,2-diphytanoyl-*sn*-glycero-3-phosphocholine (DPhPC, Avanti Polar Lipids) at 10 mg $mL^{-1}$ (w/v) in pentane was applied to the top of the aqueous phase of both the *cis* and *trans* compartments. A lipid monolayer would form at the water-air interface. Again, the pentane was allowed to evaporate and then the solution was pipetted, to lower the lipid monolayer past the level of the aperture. The solution level was then slowly raised back up, past the aperture. This was repeated in each compartment until a bilayer formed (Figure 2.4). Bilayer formation was detected as a current measurement of zero pA, indicating that an electrical seal had formed between the *cis* and *trans* compartments.

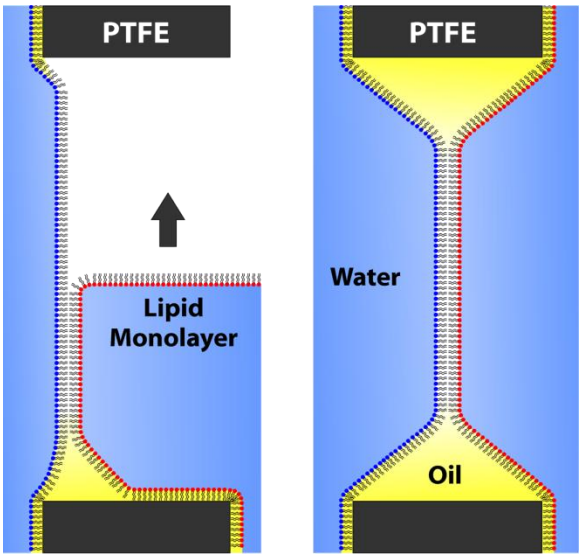



Figure 2.4. The formation of planar lipid bilayers by the Montal-Mueller method (Montal and Mueller 1972). The aperture is pre-treated with hexadecane, and the bilayer is then formed by flowing lipid monolayers over both sides of the aperture. Figure adapted from (Maglia, Heron et al. 2010).

Bilayer capacitance measurements were made to ensure that the bilayer was of sufficient size. The capacitance was measured by the application of voltage ramps to elicit constant capacitive currents (Figure 2.5). Since current ($I$) is proportional to the rate of change of voltage ($dV/dt$) for a capacitor ($I = C \times dV/dt$), an applied triangular voltage waveform produces a square-wave current output (i.e. $dV/dt$ = constant, for each sweep), which was used to determine bilayer capacitance ($I_C$ in Figure 2.5). A commercial analog waveform generator was used to input a triangular wave (~20 Hz) into the patch-clamp amplifier (front-switched input on the Axopatch 200B amplifier). The input waveform was directly calibrated by using a 100 pF bilayer model cell (MCB-1U, Molecular Devices), so that 100 pA = 100 pF. This approach provided a visual measurement of capacitance, which was used to determine the diameter of a bilayer seal (Maglia, Heron et al. 2010). Reducing bilayer area, and hence

capacitance, reduces noise (The Axon Guide). Furthermore, smaller bilayers are less susceptible to pressure variations and mechanical shock and are therefore longer-lived than larger bilayers. However, this must be balanced against the practical necessity of creating bilayers reproducibly (bilayers that are too small are often occluded by enlargement of the annulus) and the difficulty of inserting channels into bilayers with a small area. A good compromise was to use bilayers with a capacitance of ~60-100 pF, which corresponds to a diameter of ~105-135 μm (DPhPC/hexadecane, assuming a specific capacitance of 0.7 μF $cm^{-2}$ (Peterman, Ziebarth et al. 2002; Maglia, Heron et al. 2010).

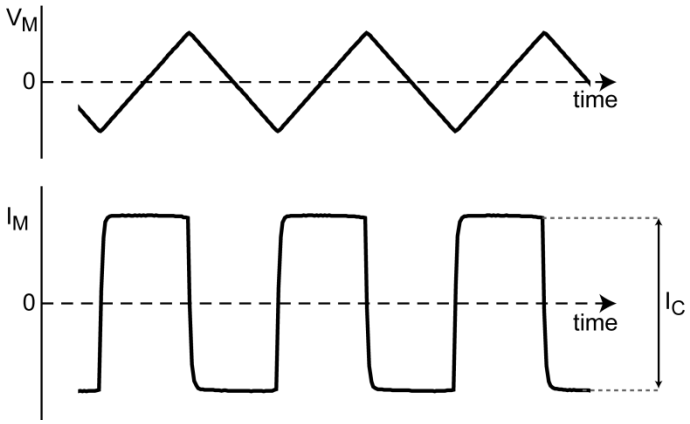



Figure 2.5. Measuring bilayer capacitance. A triangular-wave voltage input (upper) applied across a capacitor produces a square-wave current output (lower). Bilayer capacitance can be determined by measuring the peak-to-peak magnitude of the square-wave current output ($I_C$). Figure adapted from (Maglia, Heron et al. 2010).

### 2.2.6. Inserting pores and perfusion

Once a satisfactory bilayer seal was formed, αHL pore insertion was initiated by the addition of 0.2 μL (typical) of aliquoted protein (which is at ~ 1 ng $\mu L^{-1}$; see section 2.1.2) to the *cis* compartment. The solution in the *cis* compartment was then mixed by pipetting. The insertion of a protein into the

bilayer, under an applied potential, was observed as a step-wise change in the current (Figure 2.6). This current step was dependent on the applied voltage (typically +200 mV; with the *trans* positive relative to the grounded *cis*), the concentration of electrolyte (typically 1 M KCl in both *cis* and *trans* compartments) and the αHL mutant used (typical conductance of WT αHL is ~ 1 nS at +100 mV in 1 M KCl solution). Once a single pore had inserted, the *cis* compartment was perfused by manual pipetting (removing protein-containing buffer solution and replacing it with protein-free buffer solution) to reduce the protein concentration in the *cis* compartment and therefore reduce the probability of subsequent protein insertion events.

## 2.3. Data acquisition and analysis

### 2.3.1. Data digitization

The ionic current flowing through single αHL nanopores was measured using the commercial Axopatch 200B patch-clamp amplifier (Molecular Devices). This instrument transformed the continuous analog current signal that arose into an analog voltage signal, which was fed to the 1440A 16-bit digitizer (Molecular Devices) that converted it into binary numbers with finite resolution (digitization). The digitizer was connected to a computer that converted and displayed the digitized current information. The dynamic input range of the digitizer is ±10 V, which divides the signal into $2^{16}$ (65536) values (bins), providing a resolution of 0.305 mV per bin (20 V/$2^{16}$). If no output gain is applied by the amplifier, this corresponds to a dynamic range of ±10 nA and a resolution of 0.305 pA per bin.

However, the output gain of the amplifier can be changed (by up to 500X) to increase the resolution (decreased bin-width), at the expense of dynamic range. Typically, the output gain was set to 10X. This corresponds to a dynamic range of ±1 nA and a resolution of 0.0305 pA per bin (Maglia, Heron et al. 2010).

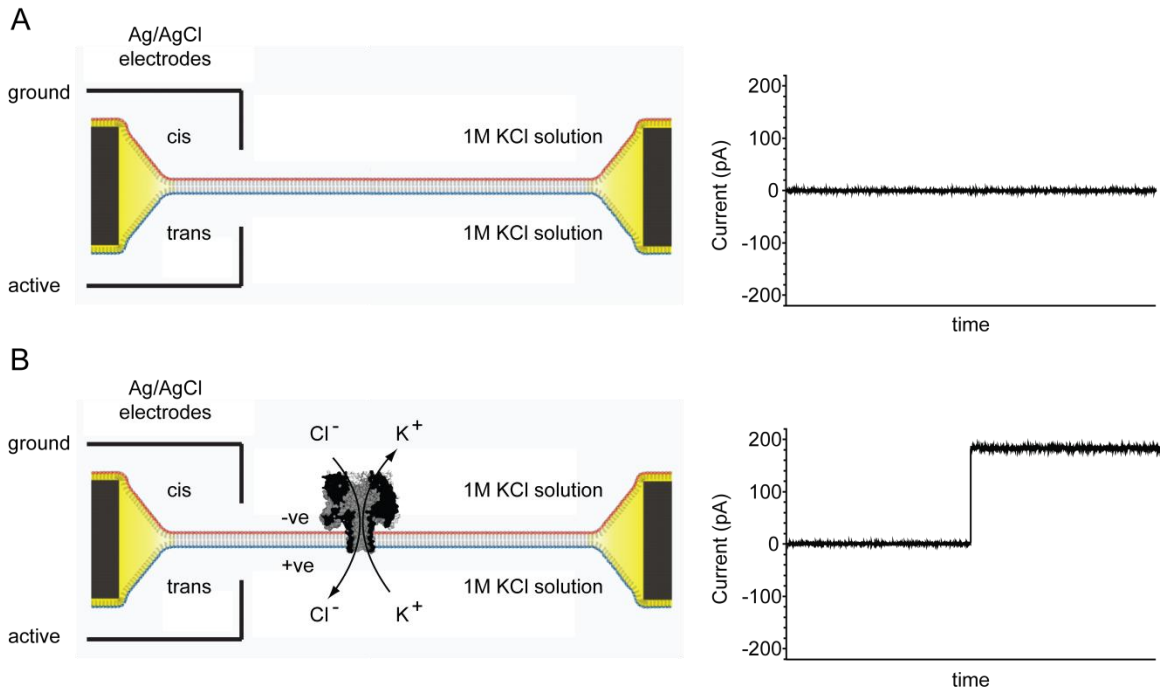



Figure 2.6. Current measurements upon bilayer formation and pore insertion. A) The formation of a planar lipid bilayer is characterized by a current measurement of zero pA, even under a positive applied potential, between the *cis* and *trans* compartments, due to the electrical isolation of the two compartments. B) Upon αHL pore (grey cut-through) insertion, the two compartments become electrically connected and an ionic current can flow across the bilayer between the two electrodes. This event is characterized by a step-wise change in the current measurement.

### 2.3.2. Filtering and sampling

High frequency electrical noise was removed from the signal collected by the amplifier, using an in-built low-pass, 4-pole Bessel filter which limits the bandwidth of the data by eliminating signals above the corner frequency (or cut-off frequency) of the filter. Bessel filters are advantageous over other filters (e.g.

a Butterworth filter), as they preserve the shape of the single channel response well (The Axon Guide) (Silberberg and Magleby 1993). In order to keep errors in estimates of current levels to <3%, the applied filter cut-off should be greater than five times the inverse of the mean event time (Silberberg and Magleby 1993). Over filtering of the signal can lead to inaccurate estimates of $I_B$. The Nyquist Sampling Theorem states that the rate at which the filtered signal is sampled (in this case by a 1440A 16-bit digitizer from Molecular Devices) should be twice the cut-off frequency of the applied filter (The Axon Guide). In practice however, it is useful to oversample (i.e. sample at a rate that exceeds the Nyquist minimum) (Sakmann and Neher 1995). A sampling rate of ten times that of the cut-off frequency of the filter is desirable, but five times the cut-off frequency is usually sufficient (Maglia, Heron et al. 2010).

### 2.3.3. Acquisition protocols

The software used to collect and display the current signals was Clampex (Molecular Devices). Clampex can also control the 1440A digitizer to apply a potential through the amplifier (see below). Although Clampex has five data acquisition modes, only the “gap-free” and “episodic stimulation" modes were used.

In the gap-free protocol, data are passively and continuously digitized, displayed and saved. This mode was used while waiting for protein pores to insert.

The episodic stimulation protocol was used to output a voltage signal from the digitizer, and apply it to the bilayer system, while simultaneously acquiring the

resulting current signal. This mode was used to apply voltage steps (e.g. when producing I-V curves) and voltage ramps (Dudko, Mathe et al. 2007). It was also used when studying the repeated immobilization of DNA-btn•streptavidin complexes inside the αHL pore. In this case, a voltage bias was applied, for a fixed amount of time, to drive a single DNA-btn•streptavidin complex into the pore. The bias was then reversed to eject the complex from the pore (see below) (Purnell, Mehta et al. 2008; Purnell and Schmidt 2009; Stoddart, Heron et al. 2009; Stoddart, Heron et al. 2010; Stoddart, Maglia et al. 2010). The changes in the current level that occurred as a result of the voltage protocol were simultaneously acquired. The protocol could be repeated multiple times per run to allow many DNA blockade events to be recorded (Maglia, Heron et al. 2010).

### 2.3.4. Analysis of DNA-btn•streptavidin molecules

ssDNA molecules, with a biotinyl group covalently attached to the 3' end through a triethyleneglycol (TEG) linker, were obtained from Sigma-Aldrich (UK) and ATDBio Ltd (UK) (Figure 2.7). Solutions of the biotinylated ssDNAs, at 100 μM in 10 mM Tris.HCl, pH 8.0, 0.1 mM EDTA, were mixed with equal volumes of 25 μM streptavidin (SA) (Sigma-Aldrich) in the same buffer (see appendix section A.4 for sequences of all biotin modified oligonucleotides).

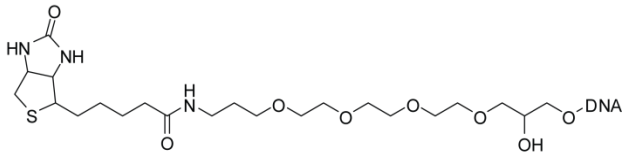



Figure 2.7. The chemical structure of the biotin-TEG linker used to biotinylate the 3' terminus of the DNA oligonucleotides. The structure was produced with ChemBioDraw Ultra 11. The length of the linker, from biotin to 3' phosphate of the first DNA base is ~3 nm, as measured using Pymol.

Each oligonucleotide (pre-incubated with streptavidin for at least five minutes) was added to the *cis* compartment to a final concentration of typically 200 nM. In a typical experiment, initially, +160 mV was applied to the *trans* side for 900 ms to drive the negatively charged, DNA-btn•streptavidin complex into the pore. The capture and immobilization of ssDNA strands by an αHL pore is observed as a stepwise decrease in the open pore current level ($I_O$) to a lower, but stable, current level ($I_B$) (Figure 2.8). A voltage of -140 mV was then applied for 50 ms to eject the immobilized DNA from the pore. The applied potential was then stepped to 0 mV for 50 ms. This one-second sequence was repeated typically for ≥100 cycles for each ssDNA species added. The amplified signal (arising from the ionic current passing through the pore) was low-pass filtered at 1 kHz and sampled at 5 kHz.

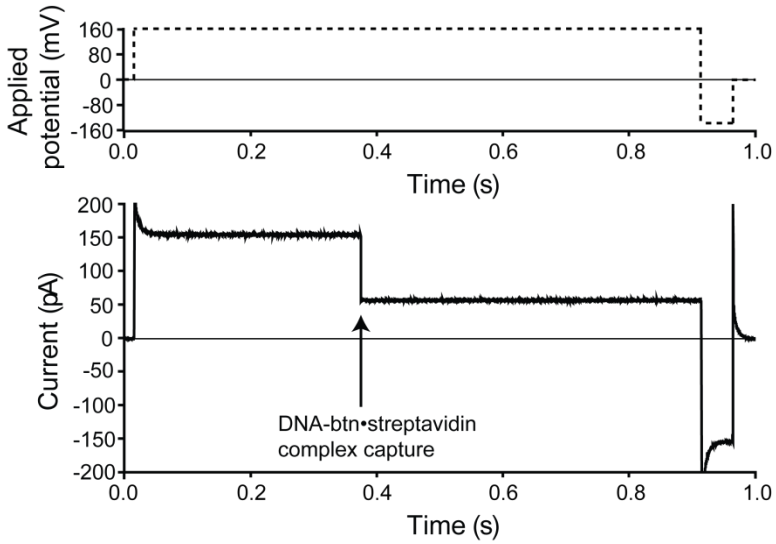



Figure 2.8. Episodic stimulation mode allows both an output voltage (dashed line - upper) to be applied, while simultaneously recording the resulting changes in ionic current (lower - black trace). A positive potential bias of +160 mV is applied to the system, resulting in an open pore current level of ~160 pA. Under the applied potential, a DNA molecule is driven into the pore and becomes immobilized (through a terminal biotin•streptavidin complex), and the current level is reduced. The potential bias is reversed (-140 mV) and the DNA molecule is ejected. The bias is then removed. The protocol can be repeated numerous times. Figure adapted from (Maglia, Heron et al. 2010).

Data were analyzed and prepared for presentation with pClamp software (version 10.1, Molecular Devices). Single-channel searches were performed to obtain the current level for each individual ssDNA blockade event ($I_B$) (Figure 2.9); firstly, pre-defined threshold levels were set for the open pore and blocked pore states, e.g. Level 1 was set to an estimation of the open pore current value and Level 2 was set to an estimation of the blocked pore current level. The event search software detects as the current changes from a Level 1 event (open pore event) to a Level 2 event (blocked pore event) and determines the arithmetic mean of each event from all of the current samples acquired, belonging to a given event (Figure 2.9). The $I_B$ values from repeated blockades, for a given DNA sequence have approximately a Gaussian distribution (Figure 2.10A). The mean

$I_B$ value for each oligonucleotide sequence was determined by performing a Gaussian fit to a histogram of all the $I_B$ values from the repeated blockades. The current blockade for each oligonucleotide sequence was also described as the residual current ($I_{RES}$), which is expressed as a percentage of the open pore current ($I_O$): $I_{RES} = (I_B/I_O)$ X 100 (Figure 2.10B).

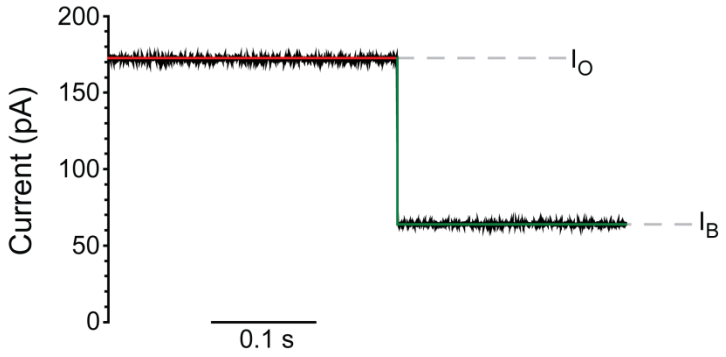



Figure 2.9. Typical DNA-btn•streptavidin blockade event analyzed by Clampfit software. The current level changes from an open pore event, to a blocked pore event. The software assigns the events and determines the arithmetic mean for both the $I_O$ (green line) and the $I_B$ (red line) events.

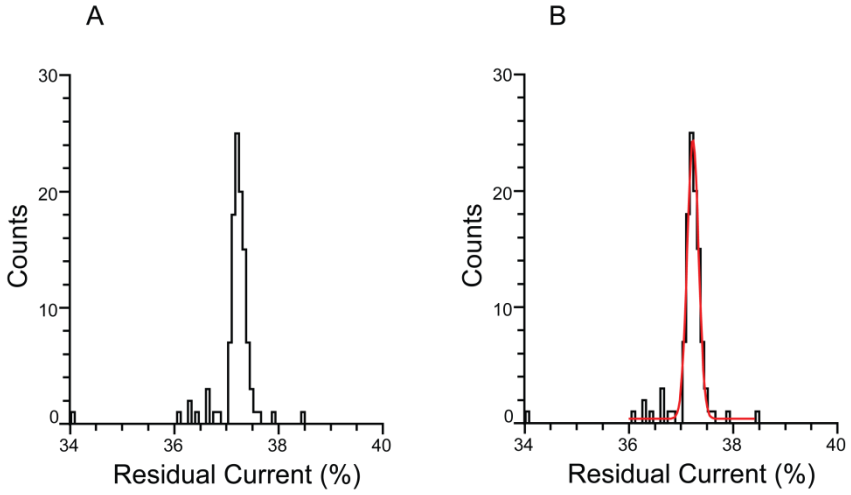



Figure 2.10. Residual current histograms to determine the average blockade level for an oligonucleotide sequence. A) Histogram displaying residual current blockades for a single oligonucleotide sequence, for a typical experiment. B) The average blockade for the oligonucleotide sequence, for this experiment, is taken from the centre of the peak of the fitted Gaussian.

When comparing several oligonucleotide sequences, a single oligonucleotide sequence was first added to the *cis* chamber and the current trace required for the determination of $I_B$ and $I_{RES}$ was recorded. Subsequently, a second (and if required, a third and a fourth) oligonucleotide sequence was added and additional currents recorded. When such experiments were repeated, the oligonucleotides were added to the chamber in a different order (Stoddart, Heron et al. 2009; Stoddart, Heron et al. 2010; Stoddart, Maglia et al. 2010). To establish if the current blockades that were produced by two different oligonucleotides were sequence dependent (i.e. to determine if it was possible to discriminate between the two sequences based on single-channel current measurements), the difference between the $I_{RES}$ levels (taken from the histogram) was determined, with the poly(dC) oligonucleotide typically acting as the reference point. For example, consider a comparison between poly(dA) and poly(dC) oligonucleotides. If $I_{RES}^{poly(dA)} - I_{RES}^{poly(dC)}$ ($\Delta I_{RES}^{poly(dA) - poly(dC)}$) = 0, then it is stated that the oligonucleotides are not discriminated, for a given mutant when measured under given conditions. However, if $\Delta I_{RES}^{poly(dA) - poly(dC)} \neq 0$, then it is stated that the oligonucleotides are discriminated. It should be noted that the $\Delta I_{RES}$ values showed little experimental variation, while the absolute current values often showed variation that exceeded $\Delta I_{RES}$. In practice, the $\Delta I_{RES}$ values were readily determined from event histograms (Figure 2.11) (Stoddart, Heron et al. 2009).

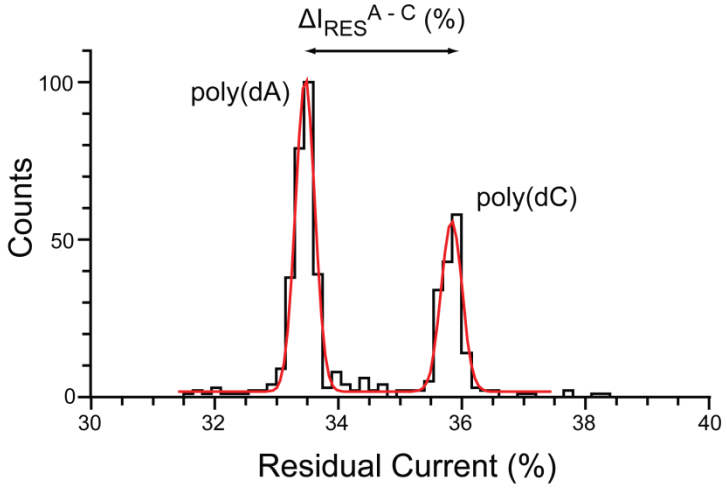



Figure 2.11. Histogram displaying how $\Delta I_{RES}$ values were determined. The $I_{RES}$ value of each oligonucleotide sequence, in this case, poly(dA) and poly(dC) were taken from Gaussian fits to the blockade events, and the $\Delta I_{RES}$ value is calculated as the difference between them: $\Delta I_{RES}^{A-C} = I_{RES}^{A} - I_{RES}^{C}$ (Stoddart, Heron et al. 2009).

To determine how well a particular mutant could discriminate between single nucleobase changes (e.g. each of the four bases; C, T, A and G), two extra parameters were used; the overall dispersion was measured (the difference between the most widely dispersed residual current levels in the histogram: $\Delta I_{RES}^{OVERALL}$, Figure 2.12). Also, the product of the sequential differences between each of the residual current levels (one level for each oligonucleotide) was calculated (δ, Figure 2.12). An αHL pore that is unable to discriminate between all nucleobases (i.e. the current levels from two or more oligonucleotides overlap) has $\delta = 0$. However, pores that are able to discriminate between all bases have $\delta > 0$ (Stoddart, Heron et al. 2010).

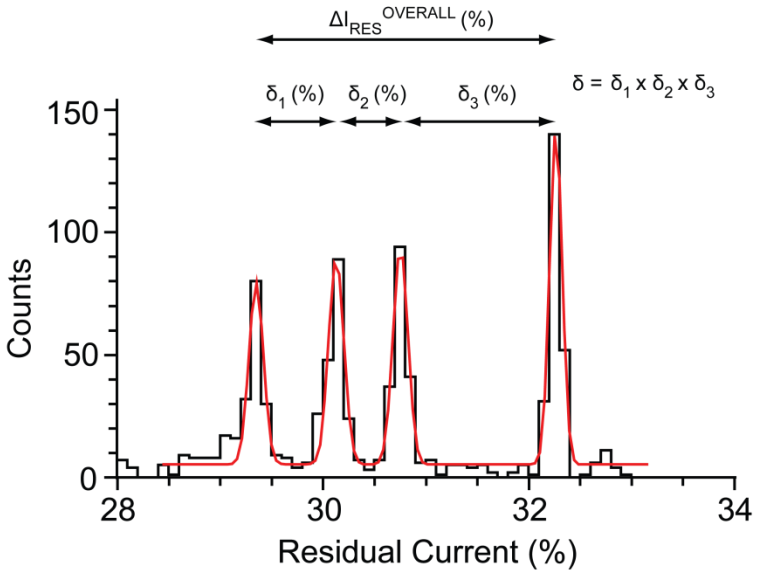



Figure 2.12. Parameters used to measure base identification properties of αHL protein pores. Histogram showing the $I_{RES}$ levels observed when an E111N/K147N/M113Y pore is probed with four poly(dC) oligonucleotides that contain a single substituted nucleotide at position 9 (relative to the 3' biotin-tag) (Stoddart, Maglia et al. 2010). The parameters used to determine the discrimination of single nucleobase substitutions are shown. $\Delta I_{RES}^{OVERALL}$ is the difference in residual current between the two most widely dispersed current peaks. The δ value is the product of the successive differences between the peaks. If any two peaks overlap, then the δ value is equal to zero. Figure adapted from (Stoddart, Heron et al. 2010).

References

Cheley, S., O. Braha, et al. (1999). "A functional protein pore with a "retro" transmembrane domain." Protein Sci. **8**: 1257-1267.

Cheley, S., M. S. Malghani, et al. (1997). "Spontaneous oligomerization of a staphylococcal a-hemolysin conformationally constrained by removal of residues that form the transmembrane b barrel." Protein Eng. **10**: 1433-1443.

Dudko, O. K., J. Mathe, et al. (2007). "Extracting Kinetics from Single-Molecule Force Spectroscopy: Nanopore Unzipping of DNA Hairpins." Biophys J **92**(12): 4188-4195.

Howorka, S. and H. Bayley (1998). "Improved protocol for high-throughput cysteine scanning mutagenesis." Biotechniques **25**(5): 764-766, 768, 770 passim.

Janz, G. J. (1961). Silver-Silver Halide Electrodes. Reference Electrodes, Theory and Practice. D. G. J. a. J. Ives, G. J**:** 179-230.

Maglia, G., A. J. Heron, et al. (2010). "Analysis of single nucleic acid molecules with protein nanopores." Methods Enzymol **475**: 591-623.

Mayer, M., J. K. Kriebel, et al. (2003). "Microfabricated teflon membranes for low-noise recordings of ion channels in planar lipid bilayers." Biophys.J. **85**: 2684-2695.

Montal, M. and P. Mueller (1972). "Formation of bimolecular membranes from lipid monolayers and study of their electrical properties." Proc.Natl.Acad.Sci.USA **69**: 3561-3566.

Peterman, M. C., J. M. Ziebarth, et al. (2002). "Ion channels and lipid bilayer membranes under high potentials using microfabricated apertures." Biomed.Microdevices **4**: 231-236.

Purnell, R. F., K. K. Mehta, et al. (2008). "Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore." Nano Lett **8**(9): 3029-3034.

Purnell, R. F. and J. J. Schmidt (2009). "Discrimination of single base substitutions in a DNA strand immobilized in a biological nanopore." ACS Nano **3**: 2533-2538.

Purves, R. D. (1981). Microelectrode methods for intracellular recording and ionophoresis, Academic Press.

Sakmann, B. and B. Neher (1995). Single-channel recording. New York, Plenum.

Schagger, H. and G. von Jagow (1991). "Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form." Anal Biochem **199**(2): 223-231.

Silberberg, S. D. and K. L. Magleby (1993). "Preventing Errors When Estimating Single Channel Properties from the Analysis of Current Fluctuations." Biophys. J. **65**: 1570-1584.

Stoddart, D., A. Heron, et al. (2009). "Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore." Proc. Natl. Acad. Sci. USA **106**: 7702-7707.

Stoddart, D., A. J. Heron, et al. (2010). "Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore." Nano Lett **10**(9): 3633-3637.

Stoddart, D., G. Maglia, et al. (2010). "Multiple base-recognition sites in a biological nanopore– two heads are better than one." Angew Chem Int Ed **49**: 556-559.

Wonderlin, W. F., A. Finkel, et al. (1990). "Optimizing planar lipid bilayer single-channel recordings for high resolution with rapid voltage steps." Biophys.J. **58**: 289-294.

# CHAPTER 3: NUCLEOBASE DISCRIMINATION IN IMMOBILIZED OLIGONUCLEOTIDES WITH THE α-HEMOLYSIN PROTEIN NANOPORE

## Declaration

The work in this chapter is also described in the manuscript:

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. and Bayley, H. (2009). "Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore." Proc. Natl. Acad. Sci. USA **106**: 7702-7707.

Author contributions: D.S., A.H., G.M. and H.B. designed the concept; D.S. performed all single-channel recordings and E.M. performed the mutant and protein preparations; D.S. analyzed data; and D.S. and H.B. wrote the paper.

## Abstract

It is possible to increase the residence time of oligonucleotides within the α-hemolysin (αHL) protein nanopore by immobilization through a 3′ terminal biotin•streptavidin plug. This allows for more precise measurement of the current signal which in turn facilitates the observation of small current differences between oligonucleotides with different sequences. The discrimination of oligonucleotides can be altered and improved by genetic engineering of the αHL protein. It is shown in this chapter that the transmembrane β barrel of the pore contains three recognition sites and that by using the αHL E111N/K147N (NN) mutant, it is possible to identify all four DNA bases in an immobilized single-

stranded DNA molecule, whether they are located in an otherwise homopolymeric DNA strand or in a heteropolymeric strand.

## 3.1. Introduction

One of the fundamental challenges that must be met in order to develop a functional nanopore DNA sequencing device is the capability to identify single bases, based on current measurements. With this mind, several advances suggest that nanopore sequencing with protein pores is feasible. Initially it was observed that it was possible to discriminate between the current signatures of homopolymeric oligonucleotides, translocating through the α-hemolysin (αHL) nanopore. Working with RNA, Akeson et al. showed that poly(rA) blockades produced a residual current level ($I_{RES}$) of 15% with a translocation rate of 22 μs $nt^{-1}$. However blockades from poly(rC) sequences produced $I_{RES}$ values of 5% and 9%, with a translocation rate of 5.8 μs $nt^{-1}$ meaning that the current signatures of poly(rA) and poly(rC) blockades could be readily distinguished (Akeson, Branton et al. 1999). Subsequently, Meller et al. showed that discrimination of homopolymeric DNA strands was also possible, using similar parameters. The sequences tested were poly(dA) and poly(dC) sequences, and as well as the oligonucleotides displaying different transit times (the translocation rates were measured as 1.9 and 0.8 μs $nt^{-1}$, respectively), the levels of current block, also differed with poly(dA) blockages producing an $I_{RES}$ of 12.6% and poly(dC) blockages producing an $I_{RES}$ of 13.4% (giving a $\Delta I_{RES}^{poly(dA)\ -\ poly(dC)}$ = -0.8% Table 3.1) (Meller, Nivon et al. 2000). Despite this early advance, the

observation that DNA traverses the nanopore at rates approaching 1 μs $nt^{-1}$ suggested that it would be very difficult to identify single nucleobases as a ssDNA molecule translocated through the αHL pore. During a DNA blockade event, the ionic current flowing through the pore is reduced to ~10 pA. Given that there are 6.2 x $10^{18}$ elementary charges in 1 Coulomb, then 10 pA corresponds to ~62 million translocating ions per second (I = Q/t). It follows that only 62 ions are available to establish the identity of a base (given a transit time 1 μs $nt^{-1}$). In reality, it would be even fewer ions than this, since the difference between translocating homopurines and homopyrimidines is only ~2 pA. Therefore, given the limitations of the best patch-clamp amplifiers, the small number of ions available cannot provide a sufficient signal-to-noise ratio to discriminate between single bases in freely translocating strands (i.e. on the microsecond timescale) (Deamer 2010), unless they have been modified with bulky groups (Mitchell and Howorka 2008). Indeed, the best that has been achieved in the "sequencing" of freely translocating polynucleotides is the ability to detect the transition between long homopolymeric stretches in single RNA molecules which have the sequence poly$(rA)_{30}(rC)_{70}$ (Akeson, Branton et al. 1999).

In a step towards the management of this problem, first the Ghadiri group and more recently the Akeson group, have shown that ssDNA can be ratcheted one base at a time through the αHL pore by the action of a DNA polymerase (Cockroft, Chu et al. 2008; Olasagasti, Lieberman et al. 2010). The idea is that the polymerase will control the speed at which DNA is fed or pulled through the nanopore, under an applied potential, and will ultimately operate at a speed of ~100 bases per second. This translates to a translocation rate of ~10 ms $nt^{-1}$,

which is several orders of magnitude slower than freely translocating DNA and should be sufficient to improve the signal-to-noise ratio of the current measurements and allow for the small current changes associated with single base transitions to be detected (Clarke, Wu et al. 2009).

Table 3.1. Summary of residual current levels ($I_{RES}$) for αHL pores probed with poly(dC) and poly(dA) oligonucleotides. $\Delta I_{RES}$ is defined as the difference in residual current between the poly(dA) and poly(dC) blockades ($I_{RES}^{poly(dA)}$ - $I_{RES}^{poly(dC)}$). Table adapted from (Stoddart, Heron et al. 2009).

| Experimental approach | End of DNA threaded | $I_{RES}^{poly(dA)}$ (%) | $I_{RES}^{poly(dC)}$ (%) | $\Delta I_{RES}$ (%) | Conditions (all experiments conducted between 19-25 °C) | Ref |
|---|---|---|---|---|---|---|
| btn•strep | 5' | 20.0 | 19.4 | +0.6 | 1 M KCl, 25 mM Tris.HCl, pH 8.0, with 0.1 mM EDTA / +160 mV | Current work |
| Terminal hairpin | 5' | 21.5 | 32.0 | -10.5 | 500 mM KCl, 5 mM MOPS, pH 7.5 / +170 mV | (1) |
| Terminal hairpin | 3' | 22.0 | 31.0 | -9.0 | 500 mM KCl, 5 mM MOPS, pH 7.5 / +170 mV | (1) |
| Freely translocating | unknown | 12.6 | 13.4 | -0.8 | 1 M KCl, 10 mM Tris.HCl, pH 8.5 / +120 mV | (2) |
| btn•strep | 5' | 18.3 | 17.1 | +1.2 | 1 M KCl, 10 mM Tris.HCl, pH 8, with 1 mM EDTA / +120 mV | (3) |
| btn•strep | 3' | 16.0 | 18.9 | -2.9 | 1 M KCl, 10 mM Tris.HCl, pH 8, with 1 mM EDTA / +120 mV | (3) |


1. Ashkenasy, N., Sánchez-Quesada, J., Bayley, H., & Ghadiri, M. R. (2005) Recognizing a single base in an individual DNA strand: A step toward nanopore DNA sequencing. *Angew.Chem.Int.Ed.Engl.* 44, 1401-1404.

2. Meller, A., Nivon, L., Brandin, E., Golovchenko, J., & Branton, D. (2000) Rapid nanopore discrimination between single polynucleotide molecules. *Proc.Natl.Acad.Sci.USA* 97, 1079-1084.

3. Purnell, R. F., Mehta, K. K., & Schmidt, J. J. (2008) Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Lett* 8, 3029-3034.


However, as yet no working system has been perfected and in a stopgap approach, base identification can be studied by immobilizing oligonucleotides

within the pore through a terminal biotin•streptavidin plug which prevents complete translocation of the DNA strand into the *trans* compartment (Figure 3.1) (Kasianowicz, Henrickson et al. 2001; Sánchez-Quesada, Saghatelian et al. 2004; Purnell, Mehta et al. 2008). This approach is designed to mimic the pauses between enzymatic base additions, during which the DNA strand is held, stretched under an applied potential, within the pore lumen. Also it is advantageous as it increases the residence time of the oligonucleotide within the pore. Therefore it is possible to reduce the noise of the current signal by stringent filtration, thus improving the signal-to-noise ratio of the current signal and the precision of the measurements. The following section details the investigations that were made, using the biotin•streptavidin approach, to map the base recognition properties of the αHL pore, and determine whether or not the discrimination of single nucleobases was possible.

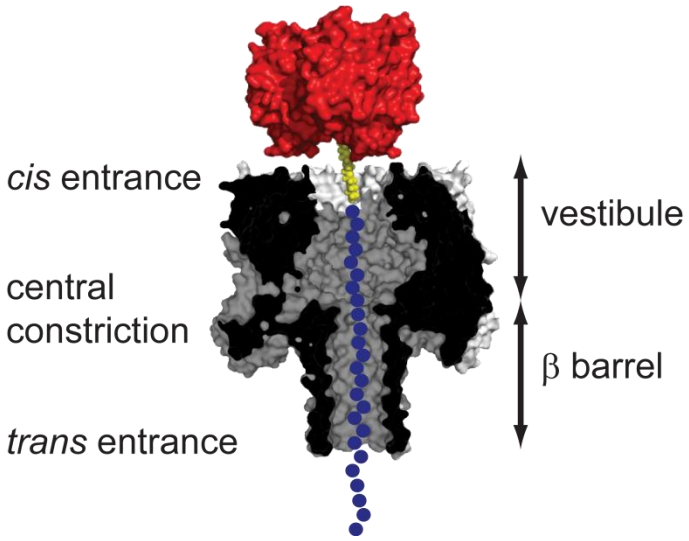



Figure 3.1. Immobilization of an ssDNA oligonucleotide within an αHL pore. Schematic representation of a ssDNA oligonucleotide (blue circles) immobilized inside an αHL pore (grey, cross-section) through the use of a biotin (yellow)-streptavidin (red) linkage. (Song, Hobaugh et al. 1996).

## 3.2. Results and discussion

### 3.2.1. Discrimination of homopolymeric oligonucleotides with the α-hemolysin (αHL) pore

Homopolymeric ssDNA oligonucleotides (poly(dA)$_{60}$ and poly(dC)$_{60}$) with biotin tags at the 3' terminus were allowed to complex with streptavidin (see Chapter 2, section 2.2.7.4) and subsequently added to the *cis* compartment. Using the episodic stimulation mode in the pClamp software (version 10.1, Molecular Devices), the Digidata 1440A digitizer (Molecular Devices) was used to apply repeated voltage steps to drive single DNA-btn•streptavidin complexes into the pore (positive potential in *trans* relative to *cis*) and subsequently eject them (negative potential in *trans* relative to *cis*). The immobilization of a DNA molecule resulted in a step-wise decrease in the current flow through the pore (Figure 2.8). The current flowing through the blocked pore, $I_B$, is also given as the residual current ($I_{RES}$) which is expressed as a percentage of the open pore current ($I_O$) (see Chapter 2, section 2.2.7.4). Initially, the wild-type (WT) αHL pore (protein prepared by E. Mikhailova) was examined. At +160 mV in 1 M KCl, 25 mM Tris.HCl, pH 8.0, containing 0.1 mM EDTA (the conditions for all SCR experiments reported, unless otherwise stated), WT αHL pores have a mean open pore current level ($I_O^{WT}$) of 171 ± 8 pA (n = 8 – Table 3.2). In each experiment, poly(dA)$_{60}$ oligonucleotides blocked WT pores to a lesser extent ($I_{RES}^{poly(dA)}$ = 20.0 ± 1.3%) than poly(dC)$_{60}$ ($I_{RES}^{poly(dC)}$ = 19.4 ± 1.4%) (Figure 3.2A). The residual current difference between the poly(dA) and the poly(dC) oligonucleotide blockades ($\Delta I_{RES} = I_{RES}^{poly(dA)} - I_{RES}^{poly(dC)}$) was +0.6 ± 0.1% (Table

3.2). In other words, the blockades caused by immobilization of oligonucleotides within the αHL are sequence dependent. It should be noted that the $\Delta I_{RES}$ values showed little experimental fluctuation, while the open pore and blocked pore current values showed variation that exceeded $\Delta I_{RES}$. The level of variation in the open and blocked pore current values that are observed between different pores and different experiments are similar to previous observations made by the Schmidt group (Purnell, Mehta et al. 2008). It has been proposed that such inter-experiment variation is due to differences in membrane resistance (Purnell, Mehta et al. 2008). Alternative sources of fluctuation may also include slight differences in temperature or solution conductivity between experiments. However, it is speculated that the largest source of variation is a result of differences in conformation between pores; although each pore is similar in overall structure and conformation, each may exhibit subtle conformational differences which produce changes in pore conductivity and result in the variations in the open pore and the blocked pore values that are observed (Table 3.2). Despite this, in each experiment the homopolymers were readily distinguishable and in practice, the $\Delta I_{RES}$ values were readily determined from event histograms (Figure 2.11). DNA homopolymers have been distinguished with the WT αHL pore by several other groups on the basis of differences in $I_{RES}$ (summarized in Table 3.1). When ssDNA was immobilized in the pore with a terminal hairpin, Ashkenasy found a $\Delta I_{RES}^{poly(dA) - poly(dC)}$ value of -10.5%, or -9.0% depending on whether the 5' end or 3' end was threaded (Ashkenasy, Sánchez-Quesada et al. 2005). On the other hand, Purnell et al., employing the DNA-btn•streptavidin approach, found that $\Delta I_{RES}$ depends on the orientation of the

threaded DNA (i.e. 5' or 3'): 5' entry, $\Delta I_{RES}^{poly(dA) - poly(dC)}$ = +1.2% and 3' entry, $\Delta I_{RES}^{poly(dA) - poly(dC)}$ = -2.9% (Purnell, Mehta et al. 2008). The dependence of the observed current blockades, and the $\Delta I_{RES}$ values, on the orientation of ssDNA within the αHL pore has also been observed with freely translocating DNA molecules (Mathé, Aksimentiev et al. 2005). Molecular dynamics simulations have suggested that the basis of this dependence is caused by the tilting of the bases towards the 5' terminus of the backbone, when DNA is confined within the channel. This base tilting would result in similar sequences adopting different conformations within the pore, depending on their conformation - 5' entry vs 3' entry, and therefore causing different current blockades.

The results described here (5' entry: $\Delta I_{RES}^{poly(dA) - poly(dC)}$ = +0.6 ± 0.1%) are in rough agreement with the work of Purnell et al., given that their results were obtained at a lower applied potential (+120 mV), and that $\Delta I_{RES}$ is voltage-dependent (Figure 3.3 – there is a decrease in $\Delta I_{RES}$ at +160 mV compared to at +120 mV). In general, $I_{RES}$ values are greater when the DNAs are immobilized within the pore, compared to freely translocating DNA (Table 3.1 and references therein). Presumably DNA is more stretched in the electric field within the pore when it is immobilized resulting in a reduced occlusion of the pore, allowing a larger current flow.

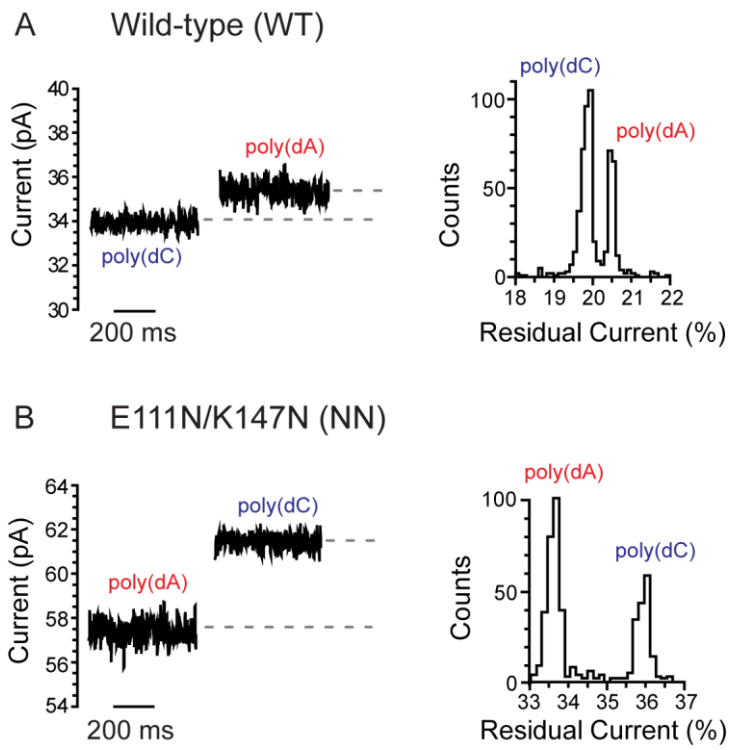
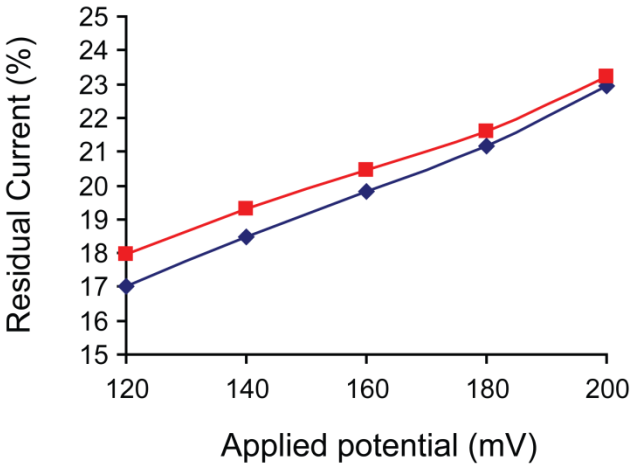



Figure 3.2. Blockades by streptavidin bound oligonucleotides are sequence dependent. Typical current levels (left) and event histograms displaying the residual current levels (right) for the WT (panel A) and E111N/K147N (NN) (panel B) pore when blocked with immobilized poly(dC) and poly(dA) oligonucleotides. Figure adapted from (Stoddart, Heron et al. 2009).

Table 3.2. Residual currents ($I_{RES}$) for poly(dC) and poly(dA) oligonucleotides immobilized within WT and NN pores. The $I_O$ and $I_{RES}$ values given for each oligonucleotide are mean values taken from Gaussian fits to event histograms for individual experiments. $\Delta I_{RES}^{A-C}$ is the difference in residual current between the poly(dA) and poly(dC) blockades ($I_{RES}^{poly(dA)}$ - $I_{RES}^{poly(dC)}$) and $\Delta I_{RES}^{OVERALL}$ is the absolute difference in residual current levels between the two current peaks (Figure 2.12). Table adapted from (Stoddart, Heron et al. 2009).

| Expt | WT | | | | | Expt | NN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $I_O$ (pA) | $I_{RES}^C$ (%) | $I_{RES}^A$ (%) | $\Delta I_{RES}$ (%) | | | $I_O$ (pA) | $I_{RES}^C$ (%) | $I_{RES}^A$ (%) | $\Delta I_{RES}$ (%) | |
| | | | | A–C | Overall | | | | | A–C | Overall |
| 1 | 163 | 17.1 | 17.8 | +0.7 | 0.7 | 1 | 172 | 37.1 | 34.1 | -3.0 | 3.0 |
| 2 | 172 | 17.6 | 18.3 | +0.7 | 0.7 | 2 | 162 | 36.4 | 33.9 | -2.5 | 2.5 |
| 3 | 187 | 21.4 | 21.9 | +0.5 | 0.5 | 3 | 166 | 35.5 | 32.6 | -2.9 | 2.9 |
| 4 | 170 | 19.9 | 20.5 | +0.7 | 0.7 | 4 | 176 | 37.1 | 34.1 | -3.0 | 3.0 |
| 5 | 169 | 19.8 | 20.5 | +0.6 | 0.6 | 5 | 167 | 36.9 | 33.8 | -3.1 | 3.1 |
| 6 | 163 | 20.2 | 20.8 | +0.6 | 0.6 | 6 | 165 | 36.9 | 33.9 | -3.0 | 3.0 |
| 7 | 172 | 19.6 | 20.0 | +0.4 | 0.4 | 7 | 192 | 37.2 | 35.3 | -1.9 | 1.9 |
| 8 | 173 | 19.9 | 20.5 | +0.6 | 0.6 | 8 | 172 | 35.8 | 33.5 | -2.3 | 2.3 |
| Mean | 171 | 19.4 | 20.0 | +0.6 | 0.6 | Mean | 172 | 36.6 | 33.9 | -2.7 | 2.7 |
| SD | 8 | 1.4 | 1.3 | 0.1 | 0.1 | SD | 9 | 0.6 | 0.7 | 0.4 | 0.4 |




| Voltage (mV) | $I_O$ (pA) | $I_{RES}^{poly(dC)}$ (%) | $I_{RES}^{poly(dA)}$ (%) | $\Delta I_{RES}^{A-C}$ (%) |
|---|---|---|---|---|
| 200 | 216 | 22.9 | 23.2 | +0.3 |
| 180 | 194 | 21.2 | 21.6 | +0.4 |
| 160 | 169 | 19.8 | 20.5 | +0.7 |
| 140 | 148 | 18.5 | 19.3 | +0.8 |
| 120 | 125 | 17.0 | 18.0 | +1.0 |

Figure 3.3. Voltage dependence of $I_{RES}$ for WT pores threaded with either poly(dA) (red line) or poly(dC) (blue line). The data for the graph was obtained by taking mean values from Gaussian fits to histograms of residual current levels for multiple blockades for each oligonucleotide, at various applied potentials. Figure adapted from (Stoddart, Heron et al. 2009).

### 3.2.2. Improved discrimination with an αHL mutant protein

The E111N/K147N (NN) mutant (mutant and protein prepared by E. Mikhailova) forms stable pores despite the removal of the electrostatic interactions between Glu-111 and Lys-147 residues at the central constriction (Gu, Dalla Serra et al. 2000; Gu, Cheley et al. 2001). Loosening the central constriction in this way (Figure 3.4) was expected to increase space at the constriction and hopefully result in more current to flow in the presence of DNA which may produce a greater dispersion of $I_{RES}$ values (i.e. provide an increase in $\Delta I_{RES}^{OVERALL}$).

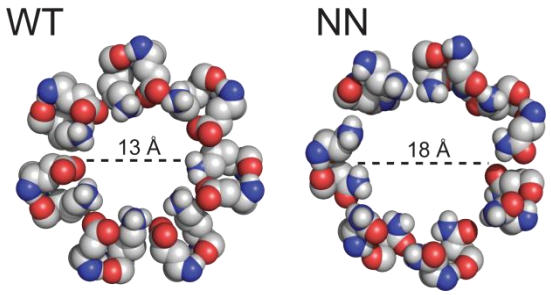



Figure 3.4. Mutation of the central constriction of the αHL pore. A view of the WT residues Glu-111 and Lys-147 (left - atoms shown as spheres) from the *cis* side of the pore. These residues form a constriction with diameter of ~13 Å. The Glu-111 and Lys-147 residues were changed to Asn with MODELLER, by J. Klingelhoefer (Stoddart, Heron et al. 2010). The atoms in the Asn-111 and Asn-147 residues are viewed as spheres, from the *cis* side (right). The mutated constriction has a wider diameter, compared to that of the WT pore, which is estimated to be ~18 Å. Figure adapted from (Stoddart, Heron et al. 2010).

Under similar conditions in which the WT pore was examined, pores formed from NN gave $I_O^{NN}$ = 172 ± 9 pA (n = 8 – Table 3.2) which is similar to the WT pore. However as hoped, $I_{RES}$ values were higher when oligonucleotides were immobilized within the NN pores (Figure 3.2B): $I_{RES}^{poly(dA)}$ = 33.9 ± 0.7% and $I_{RES}^{poly(dC)}$ = 36.6 ± 0.6%. Furthermore, the dispersion ($\Delta I_{RES}^{OVERALL}$) between the two $I_{RES}$ levels was also increased, providing improved levels of discrimination (Table 3.2). In WT pores the ring of charged lysine and glutamate side chains in the constriction (residues 147 and 111, respectively, Figure 3.4), which are replaced with asparagines in the mutant, might have one or more effects during a DNA blockade, including: a coulombic block to ion transport, or a steric block based either simply on the bulk of the large amino acid side chains, which might "grip" the DNA, or a collapse of the barrel around the DNA. In either case, the current, which is carried largely by hydrated $K^+$ ions while the negatively charged DNA strand is in the pore (Sánchez-Quesada, Saghatelian et al. 2004; Zhang

and Shklovskii 2007), is reduced in the WT pore (almost half that of the NN pore) and so is base discrimination ($\Delta I_{RES}^{OVERALL}$). The fact that residual currents in the presence of ssDNA were almost twice as high in the mutant pore may be the basis of why NN gave better discrimination between poly(dA) and poly(dC).

As well as an increase in the residual current and the magnitude of $\Delta I_{RES}^{OVERALL}$, there was also a change in the sign of $\Delta I_{RES}^{A-C}$, with poly(dA) blockades giving a lower $I_{RES}$ than poly(dC) oligonucleotide blockades in the NN pores: $\Delta I_{RES}^{A-C}$ = -2.7 ± 0.4% (Figure 3.2B). The current levels that are observed cannot be readily rationalized. Akeson et al. and Meller et al. suggest that the distinct current signatures they observe for homopolymeric sequences, as they freely translocate through the pore, are caused by differences in secondary structure. An alternative conclusion is that the central constriction, comprising residues Lys-147, Glu-111 and Met-113 in the WT, forms a recognition site and mutation of one, or more, of these residues has the potential to alter the base-dependent current flow. So, when the NN mutant was made, a “base recognition site” at the central constriction was altered, causing a relative change in the pattern of current blockades produced by poly(dA) and poly(dC), compared to WT.

### 3.2.3. Defining the recognition element within the αHL pore.

In the previous section, it was postulated that base identification may occur at the central constriction. Ashkenasy et al. concluded that recognition occurs at the *trans* entrance (Ashkenasy, Sánchez-Quesada et al. 2005)

although in this case the ssDNA was immobilized by 5' or 3' terminal hairpins, which probably enter the pore vestibule and bump up against the central constriction and therefore may perturb the base recognition that occurs at the central constriction. Together, these results imply that more than one recognition site might be present in the β barrel of the αHL pore. Therefore the next goal was to conclusively define the regions of the αHL pore that interact with DNA in a base-specific manner. In order to achieve this, five oligonucleotides were acquired, each of which contained a stretch of 5 consecutive adenine nucleobases ($A_5$ oligonucleotides) in an otherwise poly(dC) sequence (Figure 3.5A, the locations of the $A_5$ sequences in the figure are justified below). The poly(dC) "background" was used, as poly(dC) sequences are thought to have minimal secondary structure (Adler, Grossman et al. 1967) and poly(dC) molecules have been postulated to translocate through the αHL in a disordered, extended state (Akeson, Branton et al. 1999). A lack of secondary structure in our model system would be advantageous, as it would mean that any single-channel current differences observed between different sequences would probably be base-specific, rather than structure-specific. These so-called $A_5$ oligonucleotides were designed to probe the length of the αHL pore in a similar approach that had been previously established by Ghadiri and colleagues, a study in which poly(dC) was also used as the background (Ashkenasy, Sánchez-Quesada et al. 2005). The idea was to determine where base recognition was occurring, by comparing the adenine containing oligonucleotides with non-adenine containing oligonucleotides. If the adenine and non-adenine oligonucleotides produce blockades with the same $I_{RES}$ value (i.e. $\Delta I_{RES} = 0$), then

it can be inferred that the adenines are not discriminated by the pore. However, if the adenine and non-adenine containing oligonucleotides produce blockades with different $I_{RES}$ values (i.e. $\Delta I_{RES} \neq 0$), then it suggests that the adenines are situated within a base recognition region in the αHL pore. By moving the position of the adenines in the adenine containing oligonucleotide (as in the set shown in Figure 3.5A), it was hoped to define where these recognition regions were located.

$\Delta I_{RES}$ values for each of the $A_5$ oligonucleotides (i-v, Figure 3.5A), with respect to a reference poly(dC) oligonucleotide, were determined for both the WT and NN pores (Figure 3.5B Table 3.3). The data suggest that when the $A_5$ sequence is closest to the streptavidin anchor (positions 1-5 from the 3' end – herein, all nucleobase positions given are relative to 3' end), the bases are not recognized by the αHL pore (i.e. $\Delta I_{RES}^{\text{A5 oligo i – poly(dC)}} = 0$, for both WT αHL and NN), and the $A_5$ sequence is likely to lie within the vestibule. However, when the $A_5$ sequence was in positions 6-10 (oligonucleotide ii), 11-15 (iii), and 16-20 (iv), the bases were recognized in both pores (Figure 3.5). When the $A_5$ sequence was in positions 6-10, the WT and NN pores recognized the DNA in a different way, i.e. for WT αHL, $\Delta I_{RES}^{\text{A5 oligo ii – poly(dC)}}$ was positive (+0.4 ± 0.2%) and for NN, $\Delta I_{RES}^{\text{A5 oligo ii – poly(dC)}}$ was negative (-1.6 ± 0.1%). This implied that the $A_5$ sequence in oligonucleotide ii lies at the central constriction, where the mutations are located. The enhanced discrimination of oligonucleotides iii and iv (in the NN mutant) may be due to the increased current flow through the NN pore (compared to WT) when it is blocked with DNA. Finally, when the $A_5$ sequence was in positions 21-25 (oligonucleotide v), no discrimination was observed

suggesting that the adenine stretch in this sequence protrudes beyond the *trans* entrance of the pore. Therefore it is concluded that the sequence of 15 bases, from positions 6 through 20 (from the 3' end of the DNA) lie within the recognition region of the pore, which is likely to consist of the narrow confines of the β barrel.

The recognition element was further defined by sequentially substituting a single adenine base through a poly(dC) background and comparing the residual current with that of poly(dC) itself, in a similar manner to what had been done for the $A_5$ oligonucleotides. A set of fourteen poly(dC) oligonucleotides was acquired, each containing a single adenine ($A_1$) nucleobase (Ashkenasy, Sánchez-Quesada et al. 2005).

Table 3.3. Residual currents ($I_{RES}$) for poly(dC) oligonucleotides containing a stretch of 5 consecutive adenine nucleobases immobilized within WT and NN pores. The $I_O$ and $I_{RES}$ values shown are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is the difference in residual current between each $A_5$ oligonucleotide (i-v) (Figure 3.5A) and poly(dC) ($I_{RES}^{A5\ oligo} - I_{RES}^{poly(dC)}$). The errors given are standard deviations. Table adapted from (Stoddart, Heron et al. 2009).

| Oligo i-v | WT | | | | Oligo i-v | NN | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $I_O$ (pA) | $I_{RES}^{i-v}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_O$ (pA) | $I_{RES}^{i-v}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| I | 168 ± 2 | 19.3 ± 0.7 | 19.3 ± 0.7 | 0.0 ± 0.0 | i | 158 ± 1 | 35.8 ± 1.8 | 35.8 ± 1.8 | 0.0 ± 0.0 |
| Ii | 171 ± 4 | 19.7 ± 0.8 | 19.3 ± 0.7 | +0.4 ± 0.2 | ii | 162 ± 7 | 35.2 ± 0.1 | 36.8 ± 0.1 | -1.6 ± 0.1 |
| Iii | 178 ± 13 | 22.3 ± 1.3 | 21.1 ± 1.1 | +1.2 ± 0.3 | iii | 169 ± 8 | 39.0 ± 1.2 | 37.5 ± 0.9 | +1.5 ± 0.4 |
| Iv | 175 ± 11 | 20.1 ± 1.3 | 21.1 ± 1.5 | -1.0 ± 0.2 | iv | 171 ± 7 | 35.3 ± 0.5 | 37.5 ± 0.9 | -2.2 ± 0.5 |
| V | 166 ± 14 | 21.3 ± 1.3 | 21.3 ± 1.3 | 0.0 ± 0.0 | v | 168 ± 8 | 37.8 ± 1.4 | 37.8 ± 1.4 | 0.0 ± 0.0 |

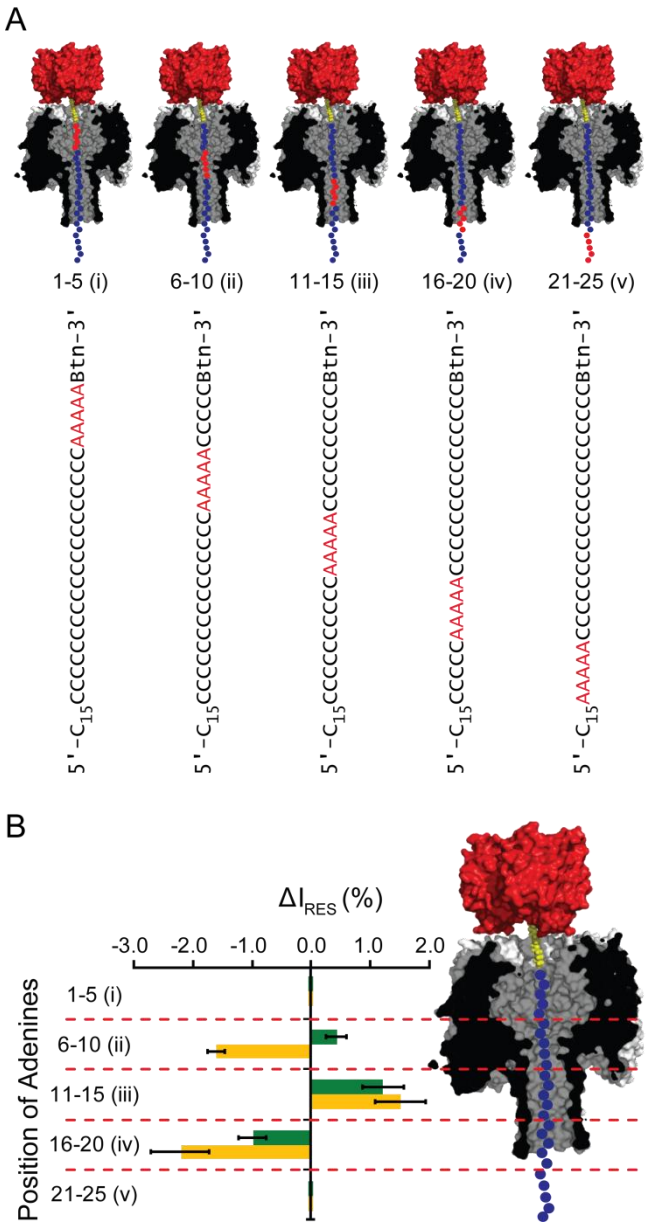



Figure 3.5. Probing the length of the αHL pore for DNA recognition. A) The five oligonucleotides (i-v) containing 5 consecutive adenine nucleobases (red circles) at different positions (numbered from the 3' biotin tag) in an otherwise poly(dC) strand (cytosine nucleobases are shown as blue circles). Only the first 25 of the 40 nucleotide-long sequences are shown. B) Residual current difference ($\Delta I_{RES}$) between the blockades produced by oligonucleotides i-v (panel A) and poly(dC) for WT (green bars) and NN (orange bars) αHL pores ($\Delta I_{RES} = I_{RES}^{A5\ oligo\ i\text{-}v} - I_{RES}^{poly(dC)}$). The probable location of the adenine ($A_5$) stretch of each oligonucleotide when immobilized with an αHL pore is indicated (right). Figure adapted from (Stoddart, Heron et al. 2009).

The $A_1$ substitutions were in positions 7 to 20 relative to the 3' biotin tag (Figure 3.6). $\Delta I_{RES}$ (with respect to poly(dC)) was plotted against the position of the adenine nucleobase for both the WT and NN pores (Figure 3.6, Table 3.4). The pattern of $\Delta I_{RES}$ values for the $A_1$ oligonucleotides mirrored the pattern seen with the $A_5$ oligonucleotides (Figures 3.5 and 3.6). Further, the data suggest that there are three recognition sites within the barrel, which were designated $R_1$, $R_2$ and $R_3$. Recognition site $R_1$ is postulated to be located at the central constriction of the αHL pore, and in the WT pore recognizes bases at positions 7 to 10 in an immobilized DNA strand with a peak at position 9 (Figure 3.6). It follows that $R_2$ is located near the middle of the β barrel and recognizes bases 13 to 15 (peak at position 14) and $R_3$ is located near the *trans* entrance of the pore and recognizes bases 16 to 19 (peak at position 17). These experiments demonstrate that a single base (A versus C) can be recognized in an otherwise identical strand at three broad sites. There are several phenomena that could contribute to the broadness of these recognition sites. It is presumed that the recognition sites are contributed, at least in part, by different regions of the transmembrane β barrel (see above). If each recognition site comprised just a single ring of amino acids, it may be expected to produce sharp recognition sites. However, if the recognition sites are comprised of a number of amino acids, acting cooperatively to provide base identification (e.g. though interactions with the bases), then the recognition sites may be expected to be broad. Alternatively, it is possible that the DNA, while lodged within the pore, may be subject to local Brownian diffusion, resulting in the target base moving in and out of the recognition site. This “rattling” of the base may result in the recognition sites appearing to be

blunt. If such motion is occurring, it is not observed in the current trace, and therefore may be occurring on a faster timescale than the current signal was acquired (5 kHz).

When the WT and NN pores are compared, the $A_1$ scans appear to be about 1 nt out of phase. In the WT pore, compared to the NN, the DNA reaches the recognition element (presumably the β barrel) sooner (position 7 is the first position recognized in WT, compared with position 9 in NN) and leaves earlier (position 18 is the last position recognized in WT, whereas position 19 is the last position recognized in NN), although the total number of bases in the β barrel appears to be similar for both pores. In the NN pore, the electrostatic interactions between Glu-111 and Lys-147 residues at the central constriction have been lost and this may result in a different positioning of the β barrel, with respect to the cap domain; for example, the entire β barrel may have moved downwards slightly (for example, by ~0.3 nm). This slight change in quaternary structure may explain why the recognition elements in the WT and NN pore are out of phase by 1 nt (which is ~0.3 nm, assuming an internucleotide distance similar to that found in B-form dsDNA). An alternative conclusion is that while the DNA in the β barrel of both pores has a similar elongation (the β barrel of both pores contain a similar number of bases), the extent of the elongation of the ssDNA in the vestibule differs slightly, with the DNA in the vestibule of the WT pore being more stretched than in the NN pore. As proposed earlier, the constriction in the WT pore may “grip” or collapse around the DNA, pulling it from within the vestibule. Another possibility is that the ring of positively charged K147 residues in the WT may electrostatically pull on the negatively DNA, causing a slight stretching of the

DNA in the vestibule. The residue at 147 in the NN mutant is a neutral asparagine, which would not exert any such electrostatic force on the DNA, hence the DNA is more relaxed in the vestibule of the NN pore, compared to the WT. However, it may not be expected that the DNA in the vestibule would be electrostatically attracted to a ring of positive lysine residues at the central constriction at the high concentration of electrolyte used (1 M KCl) as the Debye length is only ~0.3 nm (Maglia, Rincon Restrepo et al. 2008).

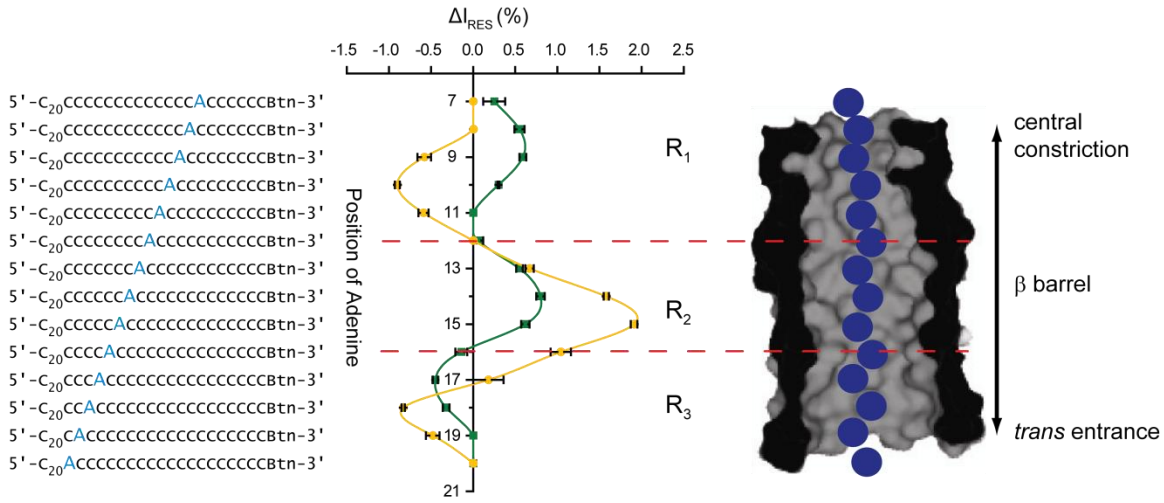



Figure 3.6. Defining the recognition element of the αHL pore. The graph (middle) indicates the differences in residual current ($\Delta I_{RES}$ values) between blockades caused by a poly(dC) oligonucleotide containing a single adenine nucleobase (the sequence of each oligonucleotide is shown to the left) and blockades caused by poly(dC) for WT (green) and NN (orange) αHL pores (see also Table 3.4). $R_1$, $R_2$ and $R_3$ represent the three proposed recognition sites in the αHL nanopore. Their probable locations are indicated on the cross-section of the β barrel domain of the αHL pore (right). Figure adapted from (Stoddart, Heron et al. 2009).

Table 3.4. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single adenine nucleobase. The position of the adenine base in the $A_1$ oligonucleotide (nucleobases 7-20) is numbered relative to the 3' biotin tag. The $I_{RES}$ and $\Delta I_{RES}$ values are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an $A_1$ oligonucleotides and poly(dC) ($I_{RES}^{A1\ oligo} - I_{RES}^{pC}$). The errors given are standard deviations (s.d.). Table adapted from (Stoddart, Heron et al. 2009).

| Position of adenine | WT | | | Position of adenine | NN | | |
|---|---|---|---|---|---|---|---|
| | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| 7 | 20.4 ± 0.4 | 20.2 ± 0.6 | +0.3 ± 0.2 | 7 | 37.0 ± 0.4 | 37.0 ± 0.4 | 0.0 ± 0.0 |
| 8 | 20.1 ± 0.4 | 19.6 ± 0.4 | +0.6 ± 0.1 | 8 | 34.9 ± 3.9 | 34.9 ± 3.9 | 0.0 ± 0.0 |
| 9 | 20.5 ± 0.5 | 19.9 ± 0.4 | +0.6 ± 0.1 | 9 | 34.3 ± 3.8 | 34.9 ± 3.7 | -0.6 ± 0.1 |
| 10 | 20.3 ± 0.2 | 20.0 ± 0.2 | +0.3 ± 0.0 | 10 | 36.1 ± 0.4 | 37.0 ± 0.4 | -0.9 ± 0.1 |
| 11 | 20.0 ± 0.2 | 20.0 ± 0.2 | 0.0 ± 0.0 | 11 | 36.6 ± 0.1 | 37.2 ± 0.1 | -0.6 ± 0.1 |
| 12 | 20.1 ± 0.1 | 20.0 ± 0.2 | +0.1 ± 0.1 | 12 | 35.0 ± 2.2 | 35.0 ± 2.2 | 0.0 ± 0.0 |
| 13 | 20.4 ± 0.3 | 19.8 ± 0.3 | +0.5 ± 0.1 | 13 | 37.0 ± 1.8 | 36.3 ± 1.7 | +0.7 ± 0.1 |
| 14 | 20.5 ± 0.7 | 19.7 ± 0.6 | +0.8 ± 0.1 | 14 | 37.1 ± 1.5 | 35.5 ± 1.4 | +1.6 ± 0.1 |
| 15 | 20.7 ± 0.3 | 20.0 ± 0.3 | +0.6 ± 0.1 | 15 | 38.4 ± 2.2 | 36.5 ± 2.2 | +1.9 ± 0.1 |
| 16 | 19.9 ± 0.2 | 20.0 ± 0.1 | -0.1 ± 0.1 | 16 | 37.4 ± 1.3 | 36.4 ± 1.2 | +1.0 ± 0.2 |
| 17 | 19.3 ± 0.4 | 19.8 ± 0.4 | -0.4 ± 0.0 | 17 | 37.0 ± 0.3 | 36.9 ± 0.1 | +0.2 ± 0.3 |
| 18 | 20.2 ± 0.9 | 20.5 ± 0.9 | -0.3 ± 0.1 | 18 | 36.2 ± 0.3 | 37.0 ± 0.4 | -0.9 ± 0.0 |
| 19 | 20.0 ± 0.3 | 20.0 ± 0.3 | 0.0 ± 0.0 | 19 | 36.1 ± 1.6 | 36.6 ± 1.5 | -0.5 ± 0.1 |
| 20 | 20.5 ± 0.9 | 20.5 ± 0.9 | 0.0 ± 0.0 | 20 | 35.9 ± 1.3 | 35.9 ± 1.3 | 0.0 ± 0.0 |

The above experiments suggest that between 12-15 nucleotides reside in the β barrel when an oligonucleotide is immobilized with the pore; the nucleobases at positions 7-8 (relative to the 3' biotin-tag) are at the entrance of the central constriction and the nucleobases at positions 19-20 are at the *trans* entrance (Figure 3.6). Given that ssDNA, in solution, has a shorter persistence length than dsDNA, and is therefore more contractile (Smith, Cui et al. 1996; Bustamante, Smith et al. 2000), are these assumptions reasonable? It is argued that the ssDNA in the pore is likely to be elongated compared to its conformation in solution, due to the force exerted by the applied potential and because the DNA has been confined within a channel.

Firstly, consider the assumption regarding the number of nucleobases in the β barrel (~12-15). Given that the pore is 10 nm in length (Song, Hobaugh et al. 1996) and assuming an inter-nucleotide distance of ~0.33 nm (as there would be for a strand in a double helix), it can be estimated that there are ~30 nt in the entire lumen of the pore and therefore ~15 nt in the transmembrane β barrel (which is 5 nm in length). The experimentally determined effective charge on each base is ~0.1e (Sauer-Budge, Nyamwanda et al. 2003; Mathé, Visram et al. 2004). This value is consistent with the theory of Zhang and Shklovskii (Zhang and Shklovskii 2007). Therefore, the overall charge (Q) in the β barrel is [15 X (0.1 X (1.6 X $10^{-19}$))] ~2.4 X $10^{-19}$ C. The field in these experiments was 0.16 V over the 5 nm of the barrel; the electric field strength (E) is ~3.2 X $10^{7}$ $Vm^{-1}$. Therefore, the force on the DNA from the applied potential (F = QE) is ~8 pN. Under this force, ssDNA has a similar extension to the B-form of dsDNA (Smith, Cui et al. 1996; Bustamante, Smith et al. 2000), so there would indeed be ~30 nt

in the full length of the pore and about 15 nt in the β barrel. Additionally, the effects of enforced confinement would serve to elongate the DNA still further (Han and Craighead 2000; Han and Craighead 2002; Douville, Huh et al. 2008; Wong and Muthukumar 2008). Furthermore, work conducted by Meller in 2001 also suggests that ~12 nucleotides occupy the β barrel (Meller, Nivon et al. 2001). Therefore, it seems reasonable to infer from the results of the experiments described above, that ~12-15 nucleobases reside in the β barrel.

Secondly, consider if it is reasonable to expect the nucleobases at positions 7-8 to be located near to the central constriction. Taking into account how streptavidin might dock on the *cis* surface of the αHL pore, the location of the biotin binding site within streptavidin and the length of the linker between the DNA and the biotinyl group (~3 nm) (Figure 2.7), then the 3' end of the DNA would be within the vestibule and ~15 Å below the *cis* entrance and ~35 Å above the central constriction (Figure 3.7). Therefore, it would require ~10 nucleobases to reach the central constriction, assuming that the DNA in the vestibule is stretched in a similar manner to the DNA residing within the β barrel. However, given that homopolymeric ssDNA sequences can adopt larger inter-nucleotide distances than dsDNA (e.g. poly(dT) sequences have been measured as having an inter-nucleotide distance of 0.52 nm (Mills, Vacano et al. 1999)), then the assumption that nucleobases at positions 7-8 would be located near the central constriction is not unreasonable. The Schmidt group, working with a similar experimental design came to a similar conclusion, and show similar results (Purnell and Schmidt 2009).

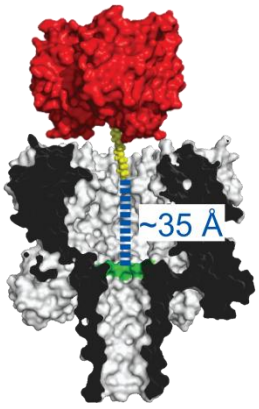



Figure 3.7. The docking of streptavidin at the *cis* entrance of the αHL pore. Using rudimentary modeling in Pymol, the biotin linker (yellow spheres - Figure 2.7) was docked into one of the biotin binding sites in a streptavidin tetramer (red). This complex was manoeuvred such that the streptavidin was brought into close association (without clashing) with the αHL pore (grey cross-section) and the biotin linker protruded into the vestibule. The distance from the end of the biotin linker and the central constriction (shaded green) was estimated as being ~35 Å (blue dashed line).

On the other hand, there is some evidence that might refute these assumptions. In Ashkenasy et al. 5′ hairpins are employed to immobilize ssDNA stretches within the αHL pore, and single nucleobase discrimination is observed at a position 20 nucleotides downstream of the hairpin. Assuming that the dsDNA hairpin is able to enter the pore vestibule and bump up against the central constriction (Vercoutere, Winters-Hilt et al. 2001; Vercoutere, Winters-Hilt et al. 2003; Winters-Hilt, Vercoutere et al. 2003) then this suggests that up to 20 nucleotides reside within the β barrel (Ashkenasy, Sánchez-Quesada et al. 2005; Henrickson, DiMarzio et al. 2010). Further to this, a study from the Kasianowicz group suggested that it takes >40 nucleotides to span the entire pore. In this study, oligonucleotides of different lengths were capped at one end, by an α-BRDU antibody (which binds to a terminal brominated deoxyuracil base), threaded through and immobilized within an αHL pore. The threaded end was

subsequently capped from the other side using avidin (which binds to a terminal biotin tag), to form a rotaxane. It was observed that the oligonucleotide needed to be over 40 nucleotides in length before such rotaxanes could be observed, and it was inferred that it therefore required >40 nucleotides to span the entire pore (Henrickson, DiMarzio et al. 2010). However, what is being measured in this case, is the number of nucleotides required to span the pore and form a rotaxane; it is possible that formation of the rotaxane is subject to certain steric constraints, which prevents the capping protein (e.g. α-BRDU antibody or avidin) from binding to its substrate unless it protrudes far enough out of the entrance of the αHL pore, i.e. the capping protein is unable to dock closely to the entrance of the αHL pore. Such constraints would lead to over estimations in the number of nucleotides required to span the entire pore.

### 3.2.4. Probing the recognition sites of αHL for four-base discrimination

In addition to the detection of individual adenine bases, the ability of αHL pores to distinguish between G, A, T and C bases within a DNA strand was also examined. Both the WT and NN pores were probed with three sets of oligonucleotides; each set comprised four poly(dC) oligonucleotides that contained a single G, A, T and C substitution (the latter oligonucleotide being entirely poly(dC)) at a specific position and each of the three sets had the base substitutions at a different position, which were designed to probe the $R_1$, $R_2$ and $R_3$ recognition sites. To determine how well a given recognition site could discriminate between single nucleobases, two parameters were used. First, the

overall dispersion was measured (the difference between the most widely dispersed residual current levels in the histogram: $\Delta I_{RES}^{OVERALL}$, Figure 2.12) and is shown in Table 3.5. Also, the product of the three sequential differences between each of the four residual current levels (one level for each oligonucleotide) was calculated (δ, Figure 2.12), as a second measure of the ability of a recognition site to discriminate between the four bases. Any site that is unable to discriminate between all four nucleobases (i.e. the current levels from two or more oligonucleotides overlap) has $\delta = 0$. However, sites that are able to discriminate between all four bases have $\delta > 0$ (Table 3.5).

The first set of oligonucleotides had the G, A, T or C substitution at position 9 (from the 3' end) and was designed to probe $R_1$ (Figure 3.8A and Table 3.5). Although there is some discrimination between the four oligonucleotides in this set, neither the WT nor the NN pore was able to distinguish all four bases (i.e. $\delta = 0$). The second set had the G, A, T or C substitution at position 14 and was designed to probe $R_2$ (Figure 3.8B). In this case, both the WT and NN pores clearly separated the bases in the order C, T, A and G (in order of increasing $I_{RES}$ - Table 3.5). The overall dispersion of current levels was far greater for the NN ($\Delta I_{RES}^{OVERALL} = 2.7 \pm 0.1\%$ and $\delta = 0.57 \pm 0.07$) than it was for WT pores ($\Delta I_{RES}^{OVERALL} = 1.1 \pm 0.1\%$ and $\delta = 0.03 \pm 0.02$). The final set had the G, A, T or C substitution at position 18 to probe $R_3$ (Figure 3.8C). In this case, only the NN pores were able distinguish the four bases, but in the reverse order, G, A, T and C (in order of increasing $I_{RES}$), although the dispersion of $I_{RES}$ values was not as large when compared with the set substituted at position 14 ($\Delta I_{RES}^{OVERALL} = 1.2 \pm 0.0\%$ and $\delta = 0.07 \pm 0.00$) (Table 3.5). Although

the actual distribution of current values that are observed cannot be easily rationalized, it does not appear that nucleobase size alone determines the observed current levels. For example, consider the distribution of current levels at $R_2$; C is smaller than G, yet it produces a larger current blockade when immobilized at $R_2$ (Figure 3.8). Therefore it is unlikely that when immobilized within the pore, the DNA is in a conformation where the bases are orientated perpendicular to the axis of the DNA backbone. If this were the case, then base size would be expected to be the biggest influence on current flow. Indeed, in such a conformation, ssDNA has an approximate diameter of 11 Å and given a minimum pore width of 13 Å, at the central constriction (Song, Hobaugh et al. 1996), there would be very little room for hydrated $K^+$ ions (diameter of 6 Å) to pass in between DNA and protein (Figure 3.9). In the absence of DNA, the αHL is a water-filled pore, with largely hydrophilic internal facing amino acids and it is unlikely that ions would become dehydrated in order to pass though the pore when it is blocked with DNA, as this would carry a high energetic penalty. Therefore presumably the amino acid side chains, and/or the bases are tilted (perhaps towards the 5' end, as discussed is section 3.2.1) and allow for hydrated ion transport. Molecular dynamics simulations suggest when ssDNA is immobilized with the pore, the bases are flexible and are able to adopt many different conformations (Jayne Wallace – personal communication). Perhaps the different bases adopt different orientations, in different regions of the pore, and this gives rise to the unique current levels that are observed for the different bases, at the different recognition sites. For example, a C base, at recognition site $R_2$, may be orientated such that it creates a larger block than a G base

(which has a larger volume than a C base). The mechanism of nucleobase discrimination is examined in more detail, with emphasis on recognition site $R_1$, in Chapter 4.

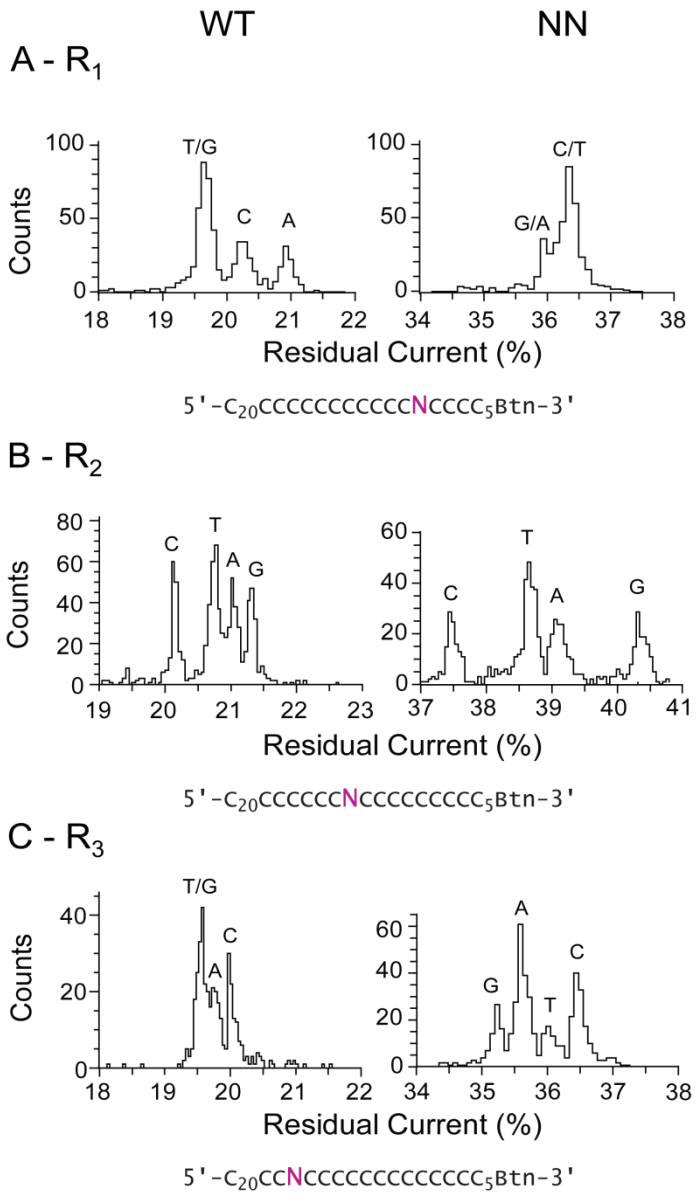



Figure 3.8. Probing the recognition sites of WT and NN pores for four-base discrimination. Histograms of the residual current levels for WT (left) and NN (right) pores are shown. Three sets of four poly(dC) oligonucleotides were used, with each set containing either a single G, A, T, or C nucleobase at a specific position; position 9 (panel – A), position 14 (panel B) and position 18 (panel – C). The results displayed in the figure are from a typical experiment (n ≥ 3). Figure adapted from (Stoddart, Heron et al. 2009).

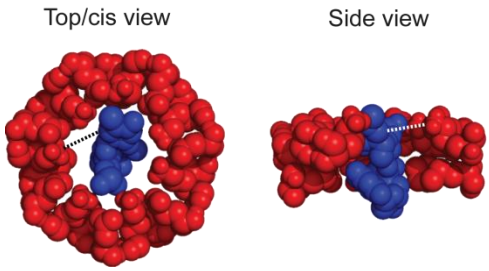



Figure 3.9. Occlusion of the pore by ssDNA. A dinucleotide (with the sequence dTdT – blue) is shown blocking at the central constriction (amino acids Glu-111, Met-113 and Lys-147 are shown – red). In the side view, some amino acids have been removed for clarity The bases are perpendicular to the ssDNA backbone and in this orientation, little room is available around the DNA for hydrated $K^+$ ions to translocate. The dashed line (black in top/cis view and white in side view) between a Lys-147 side chain and a thymine base represents a distance of 8 Å, which is just sufficient to facilitate the translocation of hydrated $K^+$ ions. However, there is insufficient room elsewhere around the base. Figure was made using PyMol.

### 3.2.5. Four-base discrimination in a heteropolymeric sequence.

Subsequently, base discrimination in an otherwise heteropolymeric sequence was tested. A nanopore sequencing instrument will require the ability to recognize individual nucleobases in a heteropolymeric background. Also, despite attempts to minimize secondary structure (i.e. by using a poly(dC) background) it was possible that the pronounced differences in residual current that were observed were the result of disruptions in the DNA structure which resulted in changes in the conformation of the DNA within the nanopore, which in turn affected current flow. If nucleobase discrimination was still observed in a heteropolymer, low in secondary structure (e.g. lacking π-stacked helices), then it would suggest that the different current levels observed were indeed due to base-specific changes, rather than changes in secondary structure.

Table 3.5. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single substituted base at position 9 (corresponding to $R_1$), 14 (corresponding to $R_2$) or 18 (corresponding to $R_3$) blocking WT or NN pores. $I_{RES}$, $\Delta I_{RES}^{OVERALL}$ and δ values are shown for n ≥ 3 experiments. $\Delta I_{RES}^{OVERALL}$ is defined as the difference in residual current between the two most widely dispersed current peaks. The δ value is the product of the successive differences between the peaks (Figure 2.12). If any two peaks overlap, then the δ value is equal to zero. Mean values (± s.d.) are given.

| | Residual Current (%) | | | | Residual Current Difference | |
|---|---|---|---|---|---|---|
| WT | $C_9$ | $T_9$ | $A_9$ | $G_9$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 19.9 ± 0.3 | 19.4 ± 0.2 | 20.5 ± 0.4 | 19.4 ± 0.2 | 1.1 ± 0.1 | 0.00 ± 0.00 |
| | $C_{14}$ | $T_{14}$ | $A_{14}$ | $G_{14}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 18.8 ± 1.5 | 19.4 ± 1.5 | 19.6 ± 1.5 | 19.9 ± 1.6 | 1.1 ± 0.1 | 0.03 ± 0.02 |
| | $C_{18}$ | $T_{18}$ | $A_{18}$ | $G_{18}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 20.1 ± 0.1 | 19.6 ± 0.1 | 19.8 ± 0.2 | 19.6 ± 0.1 | 0.5 ± 0.0 | 0.00 ± 0.00 |
| NN | $C_9$ | $T_9$ | $A_9$ | $G_9$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 35.8 ± 0.9 | 35.8 ± 0.9 | 35.4 ± 0.9 | 35.5 ± 0.9 | 0.4 ± 0.1 | 0.00 ± 0.00 |
| | $C_{14}$ | $T_{14}$ | $A_{14}$ | $G_{14}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 36.1 ± 1.3 | 37.2 ± 1.3 | 37.7 ± 1.3 | 38.8 ± 1.3 | 2.7 ± 0.1 | 0.57 ± 0.07 |
| | $C_{18}$ | $T_{18}$ | $A_{18}$ | $G_{18}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 36.9 ± 0.5 | 36.4 ± 0.5 | 36.0 ± 0.5 | 35.7 ± 0.5 | 1.2 ± 0.0 | 0.07 ± 0.00 |

To examine discrimination within a heteropolymer, the most promising site, $R_2$ in the NN mutant, was tested. The sequence used was chosen by Dr. A. J. Heron and was designed to contain minimal secondary structure, as predicted by the *m*fold algorithm (Zuker 2003), and was unlikely to form π-stacked helices

(Buhot and Halperin 2004). All four bases at position 14 in a heteropolymer were recognized with a similar distribution of residual current levels (C, T, A and G in order of increasing $I_{RES}$) as seen in the homopolymeric background (Figure 3.10 and Table 3.6). The dispersion of the current levels in the histogram ($\Delta I_{RES}^{OVERALL}$ = 3.0 ± 0.2% and δ = 0.90 ± 0.24) was similar to that seen in the homopolymeric background, although the spacing between the four peaks differed in detail.

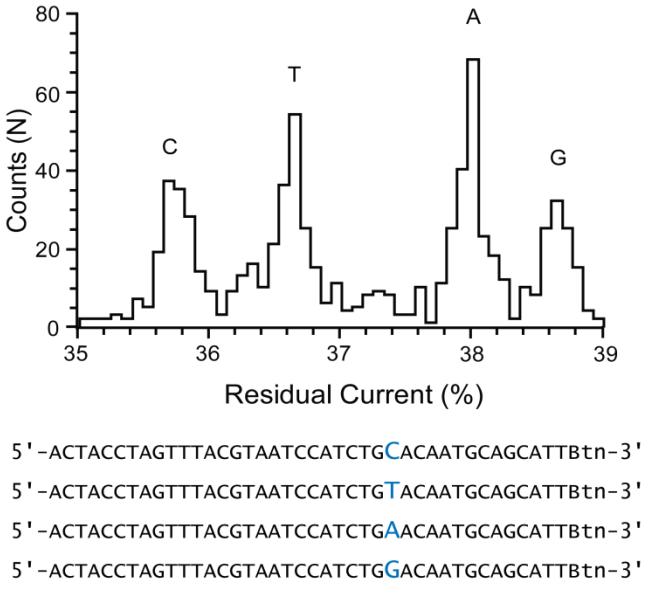



Figure 3.10. Probing the $R_2$ recognition site in the NN αHL pore for single nucleobase discrimination in a heteropolymeric oligonucleotide. Histogram (top) of residual current levels for NN pores interrogated with four heteropolymeric DNA strands (bottom) that differ at only one position (blue). The histogram shown is from a typical experiment (n = 5). Figure adapted from (Stoddart, Heron et al. 2009).

Table 3.6. Residual currents ($I_{RES}$) for heteropolymeric oligonucleotides that contain a single nucleobase at position 14 (corresponding to $R_2$) blocking NN pores. $I_{RES}$, $\Delta I_{RES}$, $\Delta I_{RES}^{OVERALL}$ and δ values are shown for five experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an oligonucleotide that contains either a T, G or A at position 14 and an oligonucleotide that contains a C at position 14 ($\Delta I_{RES} = I_{RES}^{N14} - I_{RES}^{C14}$). $\Delta I_{RES}^{OVERALL}$ is defined as the difference in residual current between the two most widely dispersed current peaks. The δ value is the product of the successive differences between the peaks (Figure 2.12). If any two peaks overlap, then the δ value is equal to zero. Mean values (± s.d.) are given.

| | Nucleobase at $R_2$ (position 14) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | C | T | | A | | G | | | |
| | $I_{RES}$ (%) | $I_{RES}$ (%) | $\Delta I_{RES}^{T-C}$ (%) | $I_{RES}$ (%) | $\Delta I_{RES}^{A-C}$ (%) | $I_{RES}$ (%) | $\Delta I_{RES}^{G-C}$ (%) | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| 1 | 37.0 | 38.0 | 0.9 | 39.4 | 2.4 | 40.0 | 2.9 | 2.9 | 0.78 |
| 2[1] | 35.8 | 36.6 | 0.9 | 38.0 | 2.2 | 38.7 | 2.9 | 2.9 | 0.78 |
| 3 | 35.6 | 36.6 | 0.9 | 37.8 | 2.1 | 38.4 | 2.8 | 2.8 | 0.76 |
| 4 | 35.3 | 36.1 | 0.8 | 37.5 | 2.2 | 38.6 | 3.4 | 3.4 | 1.32 |
| 5 | 37.3 | 38.2 | 0.9 | 39.4 | 2.1 | 40.2 | 2.9 | 2.9 | 0.85 |
| Mean | 36.2 | 37.1 | 0.9 | 38.4 | 2.2 | 39.2 | 3.0 | 3.0 | 0.90 |
| s.d. | 0.9 | 0.9 | 0.0 | 0.9 | 0.1 | 0.8 | 0.2 | 0.2 | 0.24 |

[1] The histogram corresponding to this experiment is displayed in Figure 3.10

## 3.3. Conclusions

Immobilizing DNA strands within the αHL pore using a terminal biotin•streptavidin plug increases the residence of the DNA strand within the pore, resulting in improved precision in the current measurements. Using this approach it has been possible to map the base identification properties of the

αHL pore and show that there are at least three nucleobase recognition sites within the transmembrane β barrel, which are capable of discriminating between single nucleobase differences in DNA strands, based on the different current levels that the bases induce. However, several new challenges have become apparent. The presence of multiple recognition sites could complicate the issue of base determination; consider an oligonucleotide passing through a nanopore that contains a single recognition site, one might envisage that the current observed would contain sequence information about just a single position and therefore the current output would have 4 different levels (one per base). However, the current observed when the same oligonucleotide passed through a nanopore that contains three recognition sites would contain sequence information about three positions, and this would complicate the current read-out ($4^3$ = 64 different current levels). Therefore, it may be necessary to remove superfluous recognition sites. On the other hand, the presence of multiple recognition sites within a single nanopore may improve the quality of the determined sequence, as each base would be read multiple times (once as it passed through each recognition site). Furthermore, this study shows that the three recognition sites that lie within the β barrel, each recognize a number of bases. For example, in the WT pore, recognition site $R_2$ has a peak at position 14, but also recognizes a base either side of this position. In other words, the recognition points are blunt, suggesting that it may be difficult to distinguish between the sequences CA and AC, for example. Therefore, it might be necessary to "sharpen" one or more of the recognition sites by additional mutagenesis.

References


Adler, B. A., L. Grossman, et al. (1967). "Single-stranded oligomers and polymers of cytidylic and 2'-deoxycytidylic acids: comparative optical rotatory studies." Proc Natl Acad Sci U S A **57**(2): 423-430.

Akeson, M., D. Branton, et al. (1999). "Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid and polyuridylic acid as homopolymers or as segments within single RNA molecules." Biophys.J. **77**: 3227-3233.

Ashkenasy, N., J. Sánchez-Quesada, et al. (2005). "Recognizing a single base in an individual DNA strand: a step toward nanopore DNA sequencing." Angew.Chem.Int.Ed.Engl. **44**: 1401-1404.

Buhot, A. and A. Halperin (2004). "Effects of stacking on the configurations and elasticity of single-stranded nucleic acids." Phys Rev E Stat Nonlin Soft Matter Phys **70**(2 Pt 1): 020902.

Bustamante, C., S. B. Smith, et al. (2000). "Single-molecule studies of DNA mechanics." Curr.Op.Struct.Biol. **10**: 279-285.

Clarke, J., H. Wu, et al. (2009). "Continuous base identification for single-molecule nanopore DNA sequencing." Nature Nanotechnology **4**: 265-270.

Cockroft, S. L., J. Chu, et al. (2008). "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J Am Chem Soc **130**(3): 818-820.

Deamer, D. (2010). "Nanopore analysis of nucleic acids bound to exonucleases and polymerases." Annu Rev Biophys **39**: 79-90.

Douville, N., D. Huh, et al. (2008). "DNA linearization through confinement in nanofluidic channels." Anal Bioanal Chem **391**(7): 2395-2409.

Gu, L.-Q., S. Cheley, et al. (2001). "Prolonged residence time of a noncovalent molecular adapter, b-cyclodextrin, within the lumen of mutant a-hemolysin pores." J.Gen.Physiol. **118**: 481-494.

Gu, L.-Q., M. Dalla Serra, et al. (2000). "Reversal of charge selectivity in transmembrane protein pores by using non-covalent molecular adapters." Proc.Natl.Acad.Sci.USA **97**: 3959-3964.

Han, J. and H. G. Craighead (2000). "Separation of long DNA molecules in a microfabricated entropic trap array." Science **288**(5468): 1026-1029.

Han, J. and H. G. Craighead (2002). "Characterization and optimization of an entropic trap for DNA separation." Anal Chem **74**(2): 394-401.

Henrickson, S. E., E. A. DiMarzio, et al. (2010). "Probing single nanometer-scale pores with polymeric molecular rulers." J Chem Phys **132**(13): 135101.

Kasianowicz, J. J., S. E. Henrickson, et al. (2001). "Simultaneous multianalyte detection with a nanometer-scale pore." Analyt.Chem. **73**: 2268-2272.

Maglia, G., M. Rincon Restrepo, et al. (2008). "Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge." Proc Natl Acad Sci U S A **105**: 19720-19725.

Mathé, J., H. Visram, et al. (2004). "Nanopore unzipping of individual DNA hairpin molecules." Biophys.J. **87**: 3205-3212.

Mathé, J., Aksimentiev, A., et al. (2005). "Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel." Proc.Natl.Acad.Sci.USA **102**: 12377-12382.

Meller, A., L. Nivon, et al. (2000). "Rapid nanopore discrimination between single polynucleotide molecules." Proc.Natl.Acad.Sci.USA **97**: 1079-1084.

Meller, A., L. Nivon, et al. (2001). "Voltage-driven DNA translocations through a nanopore." Phys.Rev.Lett. **86**: 3435-3438.

Mills, J. B., E. Vacano, et al. (1999). "Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA)." J Mol Biol **285**(1): 245-257.

Mitchell, N. and S. Howorka (2008). "Chemical tags facilitate the sensing of individual DNA strands with nanopores." Angew Chem Int Ed Engl **47**(30): 5565-5568.

Olasagasti, F., K. R. Lieberman, et al. (2010). "Replication of individual DNA molecules under electronic control using a protein nanopore." Nat Nanotechnol.

Purnell, R. F., K. K. Mehta, et al. (2008). "Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore." Nano Lett **8**(9): 3029-3034.

Purnell, R. F. and J. J. Schmidt (2009). "Discrimination of single base substitutions in a DNA strand immobilized in a biological nanopore." ACS Nano **3**: 2533-2538.

Sánchez-Quesada, J., A. Saghatelian, et al. (2004). "Single molecule DNA rotaxanes of a transmembrane pore protein." Angew.Chem.Int.Ed.Engl. **43**: 3063-3067.

Sauer-Budge, A. F., J. A. Nyamwanda, et al. (2003). "Unzipping kinetics of double-stranded DNA in nanopores." Phys.Rev.Lett. **90**: 238101-238101-238101-238104.

Smith, S. B., Y. Cui, et al. (1996). "Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules." Science **271**: 795-799.

Song, L., M. R. Hobaugh, et al. (1996). "Structure of staphylococcal a-hemolysin, a heptameric transmembrane pore." Science **274**: 1859-1865.

Stoddart, D., A. Heron, et al. (2009). "Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore." Proc. Natl. Acad. Sci. USA **106**: 7702-7707.

Stoddart, D., A. J. Heron, et al. (2010). "Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore." Nano Lett **10**(9): 3633-3637.

Vercoutere, W., S. Winters-Hilt, et al. (2001). "Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel." Nature Biotechnology **19**: 248-252.

Vercoutere, W. A., S. Winters-Hilt, et al. (2003). "Discrimination among individual Watson-Crick bases pairs at the termini of single DNA hairpin molecules." Nucl.Acids Res. **31**: 1311-1318.

Winters-Hilt, S., W. Vercoutere, et al. (2003). "Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules." Biophys.J. **84**: 967-976.

Wong, C. T. and M. Muthukumar (2008). "Polymer translocation through a cylindrical channel." J Chem Phys **128**(15): 154903.

Zhang, J. and B. I. Shklovskii (2007). "Effective charge and free energy of DNA inside an ion channel." Phys Rev E Stat Nonlin Soft Matter Phys **75**(2 Pt 1): 021906.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic Acids Res **31**(13): 3406-3415.

# CHAPTER 4: NUCLEOBASE RECOGNITION IN ssDNA AT THE CENTRAL CONSTRICTION OF THE α-HEMOLYSIN PROTEIN NANOPORE

Declaration

The work in this chapter is also described in the manuscripts:

Stoddart, D., Heron, A. J., Klingelhoefer, J., Mikhailova, E., Maglia, G. and Bayley, H. (2010). Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore." Nano Lett **10**(9): 3633-3637.

Author contributions: D.S., A.H., G.M. and H.B. designed the concept; D.S. performed all single-channel recordings, J.K. performed molecular dynamics simulations and E.M. performed the mutant and protein preparations; D.S. analyzed data; and D.S. and H.B. wrote the paper.

Stoddart, D., Maglia, G., Mikhailova, E., Heron, A. J. and Bayley, H. (2010). Multiple base-recognition sites in a biological nanopore – two heads are better than one." Angew Chem Int Ed **49**: 556-559.

Author contributions: D.S., A.H., G.M. and H.B. designed the concept; D.S. performed all single-channel recordings, and E.M. performed the mutant and protein preparations; D.S. analyzed data; and D.S. and H.B. wrote the paper.

Abstract

DNA strands can be immobilized within the αHL pore by using a terminal DNA-btn•streptavidin complex (that prevents full translocation of the DNA) and this improves the resolution of the current signatures associated with individual nucleobase substitutions, because of the prolonged observation time. The αHL pore contains three nucleobase recognition sites within the transmembrane β

barrel domain (see previous chapter). In this chapter, recognition site $R_1$, which is located at the central constriction, was modified by extensive mutagenesis of residue 113. Amino acids that provide an energy barrier to ion flow (e.g. bulky or hydrophobic residues) strengthen base identification, while amino acids that lower the energy barrier weaken identification. Amino acids with related side chains produce similar patterns of nucleobase recognition providing a rationale for the redesign of recognition sites. It is also shown that additional sequence information can be gained when two recognition sites, rather than one, are employed within a single nanopore.

## 4.1. Introduction

It has been shown that the WT αHL pore is capable of discriminating between different oligonucleotide sequences, based on single-channel current measurements (Akeson, Branton et al. 1999; Meller, Nivon et al. 2000) (Ashkenasy, Sánchez-Quesada et al. 2005) (Purnell and Schmidt 2009; Stoddart, Heron et al. 2009). Although base identification in ssDNA molecules has been studied by a variety of techniques (e.g. freely translocating DNA, immobilization through terminal hairpin), the DNA-btn•streptavidin approach is advantageous, as it is easy to control (DNA can be immobilized, essentially permanently, and it is possible to predict the location of target bases within the pore), it mimics the pauses between enzymatic base additions (during which the DNA strand is held within the pore lumen, stretched in the applied potential) and it shouldn't interfere with the base recognition machinery of the transmembrane β

barrel in the αHL pore (unlike the use of hairpins or other dsDNA “brakes” which could interfere with the recognition mechanism (Ashkenasy, Sánchez-Quesada et al. 2005; Derrington, Butler et al. 2010)). Using the DNA-btn•streptavidin approach, extensive mapping of the αHL pore narrowed down the recognition element to the transmembrane β barrel domain, which contains ~12-15 nucleobases and has three broad nucleobase recognition sites (Stoddart, Heron et al. 2009), termed $R_1$, $R_2$ and $R_3$ (Chapter 3). It is likely that the properties of these recognition sites will have to be modified before the pore could be used in any working DNA sequencing device: recognition sites may need to be enhanced, sharpened, or removed entirely in order to improve and/or simplify the current signals that are outputted. However, in order to be able to redesign base recognition in a controlled manner, it is necessary to understand the mechanism behind it. The central constriction, which in the WT αHL pore, comprises of Glu-111, Met-113 and Lys-147 residues (Song, Hobaugh et al. 1996), is involved in nucleobase recognition at site $R_1$. Studies in the previous chapter showed that the E111N/K147N mutant (has amino acid substitutions at the central constriction) has different nucleobase recognition properties at site $R_1$, compared to WT. The work in this chapter continues to focus on $R_1$, and site-directed mutagenesis of Met-113 (the third amino acid at the central constriction) is used to investigate the effects that different amino acid side chains have on base recognition in homoheptameric pores, in an attempt to provide a rationale for the redesign of the nucleobase recognition sites.

## 4.2. Results and discussion

### 4.2.1. Effects of amino acid substitutions at position 113 on nucleobase recognition at recognition site $R_1$

The base discrimination properties of fifteen E111N/K147N/M113X (NNX) mutant homoheptameric pores (mutants generated and proteins prepared by E. Mikhailova) were tested with the set of four poly(dC) oligonucleotides that contained single nucleobase substitutions at position 9 (Table 4.1), that had been previously used to the probe the $R_1$ recognition site of WT and NN pores (see Chapter 3, section 3.2.4 and (Stoddart, Heron et al. 2009)). The experiments to probe the $R_1$ recognition site of the NNX pores were conducted under similar conditions as previously described (see Chapter 3 section 3.2.1) and it was observed that although altering the volume of the amino acid side chain at position 113 had no predictable effect on the open pore current level (Figure 4.1A) there was a slight decrease in the residual current of poly(dC) blockades as the volume of the amino acid side increased (Figure 4.1B). One proposal is that in the unblocked pore, the changes in the volume of the amino acid side chain at position 113 have no significant effect on the volume available for the translocation of hydrated ions. However, in the blocked pore much of the pore volume is occluded (by the DNA) and the volume available for hydrated ions is reduced. Under these conditions, changes in the volume of the amino acid at position 113 will have a more significant effect on the volume available for the translocation of hydrated ions, and this is manifested in a modest decrease in the

residual current level for mutants with large amino acids at position 113. An alternative explanation is based on the electrostatic self-energy of the translocating ions. As ions translocate through a nanopore they produce an electric field. It has been argued that when DNA is within the αHL pore the electric field lines of the translocating ions in the water filled channel between the ssDNA and the wall of the protein pore are confined (Bonthuis, Zhang et al. 2006). This leads to an increase in the "self-energy" of the ions (which is the leading term in the energy of a charge in a given medium (Parsegian 1969)), compared to when the pore is not blocked with ssDNA. An increase in ion self-energy represents an electrostatic barrier to ion flow, and as such will result in a diminished current flow. Such an electrostatic self-energy barrier was predicted by Parsegian (Parsegian 1969). Given this, it is envisaged that narrowing the channel with large amino acid side chains will further squeeze the electric field lines of the translocating ions, increasing their self-energy. This will lead to an increase in the electrostatic energy barrier and therefore we observe a reduction in the residual ionic current for mutants with large amino acids chains at position 113, when blocked with ssDNA. When the pores are not blocked with DNA, narrowing the pore by means of amino acid substitutions will have little effect on the self-energy of the translocating ions and therefore, we do not observe a reduction in the open pore current for mutants with large amino acids at position 113.

In addition to the modest influence that side chain volume has on residual current values, it appears that placing positively charged amino acids at position 113 (Arg or Lys) results in a more pronounced reduction in $I_{RES}$ (Figure 4.1B).

When the αHL pore is blocked with DNA, it becomes cation selective (Sanchez-Quesada, Saghatelian et al. 2004). Therefore, increasing the positive charge within the pore (by substituting Arg or Lys amino acids in at position 113) will create an additional energy barrier to the flow of cations, which will reduce the current flow through the pore when it is blocked with DNA. As well as this proposed coulombic block, it is possible that the side chains of Arg or Lys extend and clamp onto the DNA, severely reducing the solvent volume and hence the current flow at the constriction (Cheley, Gu et al. 2002). Furthermore, it is possible that partial collapse of the barrel onto the DNA strand occurs in this situation (in a similar way as discussed for the K147 residue in WT pores in Chapter 3, section 3.2.2), which would also reduce $I_{RES}$.

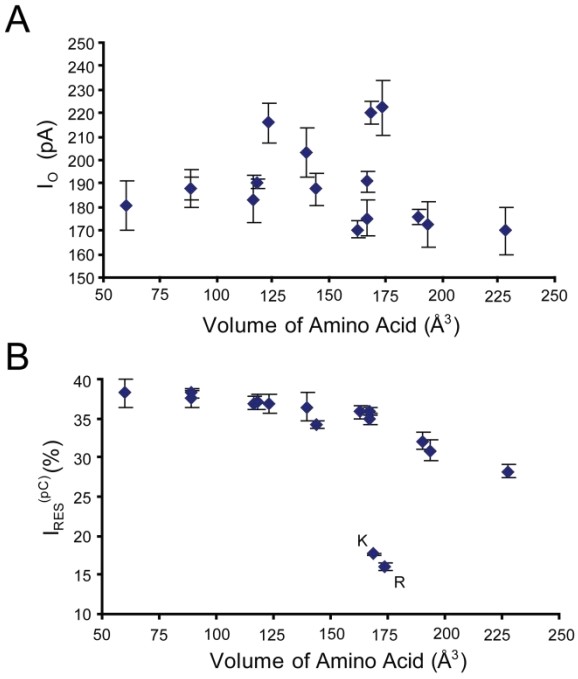



Figure 4.1. The effect of amino acid substitutions at position 113 on the open pore and residual current levels. A) The effect of the volume of the amino acid at position 113 on $I_O$. B) The effect of the volume of the amino acid at position 113 on $I_{RES}$ of a poly(dC) oligonucleotide. Figure adapted from (Stoddart, Heron et al. 2010).

An NNE mutant pore (with a negatively charged Glu amino acid at position 113) was also tested. However, no DNA capture was observed, under these experimental conditions. It has been previously reported that the introduction of negatively charged amino acids at the central constriction reduces DNA capture (Maglia, Rincon Restrepo et al. 2008). Presumably, the electrostatic repulsion between the ring of negatively charged amino acids in the homoheptamer, and the phosphate groups in the DNA backbone prevents DNA threading through the pore. Given that the NNE pore did not capture DNA, further experiments with negatively charged mutants were not pursued.

Subsequently, the effect of altering the volume of the amino acid side chains at position 113 on discrimination between single nucleobases at $R_1$ was examined. $I_{RES}$ values for each of the four oligonucleotides were determined from histograms of oligonucleotide current blockades (Table 4.1, Figure 4.2). To determine how well a particular mutant could discriminate between single nucleobases, two parameters were used. First, the $\Delta I_{RES}{}^{OVERALL}$ (Figure 2.12) was measured and plotted against the volume of the amino acid at 113 (Figure 4.3A, Table 4.1). In general, larger side chains at position 113 produce an increase in $\Delta I_{RES}{}^{OVERALL}$ and therefore improve nucleobase discrimination. Second, the δ values (Figure 2.12) were calculated and were also plotted against the volume of the amino acid at position 113 (Figure 4.B, Table 4.1). An αHL pore that is unable to discriminate between all four nucleobases (i.e. the current levels from two or more oligonucleotides overlap) has $\delta = 0$. However, pores that are able to discriminate between all four bases have $\delta > 0$, as was the case for NN pores with Val, Leu, Ile, Tyr, Trp or Arg at position 113.

Table 4.1. E111N/K147/M113X (NNX) mutant αHL pores probed with a set of four poly(dC) oligonucleotides that contain single nucleobase substitutions at position 9, relative to the 3' biotin tag. The amino acid residue volumes are from Zamyatnin (Zamyatnin 1972). The $I_O$, $I_{RES}$ (for each of the four oligonucleotides), $\Delta I_{RES}^{X-C}$, $\Delta I_{RES}^{OVERALL}$ and δ values (Figure 2.12) for each mutant were measured and the mean values (± s.d.) from the histograms from at least three experiments are displayed. $\Delta I_{RES}^{X-C}$ is the difference in residual current between a poly(dC) oligonucleotide that containing an A, T or G base at position 9, and an entirely poly(dC) oligonucleotide. $\Delta I_{RES}^{OVERALL}$ is the mean value of the differences in residual current between the two most widely dispersed current levels in the $I_{RES}$ histograms. The δ value is the mean value of the product of the three successive differences between the four current levels in the $I_{RES}$ histograms (Figure 4.2). Table adapted from (Stoddart, Heron et al. 2010).

| AA at 113 | Volume of AA at 113 ($Å^3$) | $I_O$ ± s.d. (pA) | $I_{RES}^{pC}$ ± s.d. (%) | $I_{RES}^{A}$ ± s.d. (%) | $\Delta I_{RES}^{A-C}$ ± s.d. (%) | $I_{RES}^{T}$ ± s.d. (%) | $\Delta I_{RES}^{T-C}$ ± s.d. (%) | $I_{RES}^{G}$ ± s.d. (%) | $\Delta I_{RES}^{G-C}$ ± s.d. (%) | $\Delta I_{RES}^{OVERALL}$ ± s.d. (%) | δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly (G) | 60.1 | 181 ± 11 | 38.2 ± 1.8 | 38.2 ± 1.8 | 0.0 ± 0.0 | 38.2 ± 1.8 | 0.0 ± 0.0 | 38.2 ± 1.8 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.00 ± 0.00 |
| Ala (A) | 88.6 | 188 ± 5 | 38.2 ± 0.6 | 37.8 ± 0.5 | −0.4 ± 0.1 | 39.0 ± 0.6 | +0.8 ± 0.0 | 38.2 ± 0.6 | 0.0 ± 0.0 | 1.2 ± 0.0 | 0.00 ± 0.00 |
| Ser (S) | 89.0 | 188 ± 8 | 37.5 ± 1.1 | 37.2 ± 1.1 | −0.3 ± 0.0 | 37.5 ± 1.1 | 0.0 ± 0.0 | 37.5 ± 1.1 | 0.0 ± 0.0 | 0.3 ± 0.0 | 0.00 ± 0.00 |
| Thr (T) | 116.1 | 183 ± 10 | 36.9 ± 0.9 | 36.3 ± 0.9 | −0.6 ± 0.0 | 36.3 ± 0.9 | −0.6 ± 0.0 | 36.7 ± 0.9 | −0.2 ± 0.0 | 0.6 ± 0.0 | 0.00 ± 0.00 |
| Asn (N) | 117.7 | 190 ± 2 | 37.0 ± 0.9 | 36.6 ± 1.0 | −0.4 ± 0.1 | 37.0 ± 1.0 | −0.1 ± 0.1 | 36.8 ± 0.9 | −0.2 ± 0.0 | 0.4 ± 0.1 | 0.00 ± 0.00 |
| Pro (P) | 122.7 | 216 ± 9 | 36.9 ± 1.2 | 36.2 ± 1.2 | −0.7 ± 0.1 | 36.6 ± 1.2 | −0.2 ± 0.0 | 36.9 ± 1.2 | 0.0 ± 0.0 | 0.7 ± 0.1 | 0.00 ± 0.00 |
| Val (V) | 140.0 | 203 ± 11 | 36.5 ± 1.7 | 35.5 ± 1.7 | −1.0 ± 0.1 | 36.8 ± 1.7 | +0.3 ± 0.1 | 36.0 ± 1.7 | −0.5 ± 0.1 | 1.3 ± 0.1 | 0.08 ± 0.02 |
| Gln (Q) | 143.9 | 188 ± 7 | 34.2 ± 0.5 | 34.2 ± 0.5 | 0.0 ± 0.0 | 34.2 ± 0.5 | 0.0 ± 0.0 | 33.8 ± 0.5 | −0.4 ± 0.0 | 0.4 ± 0.0 | 0.00 ± 0.00 |
| Met (M) (WT) | 162.9 | 170 ± 4 | 35.8 ± 0.9 | 35.4 ± 0.9 | −0.4 ± 0.1 | 35.8 ± 0.9 | 0.0 ± 0.0 | 35.5 ± 0.9 | −0.3 ± 0.1 | 0.4 ± 0.1 | 0.00 ± 0.00 |
| Leu (L) | 166.7 | 175 ± 7 | 35.8 ± 0.5 | 34.9 ± 0.5 | −0.9 ± 0.0 | 36.1 ± 0.5 | +0.3 ± 0.0 | 35.2 ± 0.5 | −0.6 ± 0.0 | 1.2 ± 0.1 | 0.05 ± 0.01 |
| Ile (I) | 166.7 | 191 ± 4 | 34.9 ± 0.8 | 33.3 ± 0.7 | −1.6 ± 0.1 | 35.3 ± 0.8 | +0.4 ± 0.0 | 34.0 ± 0.8 | −0.9 ± 0.0 | 2.0 ± 0.1 | 0.24 ± 0.02 |
| Lys (K) | 168.6 | 220 ± 5 | 17.6 ± 0.2 | 17.6 ± 0.2 | 0.0 ± 0.0 | 16.9 ± 0.2 | −0.7 ± 0.0 | 16.4 ± 0.2 | −1.3 ± 0.0 | 1.3 ± 0.0 | 0.00 ± 0.00 |
| Arg (R) | 173.4 | 222 ± 12 | 16.1 ± 0.5 | 15.2 ± 0.5 | −0.9 ± 0.0 | 14.8 ± 0.5 | −1.3 ± 0.1 | 13.9 ± 0.5 | −2.2 ± 0.1 | 2.2 ± 0.1 | 0.31 ± 0.04 |
| Phe (F) | 189.9 | 176 ± 3 | 32.1 ± 1.1 | 30.6 ± 1.1 | −1.5 ± 0.0 | 30.6 ± 1.1 | −1.5 ± 0.0 | 30.6 ± 1.1 | −1.5 ± 0.0 | 1.5 ± 0.0 | 0.00 ± 0.00 |
| Tyr (Y) | 193.6 | 172 ± 7 | 31.4 ± 1.3 | 30.0 ± 1.2 | −1.5 ± 0.1 | 28.6 ± 1.2 | −2.9 ± 0.1 | 29.4 ± 1.2 | −2.0 ± 0.1 | 2.9 ± 0.2 | 0.70 ± 0.11 |
| Trp (W) | 227.8 | 170 ± 10 | 28.3 ± 0.8 | 26.3 ± 0.8 | −2.0 ± 0.0 | 25.4 ± 0.8 | −2.9 ± 0.1 | 26.0 ± 0.8 | −2.2 ± 0.0 | 2.9 ± 0.1 | 0.31 ± 0.02 |

The hydrophobic (Val, Leu and Ile) and bulky (Tyr and Trp) side chains provide steric barriers to ion flow, while the positively charged Arg residues may provide both a coulombic block and a steric block (by gripping the DNA, or collapsing around it) to ion flow. The increased energy barrier at the central constriction provided by such amino acid side chains may promote interactions between the flowing ions and the immobilized bases at $R_1$, augmenting nucleobase identification. Glycine has no side chain and the E111N/K147N/M113G (NNG) pore presents a low energy barrier to ion flow (NNG pores have high $I_{RES}$ values) and exhibits the weakest base identification properties (both $\Delta I_{RES}^{OVERALL}$ and $\delta = 0$).

Interactions between the amino acids at position 113 and the ssDNA could also influence nucleobase discrimination; each of the four bases may adopt different average orientations as they seek to maximize specific interactions with the protein. If each base were to adopt a unique orientation, this could result in unique current signatures that are observed for the different bases. The amino acid side chains were categorized according to their properties and the residual current differences between the oligonucleotides containing A, T and G (at position 9) and the oligonucleotide containing C (at position 9) were assigned to them (Figure 4.3C, Table 4.1): $\Delta I_{RES}^{X-C} = I_{RES}^{X} - I_{RES}^{C}$ (where X = A, T or G).

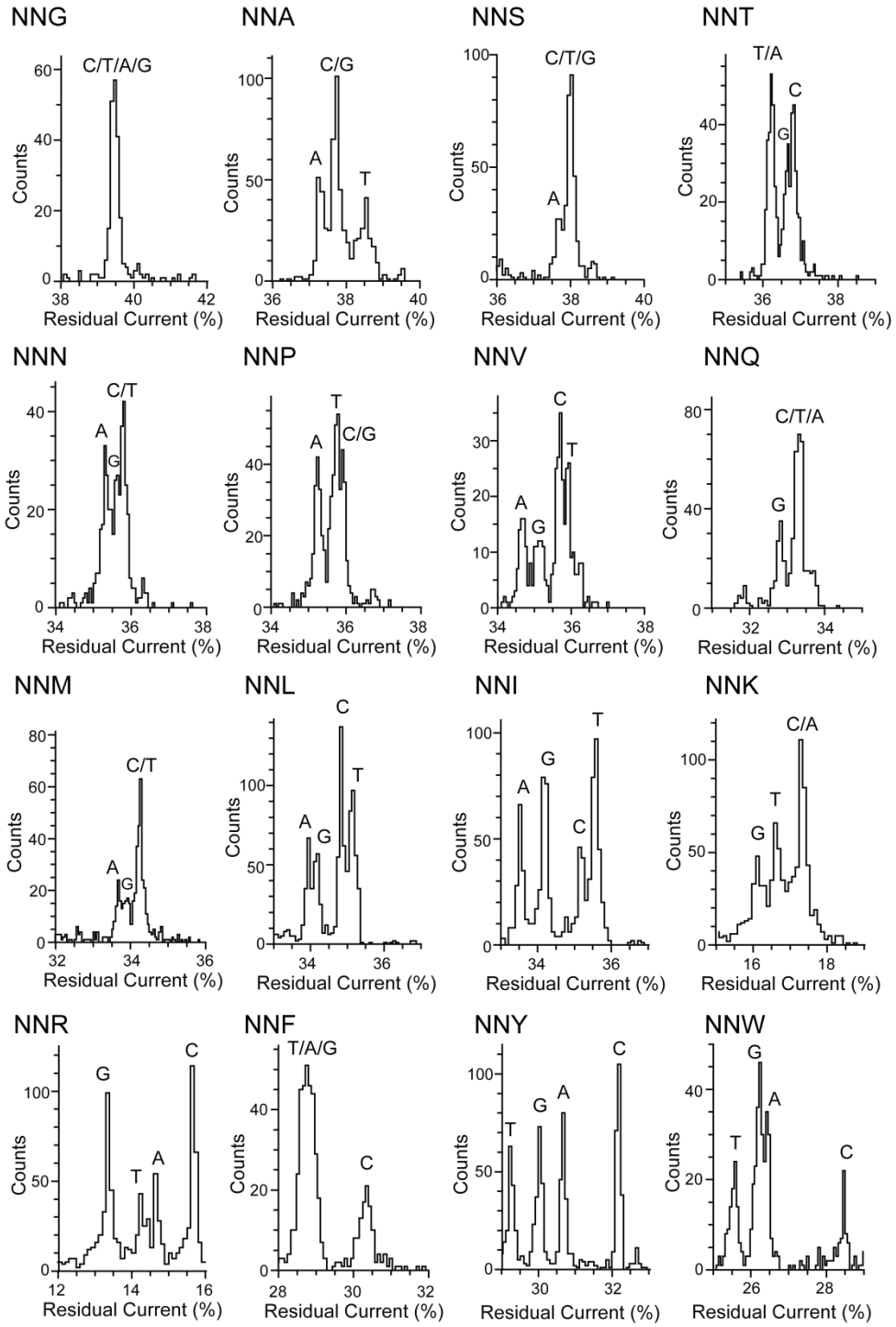



Figure 4.2. Histograms of the residual current levels observed when E111N/K147N/M113X (NNX) mutants were probed with a set of four poly(dC) oligonucleotides that contained single nucleobase substitutions at position 9, relative to the 3' biotin tag. The histogram shown is from a typical experiment (n ≥ 3). Figure adapted from (Stoddart, Heron et al. 2010).

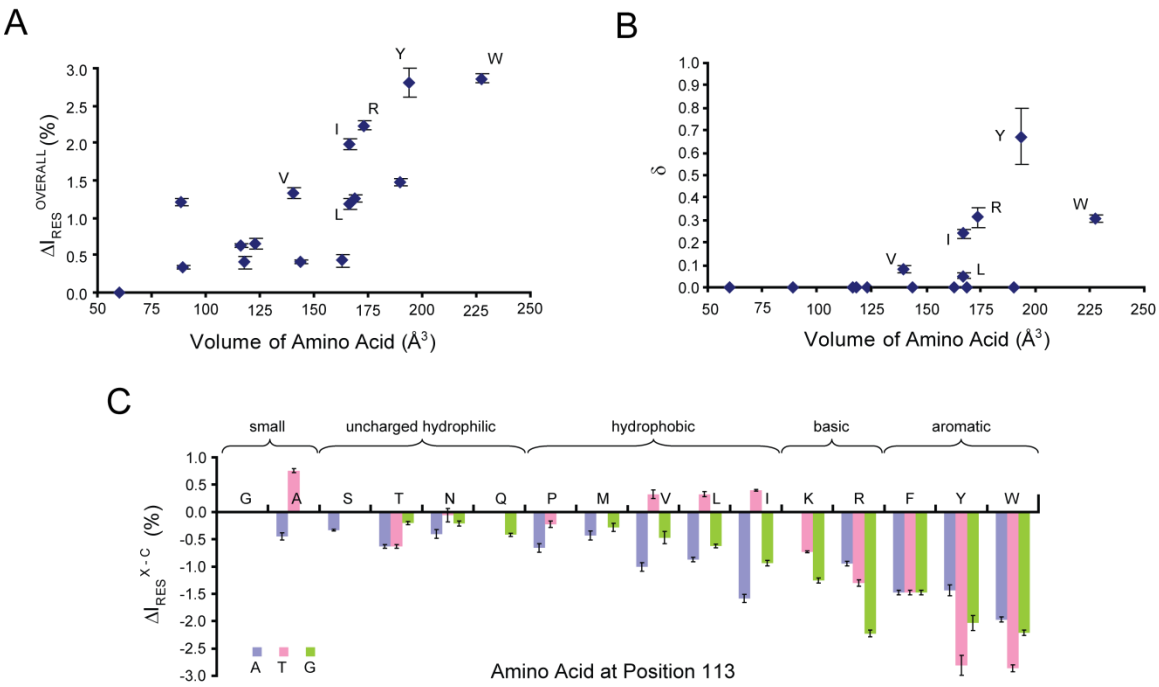



Figure 4.3. The effect of amino acid substitutions at position 113 on the residual current levels for poly(dC) oligonucleotides with single nucleobase substitutions at position 9. A) The relationship between the volume of the amino acid at position 113 and the difference between the two most widely dispersed current levels ($\Delta I_{RES}^{OVERALL}$). B) The relationship between the volume of the amino acid at position 113 and the product of the three successive differences between the four current levels in a histogram of residual currents (δ). The mutants that discriminate between all four oligonucleotides are indicated (A and B). C) The differences in residual current between the oligonucleotides containing A, T and G at position 9 and the oligonucleotide containing a C at position 9 ($\Delta I_{RES}^{X-C} = I_{RES}^{X} - I_{RES}^{C}$) are plotted for each amino acid substituted into position 113. Amino acids are categorized based on the properties of their side chains. Figure adapted from (Stoddart, Heron et al. 2010).

Patterns in base discrimination between amino acids with related side chains are readily apparent. First, amino acids with small or uncharged hydrophilic side chains discriminate between bases poorly. For example, glycine does not show discrimination of any nucleobase at position 9. By contrast, amino acids with hydrophobic, basic or aromatic side chains show superior base discrimination, which varies in character between the amino acid classes: for the hydrophobic amino acids (Val, Leu, Ile): A < G < C < T (order of $I_{RES}$ value); basic

amino acids (Lys, Arg): G < T < A ≤ C; aromatic amino acids (Tyr, Trp): T < G < A < C.

All three of these classes of amino acids have documented interactions with nucleobases. Structural data for protein•single-stranded nucleic acid (ssNA) complexes have provided great insight into how proteins interact specifically with DNA and RNA bases (see “Protein-Nucleic Acid Interactions: Structural Biology” by P. A. Rice and C. C. Correll and references therein). Hydrophobic, basic and aromatic classes of amino acids are known to form specific interactions with bases, *in vivo*. Although there are recurring trends observed in protein•ssNA complexes, often they involve diverse and extensive combinations of different types of amino acid•base interaction.

Hydrophobic residues form close van der Waals contacts with nucleic acid bases in a variety of complexes. For example, several hydrophobic residues in splicing factor SF1 make extensive contacts with substrate RNA bases (Liu, Luyten et al. 2001). A “hydrophobic binding pocket” in SF1 that specifically binds to a key adenine base in the branch-point sequence of intronic RNA contains a leucine, a valine and several isoleucines. Mutation of the leucine in this pocket reduces RNA binding affinity by ~50%, while other leucines in SF1 also make specific van der Waals contacts with other bases in the branch-point sequence. A further example of specific interactions between bases and hydrophobic amino acids is evident in the telomere end-binding protein (TEBP)-α subunit. In this case, a hydrophobic protein pocket containing a leucine and an isoleucine binds to the penultimate guanine base of telomere ssDNA (Horvath, Schweiker et al. 1998). Computational analysis suggests that van der Waals interactions are

prevalent in protein•ssNA complexes and it is likely that the driving force for such hydrophobic interactions is similar to the so-called “hydrophobic effect” that drives the folding of hydrophobic protein cores (Jones, Daley et al. 2001). It is possible that van der Waals contacts between hydrophobic amino acids substituted into position 113, such as Val, Leu and Ile, and the different nucleobases result in the four bases adopting different orientations to each other, producing the unique current signatures observed for the bases.

Amino acids which have basic side chains are capable of forming multiple types of interactions with bases. In the *trp* RNA-binding attenuation protein (TRAP)-RNA complex, the carbonyl oxygen of a guanine base hydrogen-bonds with the Nε of Lys-56 and the side chain $NH_2$ or Arg-58 (Antson, Dodson et al. 1999). As well as H-bonding interactions, lysine and arginine amino acids are also capable of forming cation-π interactions with bases. For example, the SF1 protein contains a lysine residue, adjacent to the hydrophobic binding pocket outlined above, which forms a cation-π interaction with an adenine base in the branch-point sequence (Liu, Luyten et al. 2001). Furthermore, the fact that such basic amino acids are capable of forming multiple types of interactions is important in substrate binding. For example, in the single-stranded DNA-binding domain of replication protein A-DNA complex, Arg-210 is observed to H-bond and form a cation-π interaction with the DNA. The H-bond is formed between the arginine $NH_2$ side chain group and a phosphate group. While this is not a “base-specific” interaction, it is necessary to align the Arg-210 side chain so that it can form a base specific cation-π interaction with a cytosine (Bochkarev, Pfuetzner et al. 1997). These types of interactions are common in protein•ssNA complexes

(Gallivan and Dougherty 1999; Allers and Shamoo 2001; Wang, McLachlan et al. 2002).

Similarly, tyrosine and tryptophan are also capable of H-bonding to bases, via the hydroxyl group of the phenolic ring or the pyrrole nitrogen of the indole ring, respectively. Such interactions are common in protein•ssNA complexes (Allers and Shamoo 2001). As well as forming H-bonds, tyrosine and tryptophan are also capable of forming stacking interactions with heteroaromatic bases in extended single-stranded polynucleotides. The viral methyltransferase enzyme VP39, recognizes and binds to the methylated guanine at the cap region of mRNA. This specific interaction between protein and base relies upon a stacking interaction with a tyrosine (Hu, Gershon et al. 1999). In *E.coli* SSB-ssDNA complex, there are extensive stacking interactions between the side chains of tryptophans on the protein, with DNA bases (Raghunathan, Kozlov et al. 2000).

Hence, in these cases, the NNX mutants have a ring of seven amino acids at position 113, that have the potential of forming multiple interactions with a solitary base (at position 9 in the immobilized strand); for example, the base may form an H-bond with the amino acid from one subunit, and may form a stacking interaction with the equivalent amino acid on a neighboring subunit. Such interactions may result in different bases adopting specific orientations, in an attempt to maximize the available interactions with the protein, and this may result in the unique current signatures observed. Amino acids that are capable of forming only either hydrogen-bonds (Asn, Gln, Ser and Thr) or stacking interactions with bases (Phe) (Lei, Podell et al. 2004) are poor at discriminating between nucleobases at $R_1$ (Figure 4.3B and C and Table 4.1). In these cases,

where fewer types of interaction were possible between protein and immobilized base, the different bases may not have been able to form specific orientations within the pore and therefore, did not elicit unique current blockades.

### 4.2.2. Effects of amino acid substitutions on base discrimination at distal recognition sites

It is has been shown that it is possible to modulate the base recognition properties of recognition site $R_1$ by making amino acid substitutions at the central constriction. Subsequently, it was examined whether amino acid substitutions at the central constriction alter nucleobase recognition at distal recognition sites (e.g. $R_2$ and $R_3$) as well as at recognition site $R_1$. The pores that gave the weakest discrimination and the strongest discrimination at $R_1$, namely the NNG and NNY respectively, were initially probed using the set of fourteen oligonucleotides which mimic the sequential movement of a single adenine base through a poly(dC) background, that had been previously used to map the recognition profile of the β barrel of the NN pore (Stoddart, Heron et al. 2009) (Table 3.4 and Figure 3.6). The difference in residual current between each adenine-containing oligonucleotide and poly(dC) was plotted against the position of the adenine base (Figure 4.4, Table 4.2). While the pores exhibit different adenine discrimination properties at $R_1$ (the NNY pore shows strong recognition of adenines compared to the NNG pore) the $\Delta I_{RES}$ values at $R_2$ and $R_3$ for the NNG and NNY pores were similar to those seen with the NN pore.

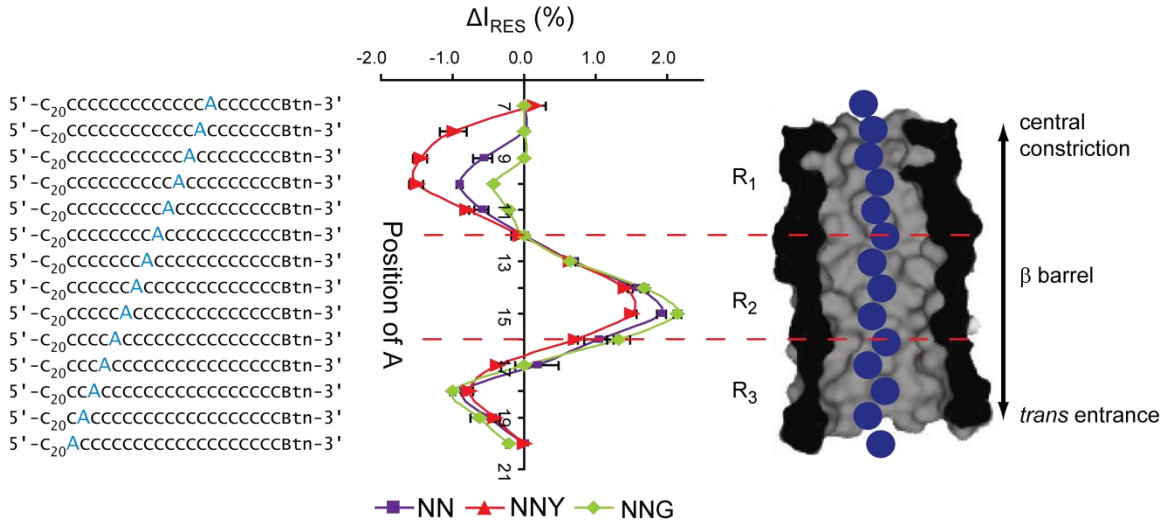



Figure 4.4. The effect of amino acid substitutions on adenine recognition at distal recognition points. The graph (middle) indicates the differences in residual current ($\Delta I_{RES}$ values) between blockades caused by a poly(dC) oligonucleotide containing a single adenine nucleobase (left) and blockades caused by poly(dC) for NN (purple), NNY (red) and NNG (green) αHL pores (Table 4.2). $R_1$, $R_2$ and $R_3$ represent the three proposed recognition sites in the αHL nanopore. Their probable locations are indicated on the cross-section of the β barrel domain of the αHL pore (right). Figure adapted from (Stoddart, Heron et al. 2010).

Subsequently, the recognition of all four DNA bases was probed at recognition sites $R_2$ and $R_3$ in the NNG and NNY pores using two sets of oligonucleotides, that had been previously used to probe the $R_2$ and $R_3$ recognition sites of WT and NN pores for four base discrimination (see Chapter 3, section 3.2.4). Both sets comprised four poly(dC) oligonucleotides that contained single G, A, T or C substitutions; the first set had the base substitution at position 14 and was designed to probe $R_2$ and the second set had the base substitution at position 18 and was designed to probe $R_3$. As already observed, base discrimination at $R_1$ is remarkably altered in the two pores, with the NNG showing weak discrimination and the NNY showing strong discrimination (Figure 4.5A and Table 4.3). However, both the NNG and the NNY pores show similar

base discrimination properties at recognition site $R_2$ (Figure 4.5B and Table 4.3) and at recognition site $R_3$ (Figure 4.5C and Table 4.3) to each other and are similar to the NN pore (Figure 3.8).

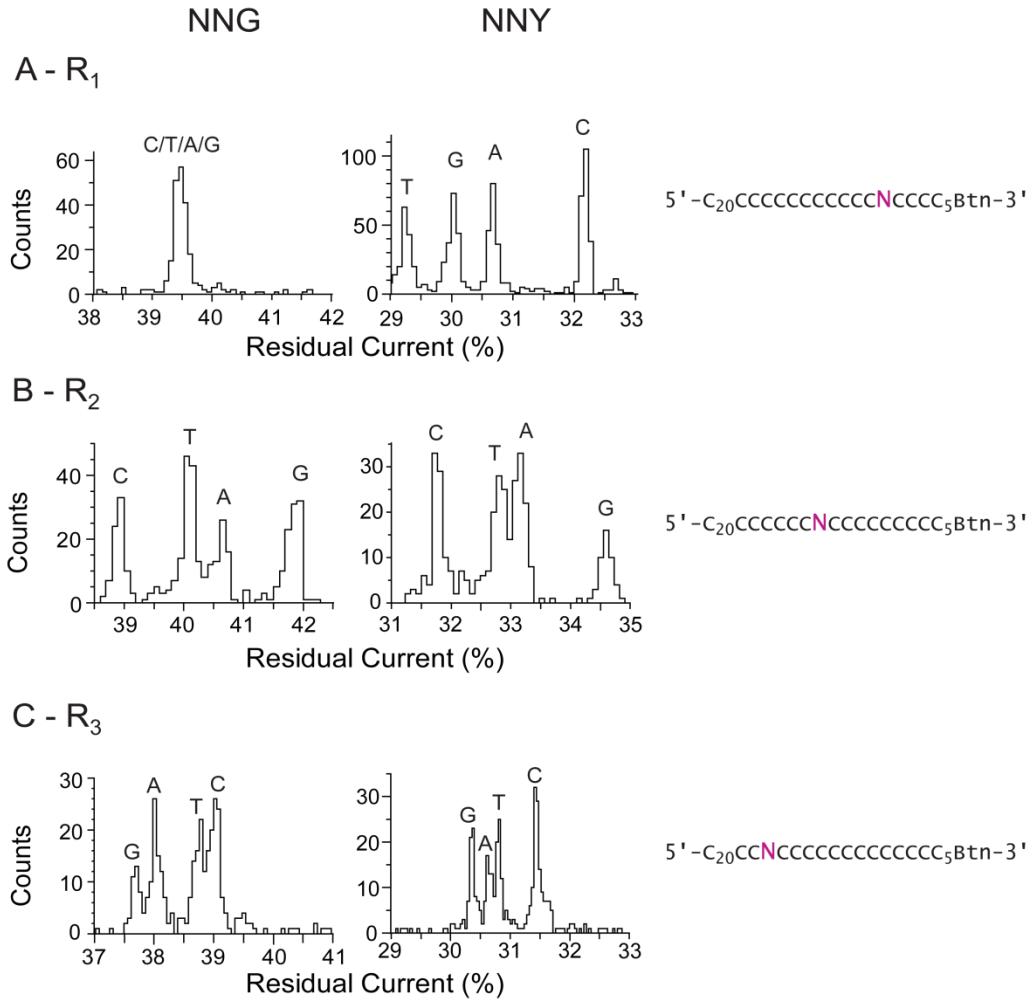



Figure 4.5. Probing the recognition sites of NNG and NNY pores for four-base discrimination. Histograms of the residual current levels for NNG (left) and NNY (right) pores are shown. Three sets of four poly(dC) oligonucleotides were used, with each set containing either a single G, A, T, or C nucleobase at a specific position; position 9 (panel – A), position 14 (panel B) and position 18 (panel – C). The results displayed in the figure are from a typical experiment (n ≥ 3).

Table 4.2. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single adenine nucleobase. The position of the adenine base in the $A_1$ oligonucleotide (nucleobases 7-20) is numbered relative to the 3' biotin tag. Data are shown for NNY and NNG pores. The $I_O$ and $I_{RES}$ values are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an $A_1$ oligonucleotides and poly(dC) ($I_{RES}^{A1\ oligo}$ - $I_{RES}^{pC}$). The errors given are standard deviations. Table adapted from (Stoddart, Heron et al. 2010).

| Position of adenine | NNY | | | Position of adenine | NNG | | |
|---|---|---|---|---|---|---|---|
| | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| 7 | 31.8 ± 1.0 | 31.7 ± 0.9 | +0.1 ± 0.2 | 7 | 40.0 ± 0.3 | 40.0 ± 0.3 | 0.0 ± 0.0 |
| 8 | 30.7 ± 1.4 | 31.7 ± 1.2 | −0.9 ± 0.2 | 8 | 39.6 ± 0.9 | 39.6 ± 0.9 | 0.0 ± 0.0 |
| 9 | 29.3 ± 1.2 | 30.7 ± 1.3 | −1.5 ± 0.1 | 9 | 38.8 ± 1.5 | 38.8 ± 1.5 | 0.0 ± 0.0 |
| 10 | 29.7 ± 1.1 | 31.2 ± 1.2 | −1.5 ± 0.1 | 10 | 37.9 ± 1.8 | 38.3 ± 1.8 | −0.4 ± 0.0 |
| 11 | 31.0 ± 0.7 | 31.8 ± 0.7 | −0.8 ± 0.0 | 11 | 37.1 ± 1.6 | 37.3 ± 1.6 | −0.2 ± 0.0 |
| 12 | 32.0 ± 1.1 | 32.0 ± 1.1 | −0.1 ± 0.1 | 12 | 37.5 ± 1.3 | 37.5 ± 1.3 | 0.0 ± 0.0 |
| 13 | 32.2 ± 0.7 | 31.6 ± 0.8 | +0.6 ± 0.1 | 13 | 38.1 ± 1.4 | 37.4 ± 1.3 | +0.6 ± 0.1 |
| 14 | 32.7 ± 0.4 | 31.3 ± 0.4 | +1.4 ± 0.0 | 14 | 40.5 ± 0.4 | 38.9 ± 0.4 | +1.7 ± 0.1 |
| 15 | 32.8 ± 0.6 | 31.3 ± 0.6 | +1.5 ± 0.1 | 15 | 41.2 ± 0.6 | 39.0 ± 0.6 | +2.1 ± 0.1 |
| 16 | 31.6 ± 0.6 | 30.9 ± 0.6 | +0.7 ± 0.0 | 16 | 39.1 ± 1.4 | 37.8 ± 1.3 | +1.3 ± 0.2 |
| 17 | 30.5 ± 0.5 | 30.9 ± 0.5 | −0.4 ± 0.1 | 17 | 38.5 ± 1.4 | 38.5 ± 1.4 | 0.0 ± 0.0 |
| 18 | 30.3 ± 0.5 | 31.0 ± 0.5 | −0.8 ± 0.0 | 18 | 37.7 ± 0.4 | 38.7 ± 0.3 | −1.0 ± 0.0 |
| 19 | 31.1 ± 0.1 | 31.6 ± 0.1 | −0.4 ± 0.0 | 19 | 37.1 ± 1.1 | 37.8 ± 1.2 | −0.6 ± 0.1 |
| 20 | 31.5 ± 0.0 | 31.5 ± 0.0 | 0.0 ± 0.0 | 20 | 37.5 ± 1.2 | 37.8 ± 1.2 | −0.2 ± 0.0 |

These data suggest that the effect of making amino acid substitutions at the central constriction on nucleobase discrimination is limited largely to $R_1$, i.e. it is possible to target a specific recognition site by site-directed mutagenesis, without grossly affecting other recognition sites.

Table 4.3. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single substituted base at position 9 (corresponding to $R_1$), 14 (corresponding to $R_2$) or 18 (corresponding to $R_3$) blocking NNG or NNY pores. $I_{RES}$, $\Delta I_{RES}^{OVERALL}$ and δ values are shown for n ≥ 3 experiments. $\Delta I_{RES}^{OVERALL}$ is defined as the difference in residual current between the two most widely dispersed current peaks. The δ value is the product of the successive differences between the peaks. If any two peaks overlap, then the δ value is equal to zero. Mean values (± s.d.) are given.

| | Residual Current (%) | | | | Residual Current Difference | |
|---|---|---|---|---|---|---|
| NNG | $C_9$ | $T_9$ | $A_9$ | $G_9$ | $\Delta I_{RES}^{OVERALL}$ (%) | Δ |
| | 38.2 ± 1.8 | 38.2 ± 1.8 | 38.2 ± 1.8 | 38.2 ± 1.8 | 0.0 ± 0.0 | 0.00 ± 0.00 |
| | $C_{14}$ | $T_{14}$ | $A_{14}$ | $G_{14}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 38.8 ± 0.3 | 40.0 ± 0.4 | 40.5 ± 0.4 | 41.7 ± 0.4 | 2.9 ± 0.0 | 0.71 ± 0.05 |
| | $C_{18}$ | $T_{18}$ | $A_{18}$ | $G_{18}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 38.7 ± 0.3 | 38.4 ± 0.3 | 37.7 ± 0.4 | 37.4 ± 0.4 | 1.3 ± 0.0 | 0.07 ± 0.01 |
| NNY | $C_9$ | $T_9$ | $A_9$ | $G_9$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 31.4 ± 1.3 | 28.6 ± 1.2 | 30.0 ± 1.2 | 29.4 ± 0.9 | 2.9 ± 0.2 | 0.70 ± 0.11 |
| | $C_{14}$ | $T_{14}$ | $A_{14}$ | $G_{14}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 31.3 ± 0.4 | 32.4 ± 0.4 | 32.7 ± 0.4 | 34.2 ± 0.4 | 2.9 ± 0.0 | 0.53 ± 0.04 |
| | $C_{18}$ | $T_{18}$ | $A_{18}$ | $G_{18}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 31.0 ± 0.4 | 30.4 ± 0.4 | 30.2 ± 0.4 | 29.9 ± 0.4 | 1.1 ± 0.1 | 0.03 ± 0.01 |

4.2.3. Utilising multiple base-recognition sites in the αHL nanopore.

It might be advantageous to use more than one recognition site for DNA sequence determination. Consider a nanopore with two recognition sites, $R_1$ and $R_2$, each capable of recognizing all four bases (Figure 4.6). If the first site, $R_1$ produces a large dispersion of current levels for the four bases and the second site, $R_2$ produces a more modest dispersion, 16 current levels, one for each of the 16 possible base combinations, would be observed as DNA molecules are translocated through the nanopore. Therefore, at any particular moment, the current signal would offer information about two positions in the sequence, rather than just one, and each base would be read twice, first at $R_1$ and secondly at $R_2$.

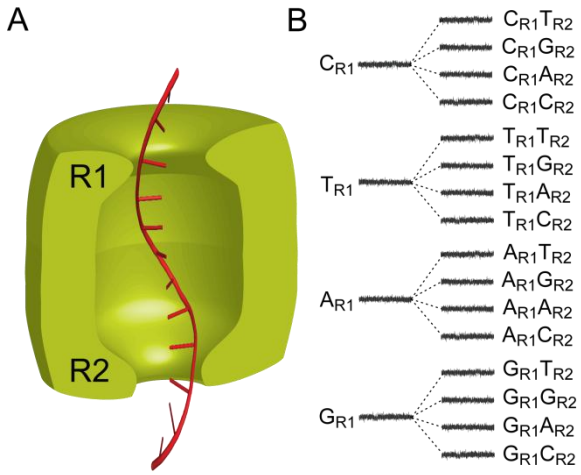



Figure 4.6. Two heads are better than one. A) A hypothetical nanopore sensor (green) with two recognition sites, $R_1$ and $R_2$, which could extract more sequence information from a DNA strand (red) than a device with a single recognition site. B) To illustrate the idea, assume that the four bases of DNA at recognition site $R_1$ produce 4 distinct current levels (widely dispersed as shown). Each of the levels is split into 4 additional levels (with a lesser dispersion, for the purpose of illustration) by the second recognition site $R_2$, yielding 16 current levels in total and providing redundant information about the DNA sequence. Figure adapted from (Stoddart, Maglia et al. 2010).

This built-in proof-reading mechanism could improve the overall quality of sequencing. In addition to this, multiple recognition sites could also provide more secure sequencing of homopolymeric stretches within a DNA strand. Consider a functioning nanopore sequencing device with a single recognition point; as a heteropolymeric ssDNA molecule traverses the pore, the current will be altered (to one of four levels, each representing; C, T, A or G) depending upon the identity of the base at the recognition site. When presented with a homopolymeric stretch, the current level will be constant for the duration of the time that the homopolymeric stretch resides at the recognition site (unless there is a current transition associated with base translocation, which is independent of the sequence). Given that DNA translocation is a stochastic process (Kasianowicz, Brandin et al. 1996; Meller, Nivon et al. 2000), merely using the dwell time of a given current level may not be sufficient to determine the length of the homopolymer (due to the stick-slip mechanism of DNA translocation). However, if a nanopore were to have a second, distal recognition site, then whilst the homopolymeric stretch translocates through one recognition site (e.g. $R_1$), the current level will continue to change as a heteropolymeric sequence passes through the other site (e.g. $R_2$). Such a two-site system would allow for effective "counting" of the length of a homopolymeric stretch, which would not be possible in a one-site system and is also a limitation of other next generation sequencing technologies (e.g. pyrosequencing). Based on these arguments, a third recognition site would, intuitively, further improve matters. However, in this case, the number of possible base combinations would increase from 16 to 64. Even if these levels could be dispersed across the entire current spectrum of the αHL

pore (from almost open to almost closed), it is unlikely that the 64 levels could be separated owing to the electrical noise in the system, even under the low bandwidth conditions used here. Under working conditions, where enzyme-mediated threading of the DNA may be operating at ~100 nt $s^{-1}$, it is likely that higher acquisition rates would be required and this would consequentially increase the noise, exacerbating the problem. Therefore, it seems likely that a two recognition-site sensor is optimal.

To determine if additional information might be gained from the use of a multiple recognition-site sensor, the NNY nanopore was examined. The NNY pore is capable of recognizing each of the four bases at $R_1$, $R_2$ and $R_3$ (Figure 4.5-right). However, the discrimination is much stronger at sites $R_1$ ($\delta$ = 0.70 ± 0.11) and $R_2$ ($\delta$ = 0.53 ± 0.04) compared to the weaker recognition site $R_3$ ($\delta$ = 0.03 ± 0.01). It was decided that it would be best to probe the two strongest recognition sites (in the case of the NNY; $R_1$ and $R_2$), rather than the weaker site ($R_3$), as it was presumed that, in practice, this would give the best chance of observing all 16 possible current levels. A library containing 16 oligonucleotides comprising poly(dC) with substitutions at position 9 (to probe $R_1$) and position 14 (to probe $R_2$) was acquired. The sequence of a given oligonucleotide was designated: $X_9X_{14}$, where X represents a defined base; G, A, T or C, and 9 and 14 gives the position of the base (relative to the biotin tag).

First, it was tested whether the identity of the base at position 14 ($R_2$) affected base recognition at position 9 ($R_1$). NNY pores were separately probed with four sets of four oligonucleotides: $N_9C_{14}$, $N_9A_{14}$, $N_9T_{14}$ and $N_9G_{14}$ (where N = G, A, T or C, Figure 4.7A-D, respectively and Table 4.4). Despite the variation of

the base at position 14, the distribution of the current levels for each set of four oligonucleotides is remarkably similar (Table 4.4). This suggests that recognition at $R_1$ (i.e. the order and dispersion of the peaks in the histograms) is only weakly influenced by the base occupying $R_2$.

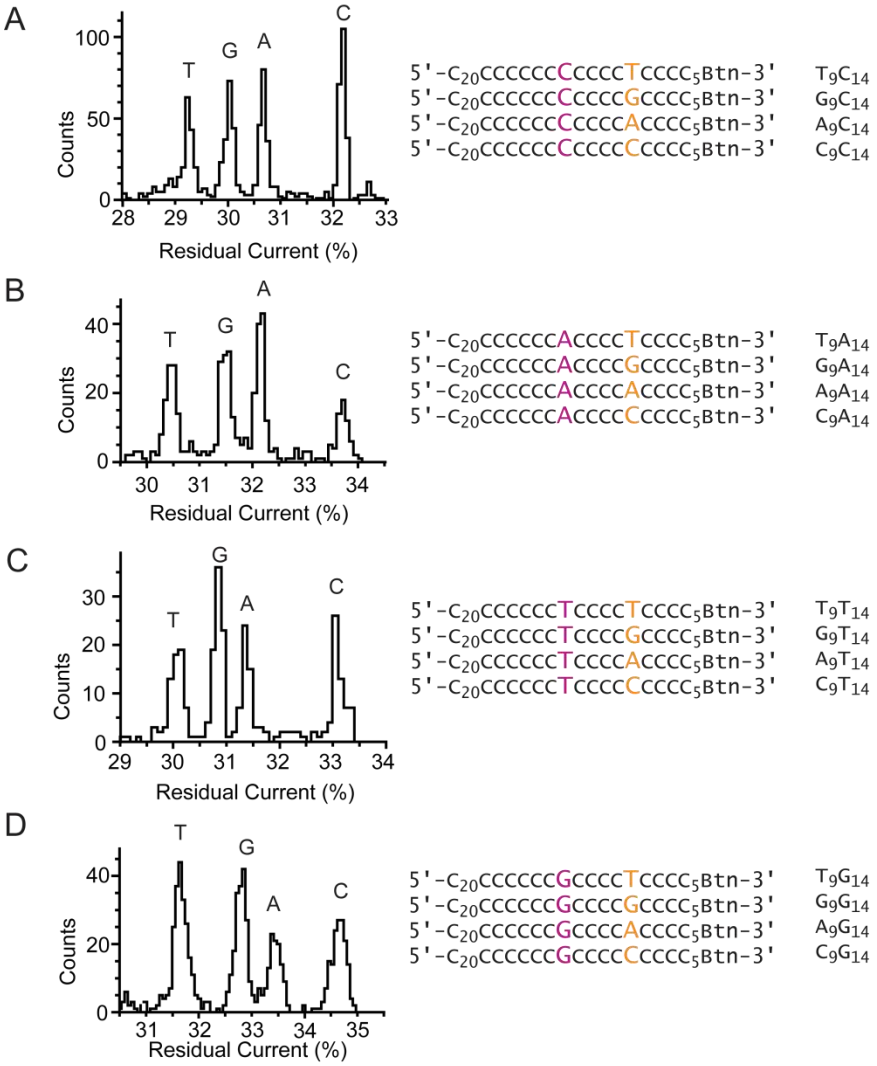



Figure 4.7. Four-base discrimination at $R_1$ when bases are altered at $R_2$, by the NNY pore. Histograms are shown (left) for a set of 4 oligonucleotides (right). Within each oligonucleotide set, the bases are altered at $R_1$ (orange) and inbetween each set, the bases are altered at $R_2$ (purple). The sequence of each oligonucleotide is abbreviated as $X_9X_{14}$. Each experiment (A-D) was conducted at least three times, and the results displayed in the figure are from a single experiment. Gaussian fits were performed for each peak in the histograms and the mean value of the residual current ($I_{RES}$) for each oligonucleotide recorded. Figure adapted from (Stoddart, Maglia et al. 2010).

Table 4.4. Mean residual current differences between poly(dC) oligonucleotides that contain nucleobase substitutions at position 9 (to probe $R_1$) and/or position 14 (to probe $R_2$). The positions of the substitutions are relative to the 3' biotin tag. The sequence of each oligonucleotide is abbreviated as $X_9X_{14}$. The mean values (± s.d.) are given for at least three experiments. $\Delta I_{RES}$ is directly measured as the difference between the residual current levels of two specified oligonucleotides (Figures 4.5 and 4.7). In the uppermost row (oligonucleotide set $C_9N_{14}$), $\Delta I_{RES} = I_{RES}^{C9N14} - I_{RES}^{C9C14}$ (Figure 4.5). In the other four rows, $\Delta I_{RES} = I_{RES}^{N9X14} - I_{RES}^{C9X14}$ (Figure 4.7A-D). Table adapted from (Stoddart, Maglia et al. 2010).

| Oligonucleotide set | Residual Current difference (%) | | | | |
|---|---|---|---|---|---|
| $C_9N_{14}$ (Figure 4.5B-right) | $\Delta I_{RES}^{C9A14-C9C14}$ | $\Delta I_{RES}^{C9T14-C9C14}$ | $\Delta I_{RES}^{C9G14-C9C14}$ | $\Delta I_{RES}^{OVERALL}$ | δ |
| | +1.4 ± 0.0 | +1.1 ± 0.0 | +2.9 ± 0.1 | 2.9 ± 0.1 | 0.53 ± 0.04 |
| $N_9C_{14}$ (Figure 4.7A) | $\Delta I_{RES}^{A9C14-C9C14}$ | $\Delta I_{RES}^{T9C14-C9C14}$ | $\Delta I_{RES}^{G9C14-C9C14}$ | $\Delta I_{RES}^{OVERALL}$ | δ |
| | −1.5 ± 0.1 | −2.9 ± 0.1 | −2.0 ± 0.1 | 2.9 ± 0.1 | 0.70 ± 0.11 |
| $N_9A_{14}$ (Figure 4.7B) | $\Delta I_{RES}^{A9A14-C9A14}$ | $\Delta I_{RES}^{T9A14-C9A14}$ | $\Delta I_{RES}^{G9A14-C9A14}$ | $\Delta I_{RES}^{OVERALL}$ | δ |
| | −1.5 ± 0.1 | −3.2 ± 0.1 | −2.1 ± 0.1 | 3.2 ± 0.1 | 0.97 ± 0.12 |
| $N_9T_{14}$ (Figure 4.7C) | $\Delta I_{RES}^{A9T14-C9T14}$ | $\Delta I_{RES}^{T9T14-C9T14}$ | $\Delta I_{RES}^{G9T14-C9T14}$ | $\Delta I_{RES}^{OVERALL}$ | δ |
| | −1.7 ± 0.1 | −2.9 ± 0.1 | −2.1 ± 0.1 | 2.9 ± 0.1 | 0.64 ± 0.14 |
| $N_9G_{14}$ (Figure 4.7D) | $\Delta I_{RES}^{A9G14-C9G14}$ | $\Delta I_{RES}^{T9G14-C9G14}$ | $\Delta I_{RES}^{G9G14-C9G14}$ | $\Delta I_{RES}^{OVERALL}$ | δ |
| | −1.2 ± 0.1 | −2.9 ± 0.2 | −1.8 ± 0.2 | 2.9 ± 0.2 | 0.82 ± 0.15 |

Assuming that the effects of each base at each recognition site on the change in current level are additive, and by using the directly determined $\Delta I_{RES}$ values in Table 4.4, it was possible to predict the distribution of $\Delta I_{RES}$ values for each of the sixteen sequences $N_9N_{14}$, relative to poly(dC), which is set as zero (Figure 4.8 and Table 4.5). For example, consider the sequence $T_9A_{14}$. It is possible to predict the unknown $\Delta I_{RES}^{T9A14-C9C14}$ (these two sequences were not compared directly, Figures 4.5 and 4.7) by using experimentally determined $\Delta I_{RES}$ values (Table 4.4): $\Delta I_{RES}^{T9A14-C9A14}$ = -3.2 ± 0.1% (Figure 4.7 and Table 4.4, 3rd row) and $\Delta I_{RES}^{C9A14-C9C14}$ = +1.4 ± 0.0% (Figure 4.5B, right and Table 4.4, 1st

row). By adding these values together, we find $\Delta I_{RES}^{T9A14-C9C14}$ = -1.8 ± 0.1%. The use of $I_{RES}$ values, rather than experimental $\Delta I_{RES}$ values, leads to unacceptable errors in predicted $\Delta I_{RES}$ values (see Chapter 3, section 3.2.1 for discussion on the potential sources of variations in $I_{RES}$).

All remaining $\Delta I_{RES}$ values were predicted in the same way (Table 4.5) and are shown in Figure 4.8 as dashed grey lines. When all 16 sequences ($N_9N_{14}$, Table 4.6) were used simultaneously to probe NNY pores, the histogram of the residual current levels contained 11 resolvable sequence-specific peaks (Figure 4.8). The predicted $\Delta I_{RES}$ values match well with the measured $\Delta I_{RES}$ values, with the observed mean $\Delta I_{RES}$ values within the error of the predicted values (Table 4.6). It is proposed that current flow is restricted at $R_1$ and $R_2$, and that the effects of the two recognition points are approximately additive, when $\Delta I_{RES}$ values are small, like the effect of two small resistances in series, in an electrical circuit.

Table 4.5. Predicted and measured residual current differences ($\Delta I_{RES}$) between poly(dC) oligonucleotides that contain single nucleobase substitutions at position 9 (to probe $R_1$) and/or position 14 (to probe $R_2$) and poly(dC). The positions of the substitutions are relative to the 3' biotin tag. The sequence of each oligonucleotide is abbreviated as $X_9X_{14}$ (see appendix, Table A.1). A measured $\Delta I_{RES}$ value is the mean of seven experiments (± s.d.). The predicted $\Delta I_{RES}$ values are calculated from $\Delta I_{RES}^{X9X14-C9C14} = \Delta I_{RES}^{X9X14-C9X14} + \Delta I_{RES}^{C9X14-C9C14}$ (using the values in Table 4.4) and the quoted error is the square root of the sum of the squares of the standard deviations for the two $\Delta I_{RES}$ values used to determine the predicted value. Table adapted from (Stoddart, Maglia et al. 2010).

| | $\Delta I_{RES}^{Oligo-C9C14}$ (%) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Oligo | $T_9C_{14}$ | $G_9C_{14}$ | $T_9T_{14}$ | $T_9A_{14}$ | $A_9C_{14}$ | $G_9T_{14}$ | $G_9A_{14}$ | $A_9T_{14}$ | $A_9A_{14}$ | $T_9G_{14}$ | $C_9T_{14}$ | $G_9G_{14}$ | $C_9A_{14}$ | $A_9G_{14}$ | $C_9G_{14}$ |
| Predicted | −2.8 ± 0.2 | −2.0 ± 0.1 | −1.8 ± 0.1 | −1.8 ± 0.1 | −1.4 ± 0.1 | −1.1 ± 0.1 | −0.7 ± 0.1 | −0.6 ± 0.1 | −0.1 ± 0.1 | +0.1 ± 0.2 | +1.1 ± 0.1 | +1.1 ± 0.2 | +1.4 ± 0.1 | +1.7 ± 0.1 | +2.9 ± 0.1 |
| Measured | −2.8 ± 0.0 | −1.9 ± 0.1 | −1.8 ± 0.1 | −1.8 ± 0.1 | −1.4 ± 0.1 | −1.0 ± 0.0 | −0.6 ± 0.1 | −0.6 ± 0.1 | 0.0 ± 0.0 | 0.0 ± 0.0 | +1.1 ± 0.1 | +1.1 ± 0.1 | +1.5 ± 0.1 | +1.8 ± 0.1 | +2.9 ± 0.1 |

## 4.3. Conclusions

It is likely that it will be necessary to augment the in-built recognition element of the WT αHL pore if it is going to be used in a nanopore DNA sequencing device. Systematic, targeted mutagenesis at the central constriction has shown that it is possible to strengthen (NNY) or weaken (NNG) base recognition at site $R_1$ specifically, without major disruptions in base identification at sites $R_2$ or $R_3$. Amino acids that provide an energy barrier to ion flow (e.g. bulky or hydrophobic residues) strengthen base identification, while amino acids that lower the energy barrier, weaken base identification. Furthermore, it was observed that amino acid side chains with similar chemical properties tend to discriminate bases in similar patterns. These data could be used to inform the redesign of recognition sites.

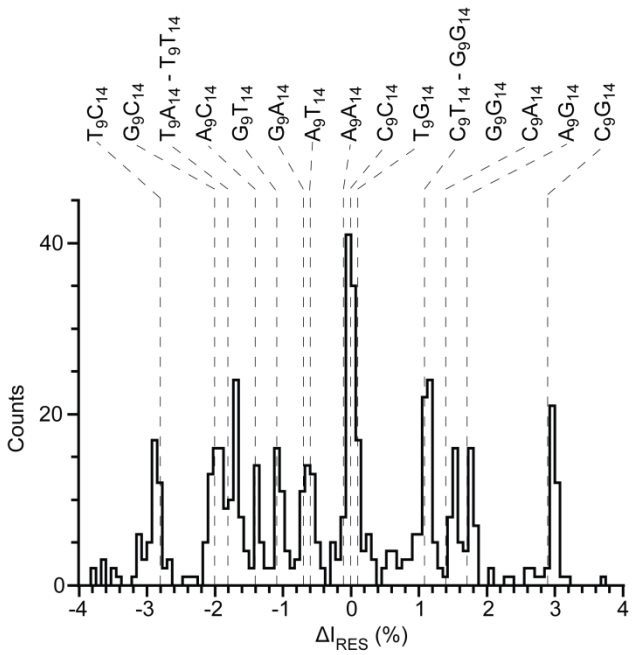



Figure 4.8. Predicted and experimental residual current level differences ($\Delta I_{RES}$) observed when NNY pores are interrogated with oligonucleotides which simultaneously probe $R_1$ and $R_2$. NNY pores were probed with 16 oligonucleotides, with the sequence 5'–CCCCCCCCCCCCCCCCCCCCCCCCCNCCCCNCCCCCCCCCBtn–3', where N is A, T, G or C ($N_9N_{14}$). A histogram displaying the residual current level differences (Table 4.6) for blockades by the various oligonucleotides, relative to the mean blockade produced by poly(dC) is shown. The current level for poly(dC) is set as zero. Blockades which have a residual current level lower than poly(dC) have negative $\Delta I_{RES}$ values and blockades which have higher residual current levels than poly(dC) have positive $\Delta I_{RES}$ values. The grey dashed lines show the predicted residual current levels, based on the $\Delta I_{RES}$ data displayed in Table 4.4 (see the text). The predicted and measured $\Delta I_{RES}$ values are displayed in Table 4.5. Figure adapted from (Stoddart, Maglia et al. 2010).

In an attempt to determine if additional sequence information could be gained from a single nanopore which contained multiple recognition sites, the NNY pore, which displays good base discrimination at both $R_1$ and $R_2$, was probed with 16 ssDNA molecules that contained base substitutions at both $R_1$ and $R_2$ recognition sites. It was possible to resolve 11 of a possible 16 current levels and although this system is imperfect, it does yield additional, redundant

information about each base, which would provide more secure base identification than a single recognition site. The NNY pore also displays base identification at $R_3$, although it is unlikely that the 64 possible current-levels that would be produced from three recognition sites could be resolved (indeed, it is difficult to resolve just 16 different current levels). Therefore, it seems likely that a two recognition-site sensor is optimal, and as such it may be necessary to remove a superfluous recognition site, for example $R_3$.

Table 4.6. Residual currents levels ($I_{RES}$) and differences ($\Delta I_{RES}$) for poly(dC) and oligonucleotides that contain nucleobase substitutions at positions 9 (to probe $R_1$) and 14 (to probe $R_2$). $I_{RES}$, and $\Delta I_{RES}$ values are shown for seven experiments. $\Delta I_{RES}$ is defined as the difference in residual current between a poly(dC) oligonucleotide containing substituted nucleobases and poly(dC) itself ($\Delta I_{RES} = I_{RES}^{oligo} - I_{RES}^{C9C14}$). Mean $I_{RES}$ and $\Delta I_{RES}$ values (± s.d.) are given. Table adapted from (Stoddart, Maglia et al. 2010).

| | | $T_9C_{14}$ | $G_9C_{14}$ | $T_9T_{14}$ | $T_9A_{14}$ | $A_9C_{14}$ | $G_9T_{14}$ | $G_9A_{14}$ | $A_9T_{14}$ | $A_9A_{14}$ | $C_9C_{14}$ | $T_9G_{14}$ | $C_9T_{14}$ | $G_9G_{14}$ | $C_9A_{14}$ | $A_9G_{14}$ | $C_9G_{14}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $I_{RES}$ (%) | 28.3 | 29.1 | 29.4 | 29.4 | 29.7 | 30.1 | 30.6 | 30.6 | 31.1 | 31.1 | 31.1 | 32.2 | 32.2 | 32.6 | 32.9 | 34.1 |
| | $\Delta I_{RES}$ (%) | −2.8 | −2.0 | −1.7 | −1.7 | −1.4 | −1.0 | −0.5 | −0.5 | 0.0 | 0.0 | 0.0 | +1.1 | +1.1 | +1.5 | +1.8 | +2.9 |
| 2 | $I_{RES}$ (%) | 26.3 | 27.1 | 27.4 | 27.4 | 27.7 | 28.0 | 28.4 | 28.4 | 29.1 | 29.1 | 29.1 | 30.0 | 30.0 | 30.3 | 30.8 | 31.9 |
| | $\Delta I_{RES}$ (%) | −2.8 | −2.0 | −1.7 | −1.7 | −1.4 | −1.1 | −0.7 | −0.7 | 0.0 | 0.0 | 0.0 | +0.9 | +0.9 | +1.2 | +1.7 | +2.8 |
| 3 | $I_{RES}$ (%) | 27.7 | 28.5 | 28.8 | 28.8 | 29.2 | 29.5 | 30.0 | 30.0 | 30.4 | 30.4 | 30.4 | 31.6 | 31.6 | 31.9 | 32.3 | 33.4 |
| | $\Delta I_{RES}$ (%) | −2.7 | −1.9 | −1.7 | −1.7 | −1.2 | −1.0 | −0.5 | −0.5 | 0.0 | 0.0 | 0.0 | +1.1 | +1.1 | +1.4 | +1.8 | +2.9 |
| 4 | $I_{RES}$ (%) | 28.6 | 29.5 | 29.5 | 29.5 | 29.9 | 30.2 | 30.7 | 30.7 | 31.3 | 31.3 | 31.3 | 32.4 | 32.4 | 32.8 | 33.2 | 34.2 |
| | $\Delta I_{RES}$ (%) | -2.8 | -1.8 | -1.8 | -1.8 | -1.5 | -1.1 | -0.7 | -0.7 | 0.0 | 0.0 | 0.0 | +1.1 | +1.1 | +1.5 | +1.9 | +2.9 |
| 5 | $I_{RES}$ (%) | 29.1 | 30.1 | 30.1 | 30.1 | 30.6 | 31.0 | 31.4 | 31.4 | 32.0 | 32.0 | 32.0 | 33.2 | 33.2 | 33.6 | 33.6 | 35.1 |
| | $\Delta I_{RES}$ (%) | -2.9 | -2.0 | -2.0 | -2.0 | -1.4 | -1.1 | -0.6 | -0.6 | 0.0 | 0.0 | 0.0 | +1.2 | +1.2 | +1.5 | +1.5 | +3.1 |
| 6 | $I_{RES}$ (%) | 28.8 | 29.8 | 29.8 | 29.8 | 30.4 | 30.8 | 31.2 | 31.2 | 31.8 | 31.8 | 31.8 | 32.9 | 32.9 | 33.4 | 33.6 | 34.8 |
| | $\Delta I_{RES}$ (%) | -2.9 | -1.9 | -1.9 | -1.9 | -1.4 | -1.0 | -0.6 | -0.6 | 0.0 | 0.0 | 0.0 | +1.2 | 1.2 | +1.6 | +1.8 | +3.0 |
| 7[1] | $I_{RES}$ (%) | 28.9 | 29.8 | 30.0 | 30.0 | 30.4 | 30.7 | 31.1 | 31.1 | 31.7 | 31.7 | 31.7 | 32.9 | 32.9 | 33.3 | 33.5 | 34.7 |
| | $\Delta I_{RES}$ (%) | -2.9 | -2.0 | -1.7 | -1.7 | -1.4 | -1.1 | -0.6 | -0.6 | 0.0 | 0.0 | 0.0 | +1.1 | +1.1 | +1.5 | +1.8 | +3.0 |
| Mean | $I_{RES}$ ± s.d. (%) | 28.2 ± 1.0 | 29.1 ± 1.0 | 29.3 ± 1.0 | 29.3 ± 1.0 | 29.7 ± 1.0 | 30.0 ± 1.0 | 30.5 ± 1.0 | 30.5 ± 1.0 | 31.1 ± 1.0 | 31.1 ± 1.0 | 31.1 ± 1.0 | 32.2 ± 1.1 | 32.2 ± 1.1 | 32.5 ± 1.1 | 32.8 ± 1.0 | 34.0 ± 1.0 |
| | $\Delta I_{RES}$ ± s.d. (%) | −2.8 ± 0.1 | −1.9 ± 0.1 | −1.8 ± 0.1 | −1.8 ± 0.1 | −1.4 ± 0.1 | −1.0 ± 0.0 | −0.6 ± 0.1 | −0.6 ± 0.1 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 | +1.1 ± 0.1 | +1.1 ± 0.1 | +1.5 ± 0.1 | +1.8 ± 0.1 | +2.9 ± 0.1 |

[1] The histogram corresponding to this experiment is displayed in Figure 4.8

References


Akeson, M., D. Branton, et al. (1999). "Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid and polyuridylic acid as homopolymers or as segments within single RNA molecules." Biophys.J. **77**: 3227-3233.

Allers, J. and Y. Shamoo (2001). "Structure-based analysis of protein-RNA interactions using the program ENTANGLE." J Mol Biol **311**(1): 75-86.

Antson, A. A., E. J. Dodson, et al. (1999). "Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA." Nature **401**(6750): 235-242.

Ashkenasy, N., J. Sánchez-Quesada, et al. (2005). "Recognizing a single base in an individual DNA strand: a step toward nanopore DNA sequencing." Angew.Chem.Int.Ed.Engl. **44**: 1401-1404.

Bochkarev, A., R. A. Pfuetzner, et al. (1997). "Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA." Nature **385**(6612): 176-181.

Bonthuis, D. J., J. Zhang, et al. (2006). "Self-energy-limited ion transport in subnanometer channels." Phys Rev Lett **97**(12): 128104.

Cheley, S., L.-Q. Gu, et al. (2002). "Stochastic sensing of nanomolar inositol 1,4,5-trisphosphate with an engineered pore." Chem.Biol. **9**: 829-838.

Derrington, I. M., T. Z. Butler, et al. (2010). "Nanopore DNA sequencing with MspA." Proc Natl Acad Sci U S A **107**(37): 16060-16065.

Gallivan, J. P. and D. A. Dougherty (1999). "Cation-pi interactions in structural biology." Proc Natl Acad Sci U S A **96**(17): 9459-9464.

Horvath, M. P., V. L. Schweiker, et al. (1998). "Crystal structure of the Oxytricha nova telomere end binding protein complexed with single strand DNA." Cell **95**(7): 963-974.

Hu, G., P. D. Gershon, et al. (1999). "mRNA cap recognition: dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains." Proc.Natl.Acad.Sci.USA **96**: 7149-7154.

Jones, S., D. T. Daley, et al. (2001). "Protein-RNA interactions: a structural analysis." Nucleic Acids Res **29**(4): 943-954.

Kasianowicz, J. J., E. Brandin, et al. (1996). "Characterization of individual polynucleotide molecules using a membrane channel." Proc.Natl.Acad.Sci.USA **93**: 13770-13773.

Lei, M., E. R. Podell, et al. (2004). "Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection." Nat Struct Mol Biol **11**(12): 1223-1229.

Liu, Z., I. Luyten, et al. (2001). "Structural basis for recognition of the intron branch site RNA by splicing factor 1." Science **294**(5544): 1098-1102.

Maglia, G., M. Rincon Restrepo, et al. (2008). "Enhanced translocation of single DNA molecules through α-hemolysin nanopores by manipulation of internal charge." Proc Natl Acad Sci U S A **105**: 19720-19725.

Meller, A., L. Nivon, et al. (2000). "Rapid nanopore discrimination between single polynucleotide molecules." Proc.Natl.Acad.Sci.USA **97**: 1079-1084.

Parsegian, A. (1969). "Energy of an ion crossing a low dielectric membrane: solutions to four relevant electrostatic problems." Nature **221**(5183): 844-846.

Purnell, R. F. and J. J. Schmidt (2009). "Discrimination of single base substitutions in a DNA strand immobilized in a biological nanopore." ACS Nano **3**: 2533-2538.

Raghunathan, S., A. G. Kozlov, et al. (2000). "Structure of the DNA binding domain of E. coli SSB bound to ssDNA." Nat Struct Biol **7**(8): 648-652.

Sanchez-Quesada, J., A. Saghatelian, et al. (2004). "Single DNA rotaxanes of a transmembrane pore protein." Angew Chem Int Ed Engl **43**(23): 3063-3067.

Song, L., M. R. Hobaugh, et al. (1996). "Structure of staphylococcal a-hemolysin, a heptameric transmembrane pore." Science **274**: 1859-1865.

Stoddart, D., A. Heron, et al. (2009). "Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore." Proc. Natl. Acad. Sci. USA **106**: 7702-7707.

Stoddart, D., A. J. Heron, et al. (2010). "Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore." Nano Lett **10**(9): 3633-3637.

Stoddart, D., G. Maglia, et al. (2010). "Multiple base-recognition sites in a biological nanopore– two heads are better than one." Angew Chem Int Ed **49**: 556-559.

Wang, X., J. McLachlan, et al. (2002). "Modular recognition of RNA by a human pumilio-homology domain." Cell **110**(4): 501-512.

Zamyatnin, A. A. (1972). "Protein volume in solution." Prog.Biophys.Mol.Biol. **24**: 107-123.

# CHAPTER 5: TRUNCATION OF THE RECOGNITION ELEMENT IN α-HEMOLYSIN PROTEIN NANOPORES BY DELETING PORTIONS OF THE TRANSMEMBRANE β BARREL

Abstract

The 5 nm-long transmembrane β barrel of the αHL pore comprises the base recognition element of the protein and the WT pore contains three broad recognition sites. In an attempt to simplify the current signals that would be observed as ssDNA molecules are sequenced, as they translocate through the αHL pore, the transmembrane β barrel region has been truncated, to reduce the length of the recognition element, and the number of recognition sites. The αHL pore is able to withstand substantial truncations in the β barrel region and still form single-channels in artificial bilayer systems. Truncations to the β barrel, reduces the number of bases that are recognized, presumably because fewer bases reside within the shortened domain. Additional site-directed mutagenesis is used to remove one of the recognition sites.

## 5.1. Introduction

Studies in previous chapters have shown that the αHL pore is capable of discriminating between ssDNA molecules which contain single nucleobase

substitutions, when immobilized within the pore through a DNA-btn•streptavidin plug. Base identification seems to occur in the transmembrane β barrel and the WT αHL pore has three broad sites capable of base recognition, termed $R_1$, $R_2$ and $R_3$. The central constriction of the protein has been identified as being involved in base recognition at $R_1$ and extensive mutagenesis at position 113 (which comprises the central constriction, along with residues 111 and 147) showed that different amino acids provide different base recognition properties. Bulky amino acids that were capable of forming multiple types of interactions with bases, such as Tyr, generally improved the base identification at $R_1$, whereas small amino acids that provided limited interactions with bases, such as Gly, weakened recognition. As well as providing strong recognition at $R_1$ the NNY pore also gave strong recognition at $R_2$ and it was shown that, in principle, it should be possible to obtain improved sequence information from a pore with multiple recognition sites; however, it was suspected that pores with more than two recognition sites would elicit current signals that would be too complex to be reliably interpreted. Therefore it is proposed that it would be advantageous to remove a recognition site, e.g. $R_3$, and this chapter details the attempts that were made to reduce the complexity of the recognition element of the αHL pore. By using site-directed and deletion mutagenesis, it has been possible to effectively remove one of the recognition sites.

## 5.2. Results and discussion

### 5.2.1. Site-directed mutagenesis at the *trans* entrance of the αHL pore

It was assumed that recognition site $R_3$ was located at the *trans* entrance of the pore. Work in earlier chapters indicated that site-directed mutagenesis at the central constriction could be a powerful tool in the redesign of recognition site $R_1$, so it was proposed that similar mutagenesis near the *trans* entrance may be useful in an attempt to augment $R_3$. The amino acid sequence at the *trans* entrance of the pore is glycine rich (Figure 5.1). Previous studies indicated that glycines provided weak base identification (at $R_1$ at least) and therefore, as attempts were focused on further weakening $R_3$, the glycine rich region was not targeted for mutation. However, it was decided that the introduction of extra glycines, and removal of bulky amino acids could help weaken $R_3$, as similar mutations at position 113 had weakened recognition site $R_1$.

Leucine is a bulky amino acid that, when located at position 113, provided good base discrimination at $R_1$ (Chapter 4, section 4.2.1, Figures 4.2 and 4.3). In WT αHL pores, the β barrel contains a leucine at position 135, which is located near the *trans* entrance of the pore, and it was proposed that this amino acid could be involved in base discrimination at $R_3$. Therefore, in order to weaken $R_3$, an L135G mutation was made in the NN gene (NN/L135G – mutant generated and protein prepared by E. Mikhailova), and the mutant pore was examined for its ability to discriminate single adenines in an otherwise poly(dC) sequence, a method which had been used previously to map the recognition properties of other αHL mutant proteins (Chapter 3, section 3.2.3 and Chapter 4, section

4.2.2). However, in the case of the NN/L135G protein, the attempt to weaken $R_3$ was not successful (Figure 5.2). On the other hand, the $R_2$ site does appear to have been weakened, suggesting that the L135 residue is a constituent of recognition site $R_2$.

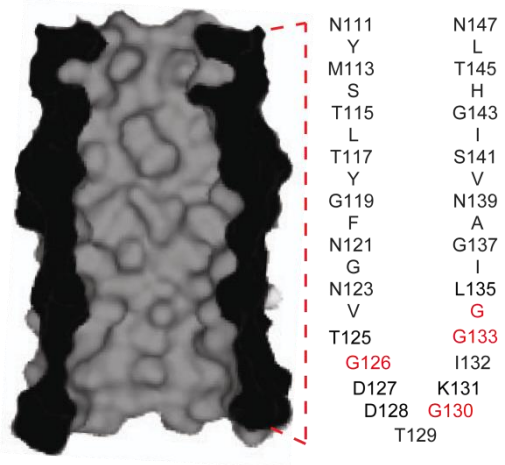



Figure 5.1. Amino acid sequence of the β barrel of the NN pore. The recognition element of the αHL pore is located in the transmembrane domain which comprises a 14 strand, antiparallel β barrel (grey cut-through). Each of the seven protomers contribute a pair of adjacent β strands (separated by a turn sequence; amino acids 126-132) to the barrel. The lower part of the β barrel is glycine rich (indicated in red), and the turn sequence contains three charged amino acid residues; D127, D128 and K131.

From the experiments shown in Figure 5.2, it appears that recognition site $R_3$ is located below L135, so the next mutant that was made in an attempt to weaken $R_3$ was at position T125, which is located just above the turn sequence at the *trans* entrance (Figure 5.1). As before, a glycine substitution was made, to generate the NN/T125G mutant (mutant generated and protein prepared by E. Mikhailova).

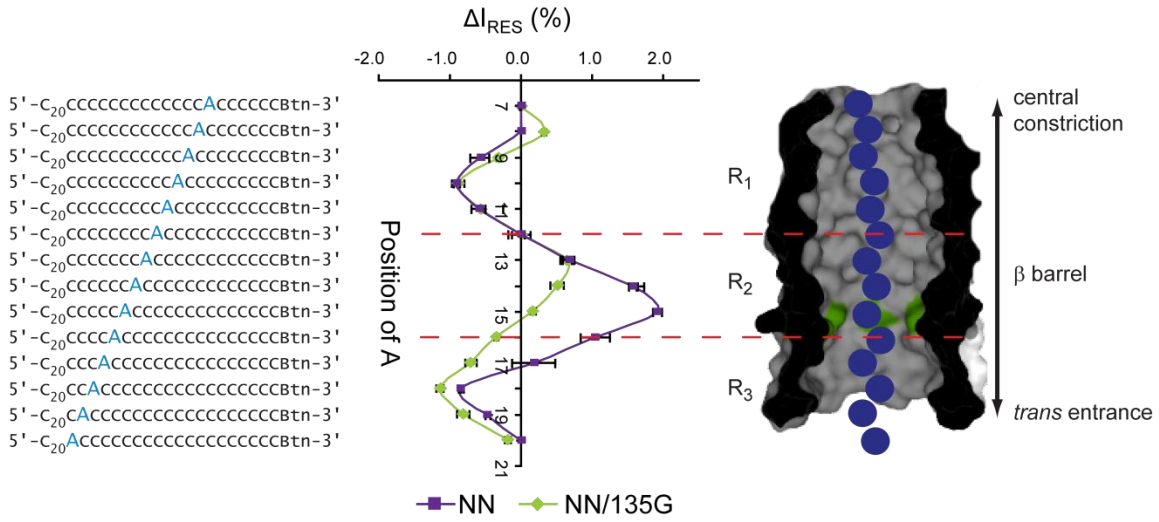



Figure 5.2. The effect of the L135G mutation on adenine recognition in the β barrel. The graph (middle) indicates the differences in residual current ($\Delta I_{RES}$ values) between blockades caused by a poly(dC) oligonucleotide containing a single adenine nucleobase (left) and blockades caused by poly(dC) for NN (purple) and NN/L135G (green) αHL pores (Table 5.1). $R_1$, $R_2$ and $R_3$ represent the three proposed recognition sites in the αHL nanopore. Their probable locations, along with the position of the L135 amino acid (green), are indicated on the cross-section of the β barrel domain of the αHL pore (right).

In single-channel recording (SCR) experiments, the NN/T125G pore did not maintain a stable open pore current level ($I_O$); rather, several sub-states were observed (Figure 5.3A). These sub-states were also present when the pore was blocked with a DNA-btn•streptavidin complex (Figure 5.3B), and this distorted the event-picking by the pClamp software (see section 2.3.4 for more information on event picking) and the histograms of the current blockades contained multiple populations, rather than just a single, dominant population, as was observed for other mutant pores (Figure 5.3CD). It is proposed that the introduction of another glycine into the already glycine rich region of the β barrel increased the flexibility of the protein to the point where it no longer forms a single, stable conformation, but instead forms multiple conformations each of which has a different

conductance; hence the changes in $I_O$ represent the conformational flexibility of the mutant protein.

Table 5.1. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single adenine nucleobase for the NN/L135G and E111N/K147N/D127N/D128N/K131Q (Q4N) pores. The position of the adenine base in the $A_1$ oligonucleotide (nucleobases 7-20) is numbered relative to the 3' biotin tag. The $I_O$ and $I_{RES}$ values are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an $A_1$ oligonucleotides and poly(dC) ($I_{RES}^{A1\ oligo} - I_{RES}^{pC}$). The errors given are standard deviations.

| Position of adenine | NN/L135G | | | Position of adenine | Q4N | | |
|---|---|---|---|---|---|---|---|
| | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_{RES}^{A}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| 7 | 39.5 ± 0.9 | 39.5 ± 0.9 | 0.0 ± 0.0 | 7 | 36.9 ± 0.2 | 36.9 ± 0.2 | 0.0 ± 0.0 |
| 8 | 40.2 ± 2.8 | 39.9 ± 2.8 | +0.3 ± 0.0 | 8 | 39.7 ± 0.2 | 36.9 ± 0.2 | −0.2 ± 0.1 |
| 9 | 38.2 ± 1.6 | 38.5 ± 1.6 | −0.3 ± 0.0 | 9 | 35.9 ± 1.0 | 36.4 ± 1.1 | −0.5 ± 0.1 |
| 10 | 38.7 ± 0.7 | 39.6 ± 0.7 | −0.9 ± 0.1 | 10 | 35.6 ± 0.6 | 36.5 ± 0.7 | −0.9 ± 0.1 |
| 11 | 39.8 ± 2.4 | 40.3 ± 2.4 | −0.6 ± 0.0 | 11 | 36.1 ± 0.7 | 36.6 ± 0.7 | −0.4 ± 0.1 |
| 12 | 40.2 ± 2.3 | 40.3 ± 2.4 | −0.1 ± 0.1 | 12 | 36.8 ± 1.0 | 36.8 ± 1.0 | 0.0 ± 0.0 |
| 13 | 39.3 ± 1.2 | 38.6 ± 1.2 | +0.6 ± 0.1 | 13 | 37.5 ± 0.6 | 36.6 ± 0.5 | +0.9 ± 0.1 |
| 14 | 40.2 ± 1.1 | 39.7 ± 1.0 | +0.5 ± 0.1 | 14 | 36.2 ± 2.7 | 35.0 ± 2.6 | +1.2 ± 0.2 |
| 15 | 39.8 ± 0.7 | 39.6 ± 0.7 | +0.2 ± 0.0 | 15 | 38.2 ± 0.9 | 37.2 ± 0.8 | +1.1 ± 0.1 |
| 16 | 37.7 ± 3.0 | 38.1 ± 3.1 | −0.3 ± 0.0 | 16 | 36.6 ± 0.2 | 36.3 ± 0.2 | +0.3 ± 0.3 |
| 17 | 39.5 ± 2.6 | 40.2 ± 2.6 | −0.7 ± 0.1 | 17 | 36.6 ± 1.1 | 37.1 ± 0.8 | −0.5 ± 0.3 |
| 18 | 37.5 ± 1.3 | 38.6 ± 1.3 | −1.1 ± 0.0 | 18 | 35.4 ± 1.0 | 36.3 ± 0.9 | −0.8 ± 0.1 |
| 19 | 37.8 ± 1.1 | 38.6 ± 1.2 | −0.7 ± 0.1 | 19 | 35.6 ± 0.3 | 36.1 ± 0.3 | −0.5 ± 0.0 |
| 20 | 37.9 ± 3.0 | 38.1 ± 3.0 | −0.2 ± 0.0 | 20 | 37.0 ± 0.6 | 37.1 ± 0.7 | −0.1 ± 0.1 |

In a further attempt to weaken $R_3$, amino acid substitutions were made in the turn sequence that separates the two β strands. This turn sequence contains charged amino acids (D127, D128 and K131 - Figure 5.1) and previous studies indicated that the neutralization of charged residues can have a marked effect on nucleobase recognition; for example, neutralizing the charged E111 and K147 residues at the central constriction by replacing them with neutral asparagines, produced a change in the base recognition properties at $R_1$ (Chapter 3).

In a separate project within the lab, a mutant had already been generated in which the charged residues at the *trans* entrance had been neutralized: E111N/K147N/D127N/D128N/K131Q (Q4N – E. Mikhailova, personal communication). It was hoped that this mutant would have altered base recognition at the *trans* entrance, compared to the NN, just as the NN had different base recognition at the central constriction, compared to the WT pore. The base recognition properties of the Q4N pore were tested as before for the NN/L135G mutant, and the results are shown in Figure 5.4 (and Table 5.1). However, it appears as though the mutations made in the turn sequence have little influence in the recognition of adenine bases at the *trans* entrance/$R_3$. A simplistic conclusion is that $R_3$ does not discriminate bases in the same way as $R_1$, and therefore a different strategy would be required to weaken this recognition site.

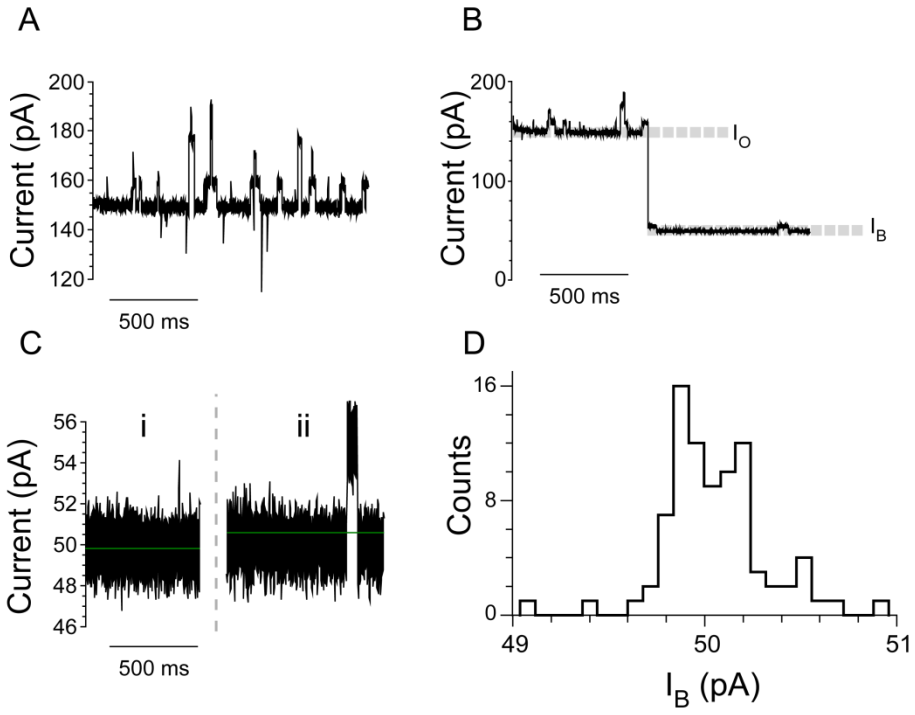



Figure 5.3. The effect of the T125G mutation on the current flow through NN αHL pores. A) Current recording of the open pore current level ($I_O$) of the NN/T125G pore. The open pore current level of the NN/T125G pore is not stable and varies over time, suggesting that the NN/T125G pore adopts more than one conformation. B) Current recording showing the stepwise change in current, as an NN/T125G pore becomes blocked with a DNA-btn•streptavidin complex. The variability in $I_O$ transmits to fluctuations in the blocked pore level ($I_B$) when the NN/T125G pore is blocked. C) Single-channel event search to find the mean $I_B$ value for two blockades (i and ii; black traces). NB both the blockades shown are from ssDNA molecules with the same sequence. The variations in $I_B$ when the pore is blocked, skews the mean value (green line) that is attributed to the block level (ii). In this case, the two blockades have similar baseline current values, but the attributed mean values are significantly different. D) Histogram of the mean $I_B$ values after a single-channel event search of many ssDNA blockades. Typically, when pores are probed with a single sequence of ssDNA, the histograms of the $I_B$ levels show one dominant population (Figure 2.10). However, due to variations in the current flow through the NN/T125G pore, the histograms of $I_B$ values show multiple populations of block pore levels.

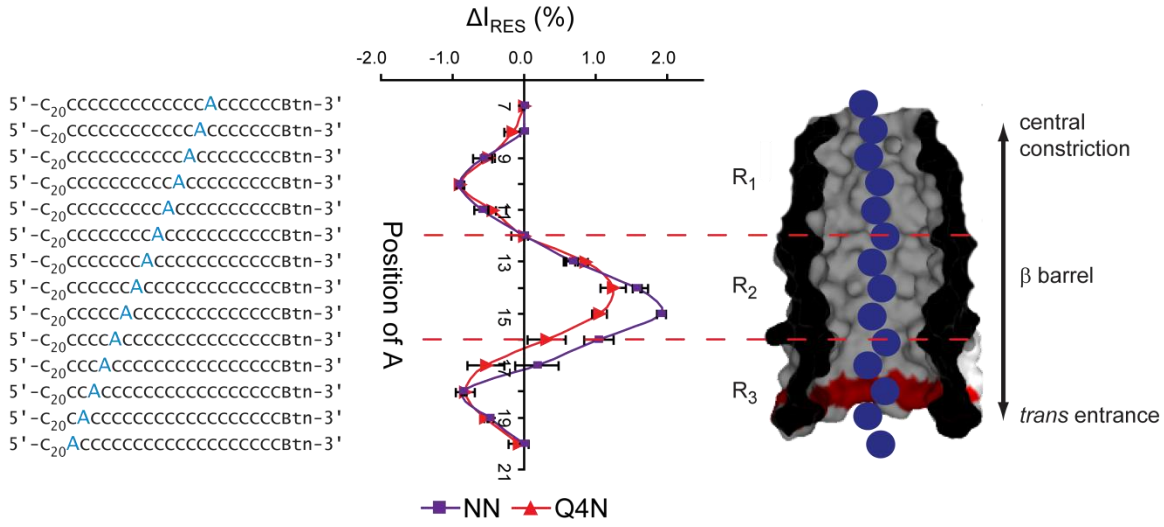



Figure 5.4. The effect of neutralizing the charged amino acids in the turn sequence, on adenine recognition in the β barrel. The graph (middle) indicates the differences in residual current ($\Delta I_{RES}$ values) between blockades caused by a poly(dC) oligonucleotide containing a single adenine nucleobase (left) and blockades caused by poly(dC) for NN (purple) and Q4N (red) αHL pores (Table 5.1). $R_1$, $R_2$ and $R_3$ represent the three proposed recognition sites in the αHL nanopore. Their probable locations along with the position of the charged amino acids in the turn (red) are indicated on the cross-section of the β barrel domain of the αHL pore (right).

### 5.2.2. Amino acid deletions at the *trans* entrance of the αHL pore to truncate the transmembrane β barrel

Given that site-directed mutagenesis had proved largely ineffective, a more drastic effort was made in an attempt to weaken recognition site $R_3$. This involved the sequential deletion of amino acids, near the *trans* entrance. The β barrel contains 14, antiparallel β strands, with each protomer, within the heptamer, contributing two adjacent strands, which are separated by a turn sequence (which approximately encapsulates amino acids Gly-126 through Ile-132 – Figure 5.5). The strands themselves largely consist of alternating hydrophilic and hydrophobic amino acids, with the hydrophilic amino acids pointing their side chains into the lumen of the pore, with the hydrophobic amino

acids pointing their side chains into the lipid bilayer/membrane (in Figure 5.5, the inward facing amino acids are numbered, and are interspersed with hydrophobic amino acids). To truncate the β barrel, rings of inward and outward facing residues from each of the two strands (per protomer) were sequentially deleted by PCR mutagenesis (leaving the turn sequence in tact), to form truncated barrel mutant (TBM) proteins. The rest of the protein sequence of the TBM proteins was left as wild-type, except that the charged residues at the central constriction are mutated to neutral asparagines (NN). In TBM Δ1, amino acids V124 and T125 were deleted from the "down" strand and amino acids G133 and G134 were deleted from the "up" strand. TBM Δ2, Δ3, Δ4 and Δ5 were formed by sequentially deleting the next pair of amino acids from each β strand (Figure 5.5A). Molecular dynamics simulations (carried out by J. Klingelhoefer) illustrate the proposed effect of the amino acid deletions; namely truncations in the transmembrane region (Figure 5.5B). Assuming that the mutant proteins adopt a WT-like fold (form heptamers, and the arrangement of the β barrel remains the same), it is estimated that with each sequential truncation, the protein becomes ~5 Å shorter in length (Table 5.2). It was felt that these proteins would be likely to form heptamers, as a previous study using atomic force microscopy (AFM), showed that αHL proteins in which almost the entire transmembrane β barrel had been removed also formed heptamers (Cheley, Malghani et al. 1997).

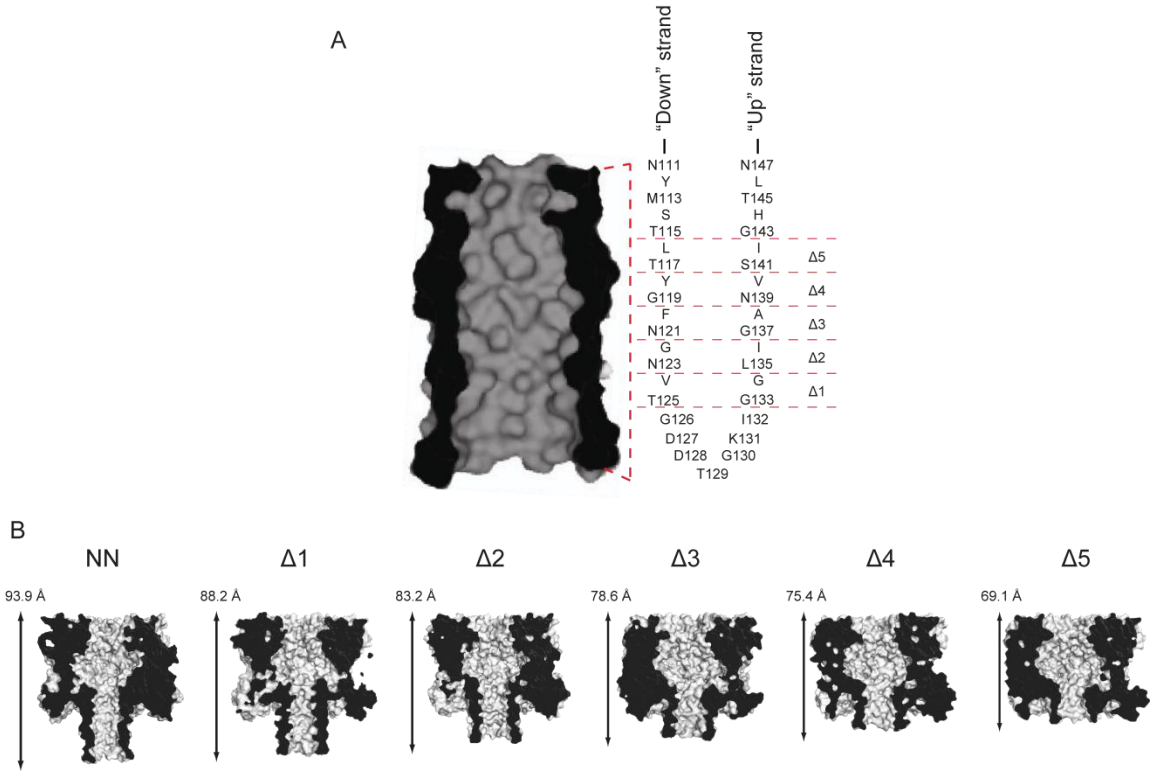



Figure 5.5. Truncations in the β barrel of the NN αHL pore. A) The transmembrane domain is a 14 stranded, antiparallel β barrel. Each of the seven protomers contribute a pair of adjacent β strands (separated by a turn sequence; amino acids 126-132) to the barrel. The amino acid sequence of the transmembrane portion of the β strands largely alternates between hydrophilic residues (which face inwards, towards the water-filled lumen of the pore) and hydrophobic residues (which face outwards towards the membrane). To truncate the β barrel, rings of inward and outward facing residues from each of the two strands (per protomer) were sequentially deleted by PCR mutagenesis (leaving the turn sequence intact), to form truncated barrel mutant (TBM) proteins. Note the E111N and K147N mutations at the top of the β barrel. B) Molecular dynamics simulations of the mutated proteins (performed by J. Klingelhoefer) suggest that the TBM proteins form truncated versions of the full length protein. Cut-through representations of each shortened mutant are shown. The length shown is the distance between the Cα atoms of amino acids N17 (located at the top of the cap domain), of the 3$^{rd}$ subunit, and T129 (located at the bottom of the transmembrane domain), of the 7$^{th}$ subunit (which is directly below N17 of the 3$^{rd}$ subunit).

The mutant genes were constructed in pT7 plasmids and were expressed by IVTT (see Materials and Methods, sections 2.1.1.2.2 and 2.1.2.1), in the absence and presence of rRBCm and were subjected to SDS-PAGE (Figure

5.6A and B, respectively). The mutant proteins, TBM Δ1, Δ2, Δ3, 3G (TBM Δ3/M113G – see below), Δ4 and Δ5 all formed oligomeric species in the absence of rRBCm. However, the full-length NN protein appears predominantly as monomer unless rRBCm are present. When comparing the oligomeric band of the full-length protein with those that are present for the TBM Δ1, Δ2, Δ3, 3G and Δ4 proteins, they all migrate at about the same speed. This would suggest that these proteins all have a similar architecture, and therefore it is likely that these truncated proteins form heptamers, in a similar fashion to the full-length protein.

On the other hand, the TBM Δ5 protein migrates anomalously to the other mutants; it forms oligomers that migrate slower through the gel, suggesting that it may form higher molecular weight structures, for example, octamers, nonamers, etc. Further, TBM Δ5 appears to form aggregates, which are not able to migrate into the gel (Figure 5.6). Notably, the truncated channel mutant (TCM) generated by Cheley et al. (Cheley, Malghani et al. 1997) which has a similar genetic architecture to the TBM Δ5 protein, was calculated to have a stoichiometry of 10.3 ± 1.8 subunits by gel filtration (based on the elution volume). However, AFM studies showed it to be heptameric. Given the similarity in the architecture of these two genes, it may be expected that the proteins would also have a similar, heptameric structure. Perhaps the TBM Δ5 species that appear to be higher in molecular weight could actually be heptamers. An alternative proposal was that if some WT-like heptamer was formed by TBM Δ5, it may be SDS-sensitive and therefore may break down on SDS-PAGE to produce the monomers that are observed on SDS-PAGE. The TCM heptamer produced by Cheley et al. displayed this type of SDS-sensitivity. It is notable that the other truncated

proteins also appear as both oligomers and as monomers in the gel shown in Figure 5.6A. Again, perhaps only a portion of the TBM protein forms SDS-stable oligomers, with the non SDS-stable oligomers degrading to monomer when the sample is dissolved in SDS-containing sample buffer and subjected to SDS-PAGE.

The full-length protein and the TBM proteins, Δ1-Δ5 (including 3G), were expressed and the products were subjected to limited proteolysis, by proteinase K (Sigma, UK – see Materials and Methods, section 2.1.4). The monomeric αHL polypeptide contains many sites that are susceptible to proteolysis (e.g. the N-terminus and the glycine-rich region), however, these sites become protease resistant upon oligomerization to the heptamer (Walker, Krishnasastry et al. 1992; Cheley, Malghani et al. 1997).

The digested fragments were heated in SDS-containing sample buffer to remove any oligomeric species before being subjected to SDS-PAGE (Figure 5.6C). When expressed in the absence of rRBCm, the NN protein is completely digested as expected (no intact monomer remains) and some digested fragments are visible in the gel. However, when expressed in the presence of rRBCm, the NN protein forms heptamers and no digested fragments are observed; only intact monomer is observed after heating. Although digested fragments are observed for the TBM Δ1 protein, some monomer remains intact and all the other TBM proteins also exhibit resistance to proteolysis.

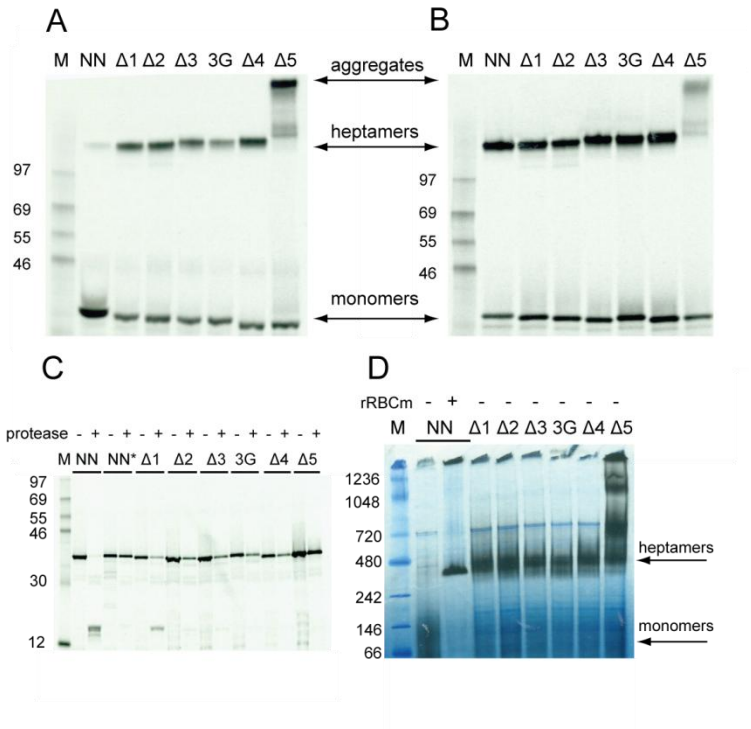



Figure 5.6. Analysis of the oligomeric state of TBM proteins. SDS-PAGE analysis after IVTT expression, in the absence (panel A) and presence (panel B) of rabbit red blood cell membranes (rRBCm), of the NN protein, the TBM Δ1-5 mutants and a TBM Δ3/M113G (3G) mutant. C) Limited proteolytic digests of freshly translated NN and TBM Δ1-Δ5 proteins (including 3G). The NN protein was translated in the two separate reactions; one in the absence of rRBCm (lanes marked NN), and the second, in the presence of rRBCm (lanes marked NN*). The apparent digestion of NN (-rRBCm) in the presence of proteinase K (+protease), suggests that the protein is in the monomeric state, and thus susceptible to proteolysis. The resistance of NN (NN* – +rRBCm) to proteinase K, suggests that the protein is in an oligomeric conformation. Each of the TBM proteins show at least some protease resistance, compared to NN (-rRBCm), suggesting that at least some of the protein is in an oligomeric conformation. D) NativePAGE analysis of the TBM proteins oligomeric structure. Radio-labelled NN proteins were translated in the presence and absence of rRBCm (where indicated) and TBM proteins were translated in the absence of rRBCm and analyzed by NativePAGE. The figure shown is of the aligned destained gel showing the pre-stained molecular markers, stapled to the developed film, showing the location of the radio-labelled protein.

The glycine-rich region is a target for proteolysis; the major cleavage sites in the glycine-rich region of full-length, monomeric protein are before I136 and Val140 (Walker, Krishnasastry et al. 1992). One or both of these sites are

absent in the TBM Δ2 (amino acids 122-125 and 133-136 deleted), Δ3 and 3G (amino acids 120-125 and 133-138 deleted), Δ4 (amino acids 118-125 and 133-140 deleted) and Δ5 (amino acids 116-125 and 133-142 deleted) proteins. The apparent protease resistance of TBM Δ2-Δ5 could therefore be attributed to the fact that major cleavage sites in the glycine-rich region are absent in these proteins. However, several other potential proteolysis sites remain (e.g. the N-terminus) but are not significantly targeted by proteinase K. This suggests that the truncated proteins form protease-resistant, oligomeric species (N.B. in the absence of rRBCm) with little monomer left over. The monomer that remains after oligomerization, may be subject to digestion, for example, as is observed for the TBM Δ1 protein.

To further probe the SDS-sensitivity of the truncated proteins, they were subjected to PAGE in the absence of SDS, using a NativePAGE gel system (Invitrogen, UK) (Figure 5.6D – and see Materials and Methods, section 2.1.5). Under native conditions, the full-length protein only formed oligomers in the presence of rRBCm, and the truncated mutants formed oligomers in the absence of membranes, as expected. The TBM Δ5 protein is observed as many bands suggesting that it forms a mixture of oligomeric species. One of the bands appears to migrate in a similar fashion to the oligomers of the other TBM proteins, suggesting that TBM Δ5 does form heptameric proteins. Note that a small amount of monomer is observed for each of the TBM proteins. This, along with the limited proteolysis data (Figure 5.6C), suggests that most of the protein oligomerizes and that most of the monomers observed in SDS-PAGE are largely due to SDS-mediated breakdown of the heptamer. This is most remarkable for

TBM Δ5, which only produces heptamer-like bands in the NativePAGE system. Furthermore, it is notable that the band for the full length heptamer is very tight and well defined, suggesting that just a single conformation exists. However, the equivalent band for the TBM proteins is much more diffuse, suggesting a range of different conformations. Perhaps the truncations made at the genetic level result in conformational instability or diversity at the protein level. This could also be the cause of their SDS instability.

As an initial assay of mutant pore function, the hemolytic activity of the proteins were tested (as described in Materials and Methods, section 2.1.3 and (Cheley, Malghani et al. 1997)). It was observed that despite the truncations in the β barrel region TBM Δ1 and Δ2 retain some hemolytic activity (Figure 5.7). However, subsequent truncations do appear to remove the ability of αHL pores to lyse rRBC, on these time scales at least. Hemolytic activity does not always correlate with the ability of mutants to form pores in artificial bilayers (E. Mikhailova – unpublished data), so despite the fact that TBM Δ3 – Δ5 lacked hemolytic activity, it was not certain that they would therefore not form single-channels, in the planar lipid bilayer system.

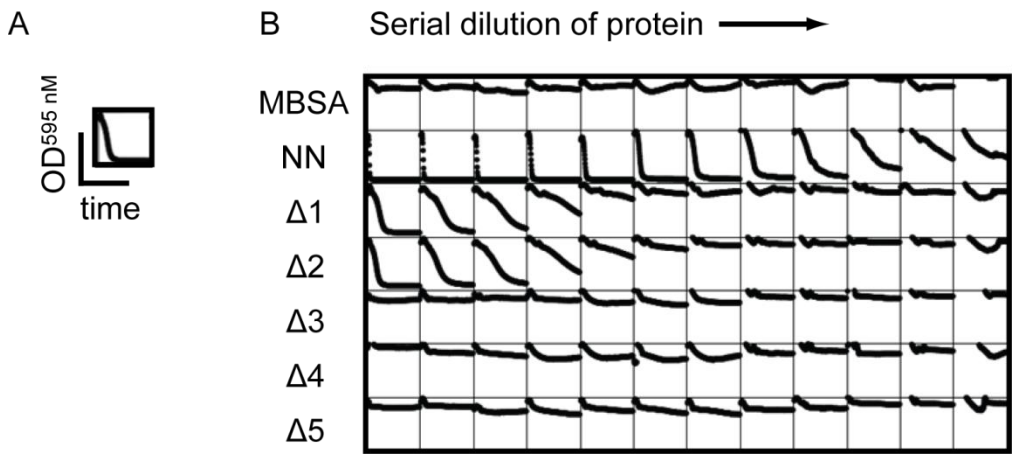



Figure 5.7. Hemolytic activity of the NN and TBM Δ1-5 αHL proteins. A) The decrease in light scattering ($OD^{595\ nM}$) over time represents lysing of rRBC due to the activity of pore forming αHL proteins. B) The hemolytic activity of αHL was assayed in 96-well plates, where 5 µL of freshly translated protein (from IVTT) was added to the first column of the plate and was subsequently serially diluted two-fold in a final volume of 100 µL, containing rRBC at ~0.5%. 5 µL of MBSA was added (top row) as a negative control; no decrease in light scattering is observed (cells are not lysed).

Oligomers of the TBM proteins were purified from SDS gels as described in Materials and Methods, section 2.1.2.1 and each of the samples was examined for "pore-forming" activity in single-channel recording experiments. In buffered electrolyte (1 M KCl, 25 mM Tris.HCl, 0.1 mM EDTA, pH8), and under applied potentials of +200 mV (*trans* compartment is positive, with respect to the grounded *cis* compartment), each of the TBM proteins displayed "pore-forming" activity, in DPhPC bilayers, with the exception of TBM Δ5 proteins, which displayed no pore-activity at all (Figure 5.8A). The average open pore current levels ($I_O$) of each of the proteins, under these conditions, is shown and compared to the full-length protein (the NN mutant) (Figure 5.8B and Table 5.2). There is no obvious correlation between pore length and $I_O$, although it could be that there are multiple factors influencing the conductance of the pores (e.g. pore

diameter and pore length). After the first and second truncations (TBM Δ1 and Δ2, respectively), the $I_O$ value increases with the decrease in pore length, and subsequent truncations (TBM Δ3 and Δ4) produce a decrease in the $I_O$ of the pores. The resistance of the pore is dependent upon pore length and pore diameter. It is proposed that an initial shortening the protein could reduce the resistance of the pore and hence increase the conductance. Further truncations may then lead to a change in the overall protein conformation, which result in a reduction in the pore conductance. For example, the dimensions of the β barrel may be altered. In the full-length pore, the β strands are tilted from the barrel axis by ~38° (Song, Hobaugh et al. 1996). In the TBM proteins, this tilt angle may be altered, such that the barrel untwists in an attempt to lengthen the transmembrane domain and overcome the genetic truncations, at the expense of pore diameter, leading to a reduction in the open pore conductance. However, this is speculative and may not be feasible, given that it may require substantial conformational alterations, such as changes in the shear number of the β barrel structure and in the H-bonding pattern.

It is also noted that the fluctuations in the open pore current signal (i.e. the "noise" of the signal) are different for each pore (Figure 5.8C). Single-channel recording experiments are very sensitive to noise from a number of both internal (Maglia, Heron et al. 2010) and external (such as vibrational and acoustic noise) sources. In order to be able to make any quantitative comparisons between different pores, it is necessary to ensure that the noise of the pore is being considered, and not the noise of the system. In order to do this, the "system noise" was subtracted from the "total noise" to correct for the noise of the system,

as follows. At an applied potential of zero mV (no current flowing through the pore), the fluctuations/noise in the current recording provide a practical estimate of the noise of the system. The noise is given as the standard deviation ($\sigma^{0\ mV}$) of the Gaussian fit to the current histogram. However, at an applied potential of +200 mV (*trans* is positive, relative to the grounded *cis*), a current will flow through the pore and the noise in the current recording contains contributions from both the noise of the pore and the noise of the system (total noise). As before, the noise is given as the standard deviation ($\sigma^{+200\ mV}$) of the Gaussian fit to the current histogram. To calculate the "pore noise", take the square root of ($\sigma^{0\ mV}$)$^2$ subtracted from ($\sigma^{+200\ mV}$)$^2$ (Equation 5.1). All the truncated proteins provide more noise to the current recording than the full length protein (Figure 5.8C and Table 5.2). Noise in the single-channel current recording is often associated with an increase in conformational movement within the protein (Bezrukov and Kasianowicz 1993; Korchev, Bashford et al. 1997) and mutations in the amino acid sequence can alter this noise (Cheley, Braha et al. 1999; Wai, Westermark et al. 2003; Chen, Khalid et al. 2008). The NativePAGE gel suggested that the TBM proteins were more conformationally diverse (Figure 5.6D - more diffuse bands on the gel, compared to the full length protein), and this therefore may cause the increase in noise in the current recordings that are observed. However, as with the open pore current levels of the TBM proteins, there appears to be no obvious correlation between pore length and the noise.

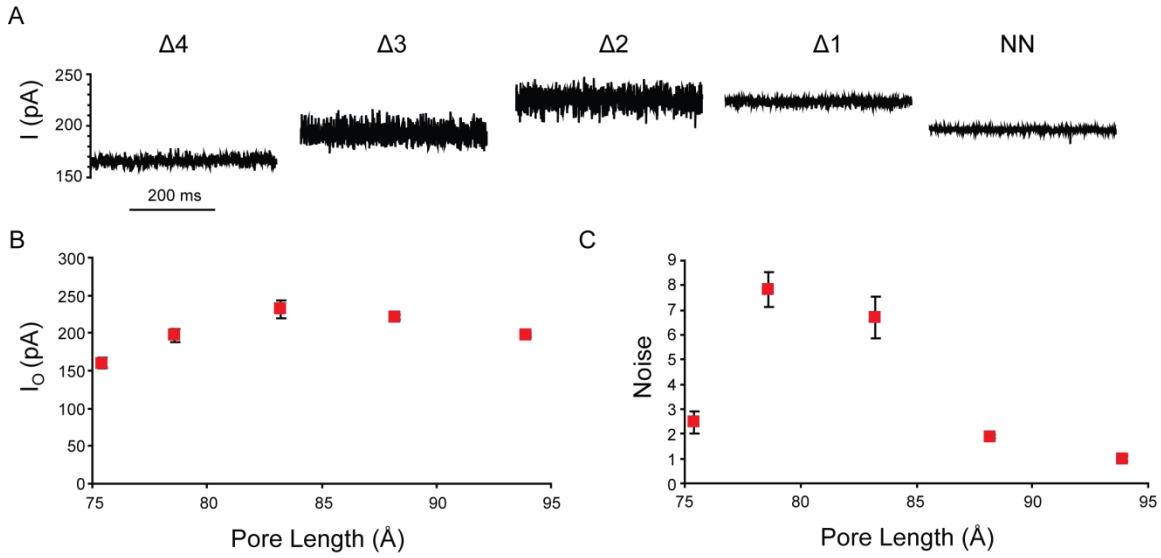



Figure 5.8. Single channel recordings (SCR) of the NN and TBM Δ1-Δ4 proteins. A) Purified proteins were inserted into DPhPC bilayers. Potentials of +200 mV were applied to the *trans* side of the bilayer (with the cis compartment connected to ground) and the current signal, which was sampled at 5 kHz and filtered at 1 kHz was recorded. Sections of typical current traces are shown for each protein, in the open pore state. B) The open pore current level ($I_O$) was measured for each of the pores (by Gaussian fitting to histograms), several times (n ≥ 3), and the mean $I_O$ level (± standard deviation) is plotted against the pore length (Figure 5.5 and Table 5.2). C) Each Gaussian fit that was made also is subject to error, or noise. The standard deviation (σ) of the Gaussian fit to the recorded current level was obtained at both +200 mV and at 0 mV and the noise was calculated (see Equation 5.1). The mean of the noise was determined for each mutant (Table 5.2 – n ≥ 3) and plotted against the pore length.

Equation 5.1. $$Noise = \sqrt{(\sigma^{+200mV})^2 - (\sigma^{+0mV})^2}$$

Table 5.2. The open pore current level and the noise for the TBM proteins. SCR recordings were acquired at +200 mV (*trans* compartment is positive, relative to the grounded *cis*) and at 0 mV (n ≥ 3 experiments). Gaussian fits were applied to the current signal and the peak of the histogram gives the current level for that pore. The average open pore level ($I_O$) at +200 mV (± s.d.) is given. The noise is calculated as described in Equation 5.1 and the average noise (± s.d.) is shown (n ≥ 3 experiments).

| Mutant | Pore Length (Å) | $I_O$ $^{+200\ mV}$ (pA) | Noise |
|---|---|---|---|
| NN | 93.9 | 197 ± 2 | 1.0 ± 0.1 |
| Δ1 | 88.2 | 221 ± 3 | 1.9 ± 0.1 |
| Δ2 | 83.2 | 231 ± 11 | 6.7 ± 0.9 |
| Δ3 | 78.6 | 197 ± 8 | 7.8 ± 0.7 |
| Δ4 | 75.4 | 160 ± 7 | 2.5 ± 0.4 |

Although pore activity was observed for TBM Δ4 proteins, the open pore current level was not stable. In DPhPC bilayers, pores formed from TBM Δ4 proteins were typically short-lived, exhibiting “burst-like” behaviour (Figure 5.9A). The fatty acid tails of DPhPC are 16 carbon atoms long, whereas the fatty acids tails of 1,2-dilauroyl-*sn*-glycero-3-phosphocholine (DLPC) lipids are only 12 carbon atoms long (Figure 5.9A). A 9:1 (DPhPC:DLPC w/w) mix of the two lipids was made, and it was observed that in these mixed bilayers the short life-time of the TBM Δ4 pores was ameliorated (Figure 5.9B). Typically, the life-time of the pores was improved from seconds to minutes, although quantitative “open-pore” probability studies were not performed. However, despite the moderate improvement in the open pore life-time in the mixed lipid bilayers, it was observed that the pores were “voltage-dependent” and would typically close at low positive potentials and at negative potentials (on the *trans* side) (Figure 5.9C). Once the pores closed under low potentials, the re-application of a high positive potential (on the *trans* side) was typically not sufficient to promote immediate pore re-opening; it was usual to wait for seconds or minutes for the pore to re-open (if it re-opened at all), even under high positive potential.

A close inspection of the predicted structure of the TBM pores (Figure 5.5B) shows that it is remarkable that the truncated barrel proteins form active pores at all. The properties of the TBM Δ4 protein are particularly intriguing. The short-lived nature of the open pore state in DPhPC bilayers suggests that whatever phenomena that are occurring to allow for the TBM proteins to insert are at the cusp of energetic favorability.

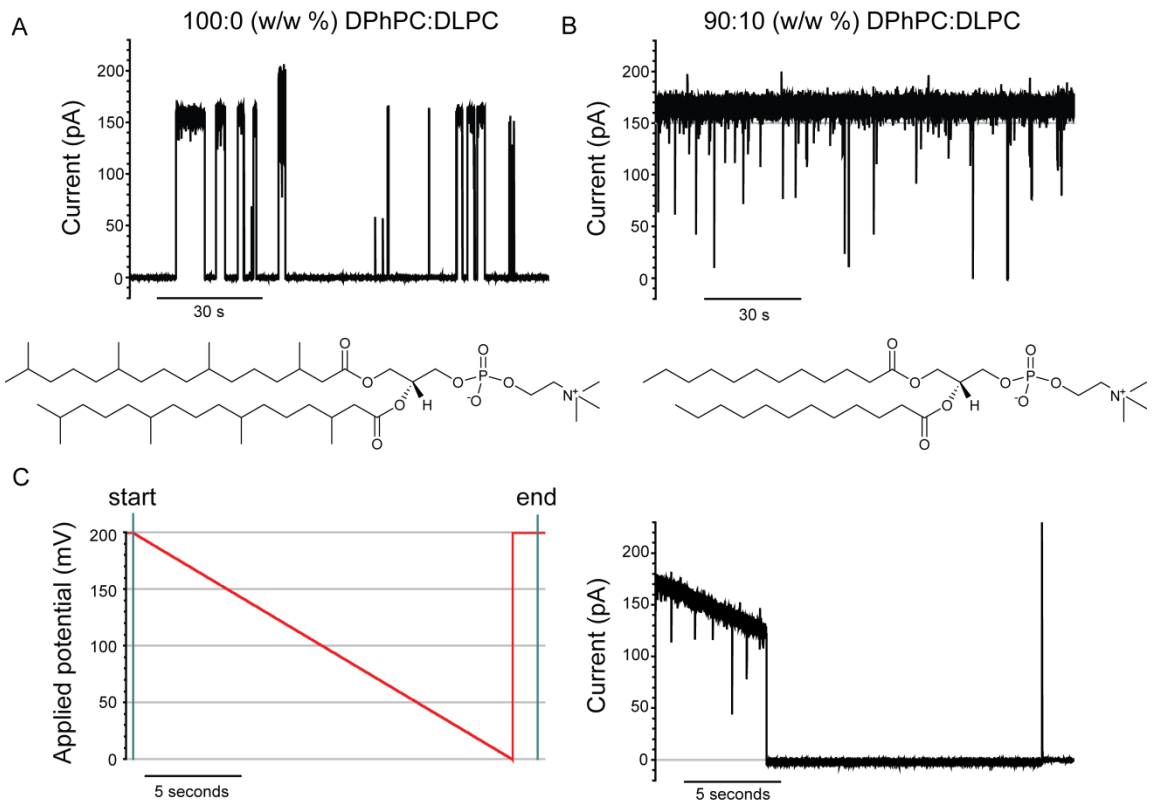



Figure 5.9. Current-voltage (I-V) characteristics of the TBM Δ4 protein pore. A) When obtaining SCR signals using bilayers that were composed of only DPhPC, TBM Δ4 proteins typically formed short-lived open pores (applied potential of +200 mV). B) The regular closing of the protein pores was ameliorated by supplementing the DPhPC bilayers with 10% (w/w) DLPC which has shorter fatty acid tails, compared to DPhPC. C) With a membrane mix of 9:1 DPhPC:DLPC (w/w), when voltage ramps were applied (where the voltage was reduced at a constant speed of 10 mV $s^{-1}$. NB membrane polarity is given as *trans*, relative to *cis*) it was observed that the pores would regularly close at potentials above +100 mV. This meant that it became impractical to carry out DNA discrimination experiments with the TBM Δ4 protein.

In order to accommodate the shortened proteins it is proposed that there must be some deformation of either the protein conformation, or the lipid bilayer. Alterations and perturbations in the physical properties of the lipid bilayer (e.g. bilayer thickness or curvature – see (Andersen and Koeppe 2007 and Phillips, Ursell et al. 2009) for review and references therein) are known to play an important role in the regulation of the activity of a wide variety of membrane

proteins. The fact that the open pore lifetime of TBM Δ4 is qualitatively improved by adding DLPC lipids to the bilayer, hints at some sort bilayer deformation, to allow protein insertion/opening, which is facilitated by the DLPC lipid (e.g. thinning or leaflet curvature). For example, disordering of the fatty acid chains could result in local “thinning” of the lipid bilayer, around the point of TBM protein insertion, and may be promoted in order to minimize the hydrophobic mismatch between the lipids and the shortened transmembrane domains of the TBM proteins (Killian 1998; Duque, Li et al. 2002). Any such deformation of the lipid bilayer may be aided by the use of shorter lipids (such as DLPC). Alternatively the protein conformation may alter at a tertiary or quaternary level. It is speculated that the β barrel domain may untwist and lengthen (at the expense of pore diameter) in order to overcome the genetic truncations. Furthermore, in the full-length protein, the 5-nm β barrel domain, is recessed within the cap domain somewhat (Figure 1.8B). In the truncated proteins, the truncated β barrel domain may be ejected fully from the cap and into the bilayer to allow for the β barrel to completely traverse the membrane. A more radical explanation is based on observations of other pore-forming proteins such as the oligomeric steptolysin-O (SLO). The SLO protein has been proposed to form large oligomers of varying stoichiometry, which penetrate the bilayer to form ring-like structures (Bhakdi, Tranum-Jensen et al. 1985; Sekiya, Satoh et al. 1993). In electron micrographs, often these ring-like structures were not fully closed and it was assumed that these represented partially formed pores and that the oligomers grew from a nucleation point. However, it has been proposed that these banana-shaped curves form membrane lesions and that the pore is formed on one side by the

protein, and by the free edge of lipid on the other side (Figure 5.10A and (Palmer, Harris et al. 1998)). Perhaps something similar is occurring with the TBM proteins; the TBM proteins associate with the lipids and the β barrel inserts part way into the lipid leaflet to form part of the conducting channel, and the remaining part of the conducting channel is made up of free lipids (Figure 5.10B). This would require an energetically expensive local curvature change for the lipids. It would be particularly costly for DPhPC lipids to bend like this due to dense packing in the chain region. Perhaps continual re-ordering of the lower leaflet to “close” and “open” is responsible for the observed “burst-like” behaviour of TBM Δ4. However, DLPC, which has a shorter, less densely packed chain region, would accept such membrane curvature readily. This could perhaps explain why DLPC promotes longer-lived TBM Δ4 pores.

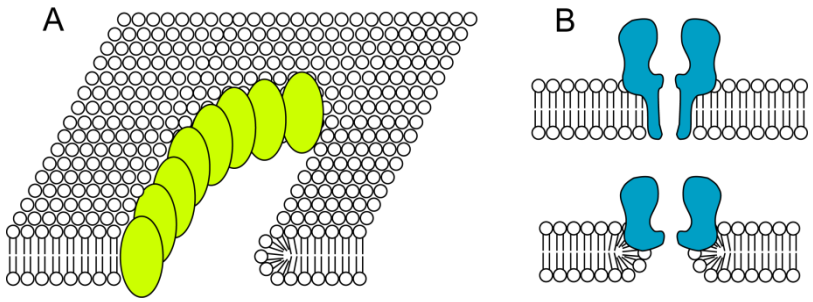



Figure 5.10. The possible conformations of TBM proteins in the lipid bilayer. A) The proposed protein-lipid channel conducting structure of SLO (green - adapted from (Palmer, Harris et al. 1998)). B) The full length αHL pore (upper - blue) forms an entirely proteinaceous channel, however TBM pores (lower - blue) may form part lipid, part protein conducting channels.

The voltage-dependency of the protein is also interesting and a similar type of behavior has been previously observed with another αHL mutant (Maglia, Heron et al. 2009). However, in this case the “7R” pore would immediately re-

open when positive potential was applied, unlike the TBM Δ4 protein, which did not always re-open. It may be speculated that defects in the bilayer, promoted at high potentials, may allow for insertion of the TBM Δ4 protein. If this were the case, it may be expected that the pore would open at high negative potentials as well; however, this was not conclusively observed. The fact that no pore activity was observed for TBM Δ5 proteins suggests that the hydrophobic mismatch between the bilayer and shortened transmembrane domain was just too large and it was energetically unlikely to that the TBM Δ5 could punch holes in the membrane.

### 5.2.3. Defining the recognition element within the truncated barrel mutant (TBM) pores

Subsequently, the β barrel region of the TBM Δ1, Δ2 and Δ3 proteins was examined for the ability to discriminate single adenine bases, in an otherwise poly(dC) oligonucleotide (Figure 5.11A and Tables 5.3 and 5.4). The ability of TBM Δ4 to capture DNA-btn•streptavidin complexes was confirmed (data not shown), however it's "voltage-gating" behavior prevented further nucleobase recognition studies using this approach, as the negative potentials (*trans* compartment is negative, relative to the grounded *cis*) required to eject the DNA-btn•streptavidin complex from the pore, would typically result in pore closure.

It was expected that with each genetic truncation, the length of the recognition element would decrease, as fewer bases would reside within the β barrel. The internucleotide distance of an immobilized ssDNA molecule, within

the β barrel is estimated to be between 3-4 Å (see Chapter 3, section 3.2.3). With each truncation of the β barrel, it is estimated that the protein will become ~5 Å shorter (Figure 5.5B). It follows that for every genetic truncation, between 1-2 nucleobases fewer, would reside within the β barrel and be recognized by the pore. Indeed, it was observed that with such subsequent truncation, the length of the recognition element of the protein is reduced. The last nucleobase recognized by the full-length NN pore is at position 19 (relative to the 3′ biotin-tag) and after this, it is assumed that the DNA protrudes from the β barrel and into the *trans* compartment (Chapter 3, section 3.2.3 and (Stoddart, Heron et al. 2009)). However, in the truncated mutants, TBM Δ1, Δ2 and Δ3, the last nucleobase position recognized is 18, 17 and 15 respectively. This suggests that the DNA is protruding from the β barrel sooner in the truncated mutants, as expected, and the length of the recognition element has been reduced. However, the pattern of adenine recognition throughout the truncated mutants is remarkable; it appears that recognition site $R_1$ (near the central constriction) has been weakened, with $R_2$ and $R_3$ remaining, despite the removal of amino acids from the near the *trans* entrance of the pore. This would suggest that base recognition at $R_2$ and $R_3$ is not solely due to the interaction of nucleobases with specific amino acid side chains located at the bottom of the β barrel. It is therefore proposed that when inserted into bilayers, lateral pressure from the lipids may induce changes in the cross-sectional area of the protein to produce constrictions (Cantor 1997; Orsi, Haubertin et al. 2008) that are not obvious from the crystal structure of the protein. Such constrictions may provide the basis for nucleobase recognition at $R_2$ and at the *trans* entrance ($R_3$), in a similar way that

increasing the volume of the amino acid side present at position 113 provided nucleobase recognition at the central constriction. It is also speculated that the entropic constraints that act upon DNA as it is confined within the β barrel, are not imposed on the DNA in the *trans* compartment and as such, the DNA structure goes from rigid (in the β barrel) to flexible (in the *trans* compartment) and at the interface between these two structures the DNA may bend, or kink. Such changes in the structure of the DNA may result in nucleobase recognition, giving rise to recognition site $R_3$. These types of phenomena would presumably be independent of amino acid deletions in the β barrel and therefore such deletions would not be expected to specifically weaken recognition site $R_2$ or $R_3$.

The main of goal of the work described in this chapter, was to remove a recognition site. Although the direct mechanism of nucleobase recognition is not known, especially at $R_2$ or $R_3$, TBM Δ3 appears to have particularly weak nucleobase recognition at the central constriction. Previous work has suggested that amino acid substitutions at position 113 influence nucleobase recognition at $R_1$ with the M113G mutant providing the weakest nucleobase recognition (see Chapter 4, section 4.2.2 and (Stoddart, Heron et al. 2010)). Given this, a combined TBM Δ3/M113G (3G) mutant was made in an attempt to remove the residual $R_1$ recognition site from the TBM Δ3 mutant to create a pore with just two recognition sites. Again, the β barrel of the combined 3G mutant was examined for the ability to discriminate single adenines (Figure 5.11B and Table 5.4) and it was observed that, as expected, the $R_1$ recognition site was removed, whilst $R_2$ and $R_3$ remain largely unchanged compared to the TBM Δ3 precursor protein.

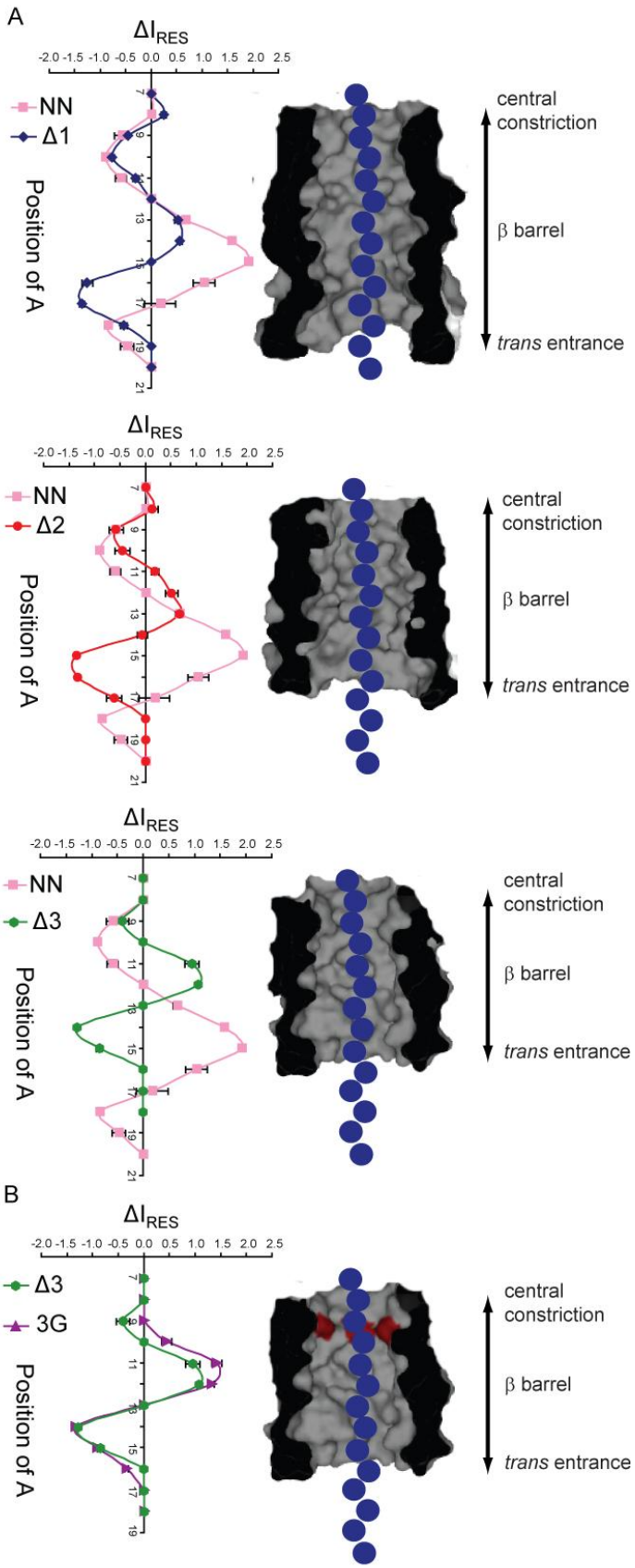



Figure 5.11. The effect of β barrel truncations on adenine recognition along the length of the β barrel. A) The difference in residual current ($\Delta I_{RES}$) between blockades caused by poly(dC) oligonucleotides containing a single adenine base and poly(dC) for the αHL pores NN (pink), TBM Δ1 (blue – upper), TBM Δ2 (red – middle), TBM Δ3 (green – lower) (Table 5.3 and 5.4). A cross-section of the β barrel domain of the truncated αHL pores, filled with DNA indicating the positions of the immobilized bases is shown. B) The difference in residual current ($\Delta I_{RES}$) between blockades caused by poly(dC) oligonucleotides containing a single adenine base and poly(dC) for the αHL pores TBM Δ3 (green) and 3G (purple). A cross-section of the β barrel domain of the truncated αHL pore (showing the position of the mutated M113G amino acid) filled with DNA indicating the position of the immobilized bases is shown (Table 5.4).

Table 5.3. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single adenine nucleobase for TBM Δ1 and Δ2 proteins. The position of the adenine base in the $A_1$ oligonucleotide (nucleobases 7-20) is numbered relative to the 3' biotin tag. The $I_O$ and $I_{RES}$ values are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an $A_1$ oligonucleotides and poly(dC) ($I_{RES}^{A1\ oligo} - I_{RES}^{pC}$). The errors given are standard deviations.

| Position of adenine | TBM Δ1 | | | Position of adenine | TBM Δ2 | | |
|---|---|---|---|---|---|---|---|
| | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| 7 | 35.2 ± 0.2 | 35.2 ± 0.2 | 0.0 ± 0.0 | 7 | 31.3 ± 0.1 | 31.3 ± 0.1 | 0.0 ± 0.0 |
| 8 | 35.4 ± 0.4 | 35.2 ± 0.3 | +0.2 ± 0.1 | 8 | 30.9 ± 1.5 | 30.8 ± 1.4 | +0.1 ± 0.1 |
| 9 | 34.9 ± 0.6 | 35.3 ± 0.6 | −0.5 ± 0.0 | 9 | 30.2 ± 1.0 | 30.8 ± 1.0 | −0.6 ± 0.0 |
| 10 | 34.2 ± 0.4 | 35.0 ± 0.5 | −0.8 ± 0.1 | 10 | 29.3 ± 2.7 | 29.8 ± 2.9 | −0.4 ± 0.2 |
| 11 | 34.9 ± 0.4 | 35.2 ± 0.4 | −0.3 ± 0.1 | 11 | 29.3 ± 2.2 | 29.1 ± 2.2 | +0.2 ± 0.1 |
| 12 | 35.1 ± 0.5 | 35.1 ± 0.5 | 0.0 ± 0.0 | 12 | 30.5 ± 3.0 | 30.0 ± 2.9 | +0.5 ± 0.1 |
| 13 | 35.2 ± 1.1 | 34.6 ± 1.2 | +0.5 ± 0.1 | 13 | 31.5 ± 2.3 | 30.8 ± 2.3 | +0.7 ± 0.0 |
| 14 | 35.9 ± 0.4 | 35.3 ± 0.4 | +0.6 ± 0.0 | 14 | 30.7 ± 1.6 | 30.7 ± 1.7 | −0.1 ± 0.1 |
| 15 | 35.5 ± 0.7 | 35.5 ± 0.7 | +0.0 ± 0.0 | 15 | 29.2 ± 1.6 | 30.5 ± 1.6 | −1.4 ± 0.0 |
| 16 | 33.5 ± 0.3 | 34.7 ± 0.4 | −1.3 ± 0.1 | 16 | 29.2 ± 1.3 | 30.5 ± 1.3 | −1.3 ± 0.0 |
| 17 | 33.9 ± 0.3 | 35.2 ± 0.3 | −1.4 ± 0.0 | 17 | 30.7 ± 1.7 | 31.3 ± 1.8 | −0.6 ± 0.1 |
| 18 | 34.0 ± 0.4 | 34.5 ± 0.4 | −0.5 ± 0.1 | 18 | 31.3 ± 1.5 | 31.3 ± 1.5 | 0.0 ± 0.0 |
| 19 | 34.1 ± 0.9 | 34.1 ± 0.9 | 0.0 ± 0.0 | 19 | 31.9 ± 1.7 | 31.9 ± 1.7 | 0.0 ± 0.0 |
| 20 | 34.4 ± 1.5 | 34.4 ± 1.5 | 0.0 ± 0.0 | 20 | 31.4 ± 0.1 | 31.4 ± 0.1 | −0.1 ± 0.1 |

Table 5.4. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single adenine nucleobase for TBM Δ3 and 3G proteins. The position of the adenine base in the $A_1$ oligonucleotide (nucleobases 7-18) is numbered relative to the 3' biotin tag. The $I_O$ and $I_{RES}$ values are the mean values from n ≥ 3 experiments. $\Delta I_{RES}$ is defined as the difference in residual current between an $A_1$ oligonucleotides and poly(dC) ($I_{RES}^{A1\ oligo}$ - $I_{RES}^{pC}$). The errors given are standard deviations.

| Position of adenine | TBM Δ3 | | | Position of adenine | 3G | | |
|---|---|---|---|---|---|---|---|
| | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) | | $I_{RES}^{A1}$ (%) | $I_{RES}^{pC}$ (%) | $\Delta I_{RES}$ (%) |
| 7 | 21.3 ± 1.4 | 21.3 ± 1.4 | 0.0 ± 0.0 | 7 | 23.5 ± 0.5 | 23.5 ± 0.5 | 0.0 ± 0.0 |
| 8 | 21.3 ± 1.8 | 21.3 ± 1.8 | 0.0 ± 0.0 | 8 | 25.5 ± 2.0 | 25.5 ± 2.0 | 0.0 ± 0.0 |
| 9 | 22.2 ± 1.0 | 22.6 ± 1.1 | −0.4 ± 0.1 | 9 | 24.6 ± 1.3 | 24.6 ± 1.3 | 0.0 ± 0.0 |
| 10 | 21.6 ± 0.3 | 21.6 ± 0.3 | 0.0 ± 0.0 | 10 | 24.1 ± 0.2 | 23.7 ± 0.3 | +0.4 ± 0.1 |
| 11 | 22.8 ± 1.3 | 23.7 ± 1.2 | +0.9 ± 0.1 | 11 | 26.0 ± 1.4 | 24.6 ± 1.3 | +1.4 ± 0.1 |
| 12 | 23.0 ± 0.5 | 21.9 ± 0.6 | +1.1 ± 0.0 | 12 | 25.0 ± 0.2 | 23.7 ± 0.2 | +1.3 ± 0.1 |
| 13 | 22.1 ± 0.6 | 22.1 ± 0.6 | 0.0 ± 0.0 | 13 | 25.9 ± 1.8 | 25.9 ± 1.8 | 0.0 ± 0.0 |
| 14 | 20.8 ± 0.2 | 22.1 ± 0.2 | −1.3 ± 0.0 | 14 | 23.1 ± 1.0 | 24.4 ± 1.0 | −1.3 ± 0.0 |
| 15 | 21.0 ± 0.1 | 21.8 ± 0.1 | −0.9 ± 0.1 | 15 | 22.8 ± 0.2 | 23.7 ± 0.2 | −0.9 ± 0.1 |
| 16 | 21.5 ± 0.1 | 21.5 ± 0.1 | 0.0 ± 0.0 | 16 | 25.5 ± 1.9 | 25.8 ± 1.9 | −0.3 ± 0.1 |
| 17 | 23.0 ± 1.4 | 23.0 ± 1.4 | 0.0 ± 0.0 | 17 | 25.2 ± 1.8 | 25.2 ± 1.8 | 0.0 ± 0.0 |
| 18 | 22.2 ± 0.1 | 22.2 ± 0.1 | 0.0 ± 0.0 | 18 | 23.5 ± 0.4 | 23.5 ± 0.4 | 0.0 ± 0.0 |

### 5.2.4. Probing recognition site $R_3$ of αHL TBM pores for four-base discrimination

As well as examining the β barrel of each of the TBM mutant pores for the ability to discriminate single adenine bases, the lowermost recognition site, $R_3$, was also tested for the capability of discriminating each of the four bases (similar

to previous studies at $R_3$, see Chapter 3, section 3.2.4 and (Stoddart, Heron et al. 2009)). It was observed, that although the each of the TBM proteins contained an $R_3$ recognition site that provided strong discrimination of adenine vs. cytosine bases, only TBM Δ3 (and the 3G derivative) retained an $R_3$ recognition site capable of discriminating the other nucleobases. Furthermore, the distribution of the four current levels (one for each of the four nucleobases) is different in the full-length pores and the truncated pores (Figure 5.12 and Table 5.5). This would imply that the amino acid deletions at the *trans* entrance do have some effect on recognition site $R_3$, although, as mentioned above, the exact contribution of the amino acid side chains is not known.

## 5.3. Conclusions

The aim of this study was to delete a recognition site, and reduce the complexity of the recognition element. In an attempt to weaken recognition site $R_3$, which is proposed to be located at the *trans* entrance of the pore, point mutations were made near the *trans* entrance and in the turn region that connects the two β strands in the monomer. These mutations proved to be either ineffectual (e.g. NN/L135G and Q4N) or had deleterious effects on pore function (e.g. NN/T125G). In a more radical approach, amino acids were removed from near the *trans* entrance by genetic deletion.

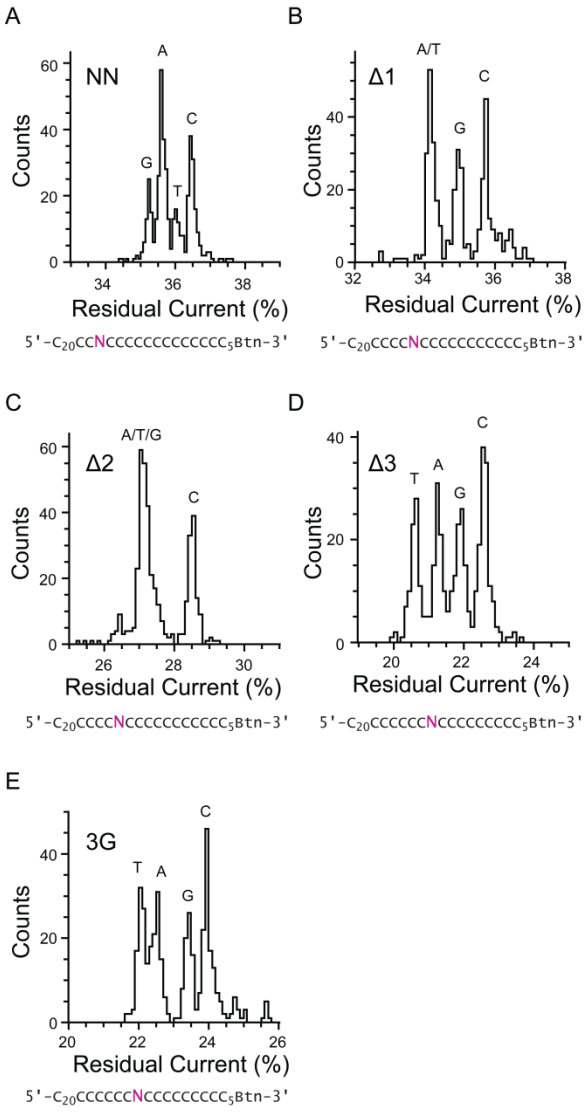



Figure 5.12. Single nucleobase discrimination in truncated αHL pores. Histograms of the residual current levels for NN, TBM Δ1, Δ2, Δ3 and 3G pores are shown. Three sets of four poly(dC) oligonucleotides were used, with each set containing either a single A, T, G or C nucleobase at a specific position. The position of the substituted base, N (purple), was designed to probe the “peak” of the lowest recognition site of each of the proteins (Figure 5.11). All experiments were conducted at least three times (Table 5.5), and the results displayed in the figure are from a typical experiment. A) NN pores were interrogated with 5'-CCCCCCCCCCCCCCCCCCCCCCNCCCCCCCCCCCCCCCCCCBtn–3'. TBM Δ1 (panel B) and TBM Δ2 (panel C) pores were interrogated with four oligonucleotides with the sequence 5'–CCCCCCCCCCCCCCCCCCCCCCCCCCNCCCCCCCCCCCCCCCCCBtn–3'. TBM Δ3 (panel D) and 3G (panel E) pores were interrogated with four oligonucleotides with the sequence 5'–CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCNCCCCCCCCCCCCCCCCCBtn–3'.

Table 5.5. Residual currents ($I_{RES}$) for poly(dC) and oligonucleotides that contain a single substituted base at position 18 (NN), 16 (TBM Δ1 and Δ2) or 13 (TBM Δ3 and 3G) blocking pores. $I_{RES}$, $\Delta I_{RES}^{OVERALL}$ and δ values are shown for n ≥ 3 experiments. $\Delta I_{RES}^{OVERALL}$ is defined as the difference in residual current between the two most widely dispersed current peaks. The δ value is the product of the successive differences between the peaks. If any two peaks overlap, then the δ value is equal to zero. Mean values (± s.d.) are given.

| | Residual Current (%) | | | | Residual Current Difference (%) | |
|---|---|---|---|---|---|---|
| NN | $C_{18}$ | $T_{18}$ | $A_{18}$ | $G_{18}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 36.9 ± 0.5 | 36.4 ± 0.5 | 36.0 ± 0.5 | 35.7 ± 0.5 | 1.2 ± 0.0 | 0.07 ± 0.00 |
| Δ1 | $C_{16}$ | $T_{16}$ | $A_{16}$ | $G_{16}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 35.7 ± 0.1 | 34.2 ± 0.1 | 34.2 ± 0.1 | 35.0 ± 0.1 | 1.5 ± 0.1 | 0.00 ± 0.00 |
| Δ2 | $C_{16}$ | $T_{16}$ | $A_{16}$ | $G_{16}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 31.4 ± 2.8 | 30.0 ± 2.7 | 30.0 ± 2.7 | 30.0 ± 2.7 | 1.4 ± 0.1 | 0.00 ± 0.00 |
| Δ3 | $C_{13}$ | $T_{13}$ | $A_{13}$ | $G_{13}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 22.9 ± 1.4 | 20.9 ± 1.3 | 21.6 ± 1.3 | 22.2 ± 1.3 | 2.0 ± 0.2 | 0.29 ± 0.06 |
| 3G | $C_{13}$ | $T_{13}$ | $A_{13}$ | $G_{13}$ | $\Delta I_{RES}^{OVERALL}$ (%) | δ |
| | 24.2 ± 0.3 | 22.4 ± 0.3 | 22.8 ± 0.3 | 23.7 ± 0.3 | 1.8 ± 0.1 | 0.19 ± 0.03 |

Remarkably, substantial portions of the β barrel could be removed without having significantly adverse effects on *in vitro* pore function. As hoped, truncations in the β barrel of the αHL pore did indeed reduce the number of bases recognized by the pore and this could hopefully simplify the current signal elicited from a working device. Surprisingly however, recognition site $R_3$ was not removed (although its properties were altered), suggesting that amino acid side chains are not the sole arbiters of nucleobase discrimination. The TBM Δ3

protein showed a weakened $R_1$ and it was possible to weaken this recognition site still further by additional site-directed mutagenesis to generate the 3G mutant pore, which displayed just two recognition sites; $R_2$ and $R_3$. Subsequent analysis of the 3G pore showed that recognition site $R_3$ was still capable of four nucleobase discrimination, as the full-length protein was, although it was observed that the distributions of the current levels differed.

References


Andersen, O. S. and R. E. Koeppe, 2nd (2007). "Bilayer thickness and membrane protein function: an energetic perspective." Annu Rev Biophys Biomol Struct **36**: 107-130.

Bezrukov, S. M. and J. J. Kasianowicz (1993). "Current Noise Reveals Protonation Kinetics and Number of Ionizable Sites in an Open Protein Ion Channel." Physical Review Letters **70**(15): 2352-2355.

Bhakdi, S., J. Tranum-Jensen, et al. (1985). "Mechanism of membrane damage by streptolysin-O." Infect Immun **47**(1): 52-60.

Cantor, R. S. (1997). "The lateral pressure profile in membranes: a physical mechanism of general anesthesia." Biochemistry **36**(9): 2339-2344.

Cheley, S., O. Braha, et al. (1999). "A functional protein pore with a "retro" transmembrane domain." Protein Sci. **8**: 1257-1267.

Cheley, S., M. S. Malghani, et al. (1997). "Spontaneous oligomerization of a staphylococcal a-hemolysin conformationally constrained by removal of residues that form the transmembrane b barrel." Protein Eng. **10**: 1433-1443.

Chen, M., S. Khalid, et al. (2008). "Outer membrane protein G: Engineering a quiet pore for biosensing." Proc Natl Acad Sci U S A **105**(17): 6272-6277.

Duque, D., X. J. Li, et al. (2002). "Molecular theory of hydrophobic mismatch between lipids and peptides." Biophysical Journal **82**(1): 515a-515a.

Killian, J. A. (1998). "Hydrophobic mismatch between proteins and lipids in membranes." Biochim Biophys Acta **1376**(3): 401-415.

Korchev, Y. E., C. L. Bashford, et al. (1997). "A novel explanation for fluctuations of ion current through narrow pores." Faseb Journal **11**(7): 600-608.

Maglia, G., A. J. Heron, et al. (2009). "Droplet networks with incorporated protein diodes show collective properties." Nat Nanotechnol **4**(7): 437-440.

Maglia, G., A. J. Heron, et al. (2010). "Analysis of single nucleic acid molecules with protein nanopores." Methods Enzymol **475**: 591-623.

Orsi, M., D. Y. Haubertin, et al. (2008). "A quantitative coarse-grain model for lipid bilayers." J Phys Chem B **112**(3): 802-815.

Palmer, M., R. Harris, et al. (1998). "Assembly mechanism of the oligomeric streptolysin O pore: the early membrane lesion is lined by a free edge of the lipid membrane and is extended gradually during oligomerization." EMBO J **17**(6): 1598-1605.

Phillips, R., T. Ursell, et al. (2009). "Emerging roles for lipids in shaping membrane-protein function." Nature **459**: 379-385.

Sekiya, K., R. Satoh, et al. (1993). "A ring-shaped structure with a crown formed by streptolysin O on the erythrocyte membrane." J Bacteriol **175**(18): 5953-5961.

Song, L., M. R. Hobaugh, et al. (1996). "Structure of staphylococcal a-hemolysin, a heptameric transmembrane pore." Science **274**: 1859-1865.

Stoddart, D., A. Heron, et al. (2009). "Single nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore." Proc. Natl. Acad. Sci. USA **106**: 7702-7707.

Stoddart, D., A. J. Heron, et al. (2010). "Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore." Nano Lett **10**(9): 3633-3637.

Wai, S. N., M. Westermark, et al. (2003). "Characterization of dominantly negative mutant ClyA cytotoxin proteins in Escherichia coli." J Bacteriol **185**(18): 5491-5499.

Walker, B., M. Krishnasastry, et al. (1992). "Assembly of the oligomeric membrane pore formed by Staphylococcal alpha-hemolysin examined by truncation mutagenesis." J Biol Chem **267**(30): 21782-21786.

# CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

## 6.1. Conclusions

The ability to immobilize DNA within the αHL pore improves the precision to which the current signal can be measured and hence has allowed for the nucleobase recognition properties of the protein to be exquisitely examined. The recognition element of the pore appears to be located within the narrow confines of the β barrel and comprises at least three separate recognition sites, termed $R_1$ (located at the central constriction), $R_2$ (located in the middle of the barrel) and $R_3$ (located at the *trans* entrance). It has been shown that the αHL pore is capable of separating all four DNA bases at one or more of these recognition sites. Furthermore, mutations at the central constriction have a profound effect on the nucleobase recognition properties of the $R_1$ recognition site, providing a method to improve, or weaken that site, without grossly altering the properties of distal sites. Amino acids with related side chains produce similar patterns of recognition, providing a rationale for the tuning of base recognition. The use of multiple recognition sites within a single nanopore was observed to provide extra sequencing information which may prove useful in a working system, although it is believed that given the limitations of the system, a two-site sensor (rather than three) would be optimal. It is shown that site-directed mutation and genetic truncations can be combined to remove one of the recognition sites from the pore in an effort to simplify the elicited current signal from the multi-site nanopore.

## 6.2. Future directions

Additional progress is still required to produce a nanopore DNA sequencing device. The primary issue that needs to be addressed, and is currently being examined by several research groups, is the control of DNA translocation through the pore. As indicated earlier, DNA passes through the pore too quickly under the potentials required for threading to allow base identification. Several variables have been investigated in attempts to slow down DNA translocation speed, such as viscosity (Kawano, Schibel et al. 2009), electrolyte composition (de Zoysa, Jayawardhana et al. 2009; Wanunu, Morrison et al. 2010) and pore mutation (Rincon-Restrepo, Mikhailova et al. 2011) but these attempts were unable to slow the DNA translocation rate to the required level of ~10 ms $nt^{-1}$. However, DNA processing enzymes, such as DNA polymerases and exonucleases, do operate at these rates. Therefore, such proteins would be ideal candidates to sit on top of the pore and feed DNA through the lumen, in a controlled fashion and at a controlled speed, even under an applied potential. Some enzymes, such as ExoI may be unsuitable for this purpose, as it has been observed that the DNA is often “pulled off” of the enzyme under high potentials (Hornblower, Coombs et al. 2007). Work in this area is highly encouraging and both the Ghadiri (Cockroft, Chu et al. 2008; Chu, Gonzalez-Lopez et al. 2010) and Akeson (Lieberman, Cherf et al. 2010; Olasagasti, Lieberman et al. 2010) groups have shown that it is possible to identify markers (phosphoPEG or abasic DNA, respectively) as they are ratcheted through the pore by a DNA polymerase, although actual base

identification has not yet been observed. It is proposed that an αHL with superior nucleobase identification properties, compared to the WT pore, e.g. the NN pore, may be able to “read” bases as they are pulled through the pore by a DNA polymerase. It is likely that by using a DNA polymerase, or a related approach, that DNA base identification in a working system could be possible soon. Once data emerges, it will be intriguing to see if the patterns that have been observed using the DNA-btn•streptavidin approach, are applicable in the working system. If so, then it should be possible to “tune” the pore such that base identification properties can be optimized, using data such as those presented in this thesis, or gained by a similar approach.

Once controlled translocation and base identification have been achieved, the next step will be to parallelize the set-up. A single pore, working at a rate of 100 bases $s^{-1}$, will be able to sequence a human genome in ~700 days, which would not be competitive with current sequencing technologies. If the number of nanopores in a device could be scaled-up to around $10^6$, then it would be possible to offer a “10 minute” genome. This is probably the greatest challenge that must be overcome in order to launch a commercial product, for many engineering reasons. First, the chip that would house the electronics needs to be sufficiently small, such that the sample volumes can be kept to a minimum. However, the chip size must be large enough that it is possible to house all the electronics and insulate all the electrodes from one and other to reduce noise.

A further road-block is the necessity for bilayers, which are delicate and prone to rupture (when prepared using current experimental methods), and for each bilayer to contain a single channel. Although several studies are being

carried out in an attempt to stabilize bilayers (for example, bilayers can be polymerized to prevent rupture (Heitz, Xu et al. 2009)), it will be hard to overcome the Poisson limit (since pore distribution is a stochastic process), which dictates that no more than 37% of bilayers will contain single pores (other bilayers will either contain multiple pores, or no pore at all). The use of a solid-state platform for protein nanopores could solve both these issues; solid-state platforms are durable (unlike current bilayer systems) and apertures within a solid-state wafer could be made just large enough to only a fit a single protein nanopore. Recent collaborative work between the Bayley group and the Dekker group in the Netherlands has shown that αHL pores can be immobilized within larger solid-state apertures which have been fabricated in a silicon nitride sheet (Hall, Scott et al. 2010). These pores can pass ssDNA, but not dsDNA (Figure 6.1A). If this approach can be expanded to include an array in which many pores can be examined simultaneously, then it could prove to be advantageous over the current method of supporting αHL pores in bilayers. Indeed this approach would allow for the truncated barrel proteins to be used as the protein nanopore, even though the shortened barrels, e.g. TBM Δ5, don't insert into bilayers (Figure 6.1B). Further to this, it may be possible to insert non-pore forming proteins inside the solid-state support. It is envisaged that ring-forming proteins, such as λ exonuclease, could act as both the mechanism to slow down the DNA, and as the "base reader" (Figure 6.1C).

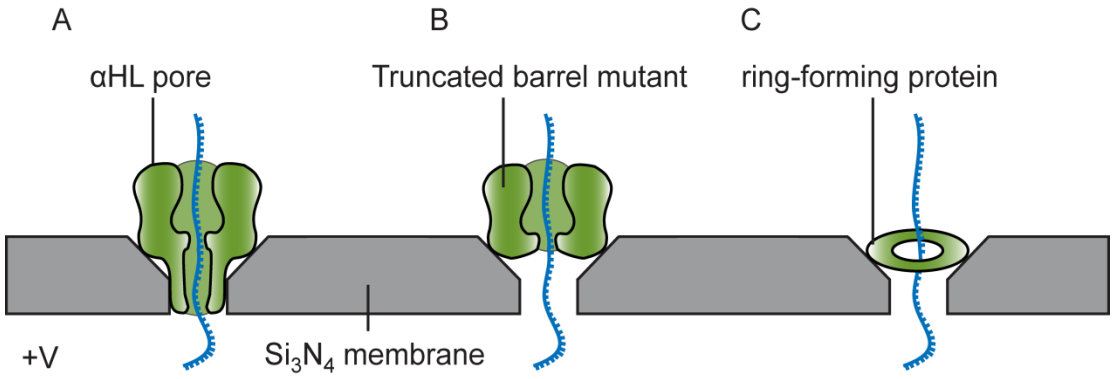



Figure 6.1. Protein•solid-state nanopore hybrids offer advantages of durability (solid-sate), and precise control and homogeneity (protein). A) Representation of an αHL nanopore embedded in a solid-state support. B) Nanopore mutants that do not insert into planar bilayers, but offer superior base-calling potential, such as the TBM proteins, could also be fitted into solid-state apertures. C) Protein•solid-state hybrids could negate the need for a transmembrane protein, and ring-forming, DNA processing enzymes, such as λ exonuclease, could be used to both slow down the DNA as it translocates while simultaneously reading the bases.

References

Chu, J., M. Gonzalez-Lopez, et al. (2010). "Real-time monitoring of DNA polymerase function and stepwise single-nucleotide DNA strand translocation through a protein nanopore." Angew Chem Int Ed Engl **49**(52): 10106-10109.

Cockroft, S. L., J. Chu, et al. (2008). "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J Am Chem Soc **130**(3): 818-820.

de Zoysa, R. S., D. A. Jayawardhana, et al. (2009). "Slowing DNA translocation through nanopores using a solution containing organic salts." J Phys Chem B **113**(40): 13332-13336.

Hall, A. R., A. Scott, et al. (2010). "Hybrid pore formation by directed insertion of alpha-haemolysin into solid-state nanopores." Nat Nanotechnol **5**(12): 874-877.

Heitz, B. A., J. Xu, et al. (2009). "Enhanced long-term stability for single ion channel recordings using suspended poly(lipid) bilayers." J Am Chem Soc **131**(19): 6662-6663.

Hornblower, B., A. Coombs, et al. (2007). "Single-molecule analysis of DNA-protein complexes using nanopores." Nat Methods **4**(4): 315-317.

Kawano, R., A. E. Schibel, et al. (2009). "Controlling the translocation of single-stranded DNA through alpha-hemolysin ion channels using viscosity." Langmuir **25**(2): 1233-1237.

Lieberman, K. R., G. M. Cherf, et al. (2010). "Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase." J Am Chem Soc **132**(50): 17961-17972.

Olasagasti, F., K. R. Lieberman, et al. (2010). "Replication of individual DNA molecules under electronic control using a protein nanopore." Nat Nanotechnol.

Rincon-Restrepo, M., E. Mikhailova, et al. (2011). "Controlled Translocation of Individual DNA Molecules through Protein Nanopores with Engineered Molecular Brakes." Nano Lett.

Wanunu, M., W. Morrison, et al. (2010). "Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient." Nat Nanotechnol **5**(2): 160-165.

# APPENDIX

## A.1 Gene sequences

All genes were generated and expressed in the pT7 vector

### Wild-type (WT)

```
   1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
  71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
 141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
 211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
 281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAGAGTATA TGAGTACTTT
 351  AACTTATGGA TTCAACGGTA ATGTTACTGG TGATGATACA GGAAAAATTG GCGGCCTTAT TGGTGCAAAT
 421  GTTTCGATTG GTCATACACT GAAATATGTT CAACCTGATT TCAAAACAAT TTTAGAGAGC CCAACTGATA
 491  AAAAAGTAGG CTGGAAAGTG ATATTTAACA ATATGGTGAA TCAAAATTGG GGACCATACG ATCGAGATTC
 561  TTGGAACCCG GTATATGGCA ATCAACTTTT CATGAAAACT AGAAATGGTT CTATGAAAGC AGCAGATAAC
 631  TTCCTTGATC CTAACAAAGC AAGTTCTCTA TTATCTTCAG GGTTTTCACC AGACTTCGCT ACAGTTATTA
 701  CTATGGATAG AAAAGCATCC AAACAACAAA CAAATATAGA TGTAATATAC GAACGAGTTC GTGATGATTA
 771  CCAATTGCAT TGGACTTCAA CAAATTGGAA AGGTACCAAT ACTAAAGATA AATGGACAGA TCGTTCTTCA
 841  GAAAGATATA AAATCGATTG GGAAAAAGAA GAAATGACAA AT
```

### E111N/K147N (NN) – prepared by E. Mikhailova

```
   1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
  71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
 141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
 211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
 281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTTT
 351  AACTTATGGA TTCAACGGTA ATGTTACTGG TGATGATACA GGAAAAATTG GCGGCCTTAT TGGTGCAAAT
 421  GTTTCGATTG GTCATACACT GAACTATGTT CAACCTGATT TCAAAACAAT TTTAGAGAGC CCAACTGATA
 491  AAAAAGTAGG CTGGAAAGTG ATATTTAACA ATATGGTGAA TCAAAATTGG GGACCATACG ATCGAGATTC
 561  TTGGAACCCG GTATATGGCA ATCAACTTTT CATGAAAACT AGAAATGGTT CTATGAAAGC AGCAGATAAC
 631  TTCCTTGATC CTAACAAAGC AAGTTCTCTA TTATCTTCAG GGTTTTCACC AGACTTCGCT ACAGTTATTA
 701  CTATGGATAG AAAAGCATCC AAACAACAAA CAAATATAGA TGTAATATAC GAACGAGTTC GTGATGATTA
 771  CCAATTGCAT TGGACTTCAA CAAATTGGAA AGGTACCAAT ACTAAAGATA AATGGACAGA TCGTTCTTCA
 841  GAAAGATATA AAATCGATTG GGAAAAAGAA GAAATGACAA AT
```

### E111N/K147N/M113X (NNX) – prepared by E. Mikhailova

```
   1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
  71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
 141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
 211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
 281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATX XXAGTACTTT
 351  AACTTATGGA TTCAACGGTA ATGTTACTGG TGATGATACA GGAAAAATTG GCGGCCTTAT TGGTGCAAAT
 421  GTTTCGATTG GTCATACACT GAACTATGTT CAACCTGATT TCAAAACAAT TTTAGAGAGC CCAACTGATA
 491  AAAAAGTAGG CTGGAAAGTG ATATTTAACA ATATGGTGAA TCAAAATTGG GGACCATACG ATCGAGATTC
 561  TTGGAACCCG GTATATGGCA ATCAACTTTT CATGAAAACT AGAAATGGTT CTATGAAAGC AGCAGATAAC
 631  TTCCTTGATC CTAACAAAGC AAGTTCTCTA TTATCTTCAG GGTTTTCACC AGACTTCGCT ACAGTTATTA
 701  CTATGGATAG AAAAGCATCC AAACAACAAA CAAATATAGA TGTAATATAC GAACGAGTTC GTGATGATTA
 771  CCAATTGCAT TGGACTTCAA CAAATTGGAA AGGTACCAAT ACTAAAGATA AATGGACAGA TCGTTCTTCA
 841  GAAAGATATA AAATCGATTG GGAAAAAGAA GAAATGACAA AT
```

## Δ124-125, Δ133-134/E111N/K147N (TBM Δ1)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTTT
351  AACTTATGGA TTCAACGGTA ATGGTGATGA TACAGGAAAA ATTCTTATTG GTGCAAATGT TTCGATTGGT
421  CATACACTGA ACTATGTTCA ACCTGATTTC AAAACAATTT TAGAGAGCCC AACTGATAAA AAAGTAGGCT
491  GGAAAGTGAT ATTTAACAAT ATGGTGAATC AAAATTGGGG ACCATACGAT CGAGATTCTT GGAACCCGGT
561  ATATGGCAAT CAACTTTTCA TGAAAACTAG AAATGGTTCT ATGAAAGCAG CAGATAACTT CCTTGATCCT
631  AACAAAGCAA GTTCTCTATT ATCTTCAGGG TTTTCACCAG ACTTCGCTAC AGTTATTACT ATGGATAGAA
701  AAGCATCCAA ACAACAAACA AATATAGATG TAATATACGA ACGAGTTCGT GATGATTACC AATTGCATTG
771  GACTTCAACA AATTGGAAAG GTACCAATAC TAAAGATAAA TGGACAGATC GTTCTTCAGA AAGATATAAA
841  ATCGATTGGG AAAAAGAAGA AATGACAAAT
```

## Δ122-125, Δ133-136/E111N/K147N (TBM Δ2)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTTT
351  AACTTATGGA TTCAACGGTG ATGATACAGG AAAAATTGGT GCAAATGTTT CGATTGGTCA TACACTGAAC
421  TATGTTCAAC CTGATTTCAA AACAATTTTA GAGAGCCCAA CTGATAAAAA AGTAGGCTGG AAAGTGATAT
491  TTAACAATAT GGTGAATCAA AATTGGGGAC CATACGATCG AGATTCTTGG AACCCGGTAT ATGGCAATCA
561  ACTTTTCATG AAAACTAGAA ATGGTTCTAT GAAAGCAGCA GATAACTTCC TTGATCCTAA CAAAGCAAGT
631  TCTCTATTAT CTTCAGGGTT TTCACCAGAC TTCGCTACAG TTATTACTAT GGATAGAAAA GCATCCAAAC
701  AACAAACAAA TATAGATGTA ATATACGAAC GAGTTCGTGA TGATTACCAA TTGCATTGGA CTTCAACAAA
771  TTGGAAAGGT ACCAATACTA AAGATAAATG GACAGATCGT TCTTCAGAAA GATATAAAAT CGATTGGGAA
841  AAAGAAGAAA TGACAAAT
```

## Δ120-125, Δ133-138/E111N/K147N (TBM Δ3)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTTT
351  AACTTATGGA GGTGATGATA CAGGAAAAAT TAATGTTTCG ATTGGTCATA CACTGAACTA TGTTCAACCT
421  GATTTCAAAA CAATTTTAGA GAGCCCAACT GATAAAAAAG TAGGCTGGAA AGTGATATTT AACAATATGG
491  TGAATCAAAA TTGGGGACCA TACGATCGAG ATTCTTGGAA CCCGGTATAT GGCAATCAAC TTTTCATGAA
561  AACTAGAAAT GGTTCTATGA AAGCAGCAGA TAACTTCCTT GATCCTAACA AAGCAAGTTC TCTATTATCT
631  TCAGGGTTTT CACCAGACTT CGCTACAGTT ATTACTATGG ATAGAAAAGC ATCCAAACAA CAAACAAATA
701  TAGATGTAAT ATACGAACGA GTTCGTGATG ATTACCAATT GCATTGGACT TCAACAAATT GGAAAGGTAC
771  CAATACTAAA GATAAATGGA CAGATCGTTC TTCAGAAAGA TATAAAATCG ATTGGGAAAA AGAAGAAATG
841  ACAAAT
```

## Δ120-125, Δ133-138/E111N/K147N/M113G (3G)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATG GGAGTACTTT
351  AACTTATGGA GGTGATGATA CAGGAAAAAT TAATGTTTCG ATTGGTCATA CACTGAACTA TGTTCAACCT
421  GATTTCAAAA CAATTTTAGA GAGCCCAACT GATAAAAAAG TAGGCTGGAA AGTGATATTT AACAATATGG
491  TGAATCAAAA TTGGGGACCA TACGATCGAG ATTCTTGGAA CCCGGTATAT GGCAATCAAC TTTTCATGAA
561  AACTAGAAAT GGTTCTATGA AAGCAGCAGA TAACTTCCTT GATCCTAACA AAGCAAGTTC TCTATTATCT
631  TCAGGGTTTT CACCAGACTT CGCTACAGTT ATTACTATGG ATAGAAAAGC ATCCAAACAA CAAACAAATA
701  TAGATGTAAT ATACGAACGA GTTCGTGATG ATTACCAATT GCATTGGACT TCAACAAATT GGAAAGGTAC
771  CAATACTAAA GATAAATGGA CAGATCGTTC TTCAGAAAGA TATAAAATCG ATTGGGAAAA AGAAGAAATG
```

```
841  ACAAAT
```

## Δ118-125, Δ133-140/E111N/K147N (TBM Δ4)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTTT
351  AACTGGTGAT GATACAGGAA AAATTTCGAT TGGTCATACA CTGAACTATG TTCAACCTGA TTTCAAAACA
421  ATTTTAGAGA GCCCAACTGA TAAAAAAGTA GGCTGGAAAG TGATATTTAA CAATATGGTG AATCAAAATT
491  GGGGACCATA CGATCGAGAT TCTTGGAACC CGGTATATGG CAATCAACTT TTCATGAAAA CTAGAAATGG
561  TTCTATGAAA GCAGCAGATA ACTTCCTTGA TCCTAACAAA GCAAGTTCTC TATTATCTTC AGGGTTTTCA
631  CCAGACTTCG CTACAGTTAT TACTATGGAT AGAAAAGCAT CCAAACAACA AACAAATATA GATGTAATAT
701  ACGAACGAGT TCGTGATGAT TACCAATTGC ATTGGACTTC AACAAATTGG AAAGGTACCA ATACTAAAGA
771  TAAATGGACA GATCGTTCTT CAGAAAGATA TAAAATCGAT TGGGAAAAAG AAGAAATGAC AAAT
```

## Δ116-125, Δ133-142/E111N/K147N (TBM Δ5)

```
  1  ATGGCAGATT CTGATATTAA TATTAAAACC GGTACTACAG ATATTGGAAG CAATACTACA GTAAAAACAG
 71  GTGATTTAGT CACTTATGAT AAAGAAAATG GCATGCACAA AAAAGTATTT TATAGTTTTA TCGATGATAA
141  AAATCACAAT AAAAAACTGC TAGTTATTAG AACAAAAGGT ACCATTGCTG GTCAATATAG AGTTTATAGC
211  GAAGAAGGTG CTAACAAAAG TGGTTTAGCC TGGCCTTCAG CCTTTAAGGT ACAGTTGCAA CTACCTGATA
281  ATGAAGTAGC TCAAATATCT GATTACTATC CAAGAAATTC GATTGATACA AAAAACTATA TGAGTACTGG
351  TGATGATACA GGAAAAATTG GTCATACACT GAACTATGTT CAACCTGATT TCAAAACAAT TTTAGAGAGC
421  CCAACTGATA AAAAAGTAGG CTGGAAAGTG ATATTTAACA ATATGGTGAA TCAAAATTGG GGACCATACG
491  ATCGAGATTC TTGGAACCCG GTATATGGCA ATCAACTTTT CATGAAAACT AGAAATGGTT CTATGAAAGC
561  AGCAGATAAC TTCCTTGATC CTAACAAAGC AAGTTCTCTA TTATCTTCAG GGTTTTCACC AGACTTCGCT
631  ACAGTTATTA CTATGGATAG AAAAGCATCC AAACAACAAA CAAATATAGA TGTAATATAC GAACGAGTTC
701  GTGATGATTA CCAATTGCAT TGGACTTCAA CAAATTGGAA AGGTACCAAT ACTAAAGATA AATGGACAGA
771  TCGTTCTTCA GAAAGATATA AAATCGATTG GGAAAAAGAA GAAATGACAA AT
```

## A.2. Primer sequences

nmR
```
1  CAGAAGTGGT CCTGCAACTT TAT
```

nmF
```
1  ATAAAGTTGC AGGACCACTT CTG
```

## Δ124-125, Δ133-134/E111N/K147N (TBM Δ1)

mF
```
1  GGTGATGATA CAGGAAAAAT TCTTATTGGT GCAAATGTTT CGATTG
```

mR
```
1  AATTTTTCCT GTATCATCAC CATTACCGTT GAATCCATAA GTTAAAG
```

## Δ122-125, Δ133-136/E111N/K147N (TBM Δ2)

mF
```
1  GGTGATGATA CAGGAAAAAT TGGTGCAAAT GTTTCGATTG GTCATAC
```

mR
```
1  AATTTTTCCT GTATCATCAC CGTTGAATCC ATAAGTTAAA GTACTC
```

### Δ120-125, Δ133-138/E111N/K147N (TBM Δ3)

mF `1 GGTGATGATA CAGGAAAAAT TAATGTTTCG ATTGGTCATA CACTG`

mR `1 AATTTTTCCT GTATCATCAC CTCCATAAGT TAAAGTACTC ATATA`

### Δ120-125, Δ133-138/E111N/K147N/M113G (3G)

mF `1 AAAAACTATG GGAGTACTTT AACTTATGG`

mR `1 AGTTAAAGTA CTCCCATAGT TTTTTGTATC`

### Δ118-125, Δ133-140/E111N/K147N (TBM Δ4)

mF `1 GGTGATGATA CAGGAAAAAT TTCGATTGGT CATACACTGA ACTATG`

mR `1 AATTTTTCCT GTATCATCAC CAGTTAAAGT ACTCATATAG TTTTTTG`

### Δ116-125, Δ133-142/E111N/K147N (TBM Δ5)

mF `1 GGTGATGATA CAGGAAAAAT TGGTCATACA CTGAACTATG TTCAAC`

mR `1 AATTTTTCCT GTATCATCAC CAGTACTCAT ATAGTTTTTT GTATC`

## A.3. Protein sequences

### Wild-type (WT)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK EYMSTLTYGF NGNVTGDDTG KIGGLIGANV
141  SIGHTLKYVQ PDFKTILESP TDKKVGWKVI FNNMVNQNWG PYDRDSWNPV YGNQLFMKTR NGSMKAADNF
211  LDPNKASSLL SSGFSPDFAT VITMDRKASK QQTNIDVIYE RVRDDYQLHW TSTNWKGTNT KDKWTDRSSE
281  RYKIDWEKEE MTN
```

### E111N/K147N (NN)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTLTYGF NGNVTGDDTG KIGGLIGANV
141  SIGHTLNYVQ PDFKTILESP TDKKVGWKVI FNNMVNQNWG PYDRDSWNPV YGNQLFMKTR NGSMKAADNF
211  LDPNKASSLL SSGFSPDFAT VITMDRKASK QQTNIDVIYE RVRDDYQLHW TSTNWKGTNT KDKWTDRSSE
281  RYKIDWEKEE MTN
```

### E111N/K147N/M113X (NNX)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
```

```
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYXSTLTYGF NGNVTGDDTG KIGGLIGANV
141  SIGHTLNYVQ PDFKTILESP TDKKVGWKVI FNNMVNQNWG PYDRDSWNPV YGNQLFMKTR NGSMKAADNF
211  LDPNKASSLL SSGFSPDFAT VITMDRKASK QQTNIDVIYE RVRDDYQLHW TSTNWKGTNT KDKWTDRSSE
281  RYKIDWEKEE MTN
```

## Δ124-125, Δ133-134/E111N/K147N (TBM Δ1)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTLTYGF NGNGDDTGKI LIGANVSIGH
141  TLNYVQPDFK TILESPTDKK VGWKVIFNNM VNQNWGPYDR DSWNPVYGNQ LFMKTRNGSM KAADNFLDPN
211  KASSLLSSGF SPDFATVITM DRKASKQQTN IDVIYERVRD DYQLHWTSTN WKGTNTKDKW TDRSSERYKI
281  DWEKEEMTN
```

## Δ122-125, Δ133-136/E111N/K147N (TBM Δ2)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTLTYGF NGDDTGKIGA NVSIGHTLNY
141  VQPDFKTILE SPTDKKVGWK VIFNNMVNQN WGPYDRDSWN PVYGNQLFMK TRNGSMKAAD NFLDPNKASS
211  LLSSGFSPDF ATVITMDRKA SKQQTNIDVI YERVRDDYQL HWTSTNWKGT NTKDKWTDRS SERYKIDWEK
281  EEMTN
```

## Δ120-125, Δ133-138/E111N/K147N (TBM Δ3)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTLTYGG DDTGKINVSI GHTLNYVQPD
141  FKTILESPTD KKVGWKVIFN NMVNQNWGPY DRDSWNPVYG NQLFMKTRNG SMKAADNFLD PNKASSLLSS
211  GFSPDFATVI TMDRKASKQQ TNIDVIYERV RDDYQLHWTS TNWKGTNTKD KWTDRSSERY KIDWEKEEMT
281  N
```

## Δ120-125, Δ133-138/E111N/K147N/M113G (3G)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYGSTLTYGG DDTGKINVSI GHTLNYVQPD
141  FKTILESPTD KKVGWKVIFN NMVNQNWGPY DRDSWNPVYG NQLFMKTRNG SMKAADNFLD PNKASSLLSS
211  GFSPDFATVI TMDRKASKQQ TNIDVIYERV RDDYQLHWTS TNWKGTNTKD KWTDRSSERY KIDWEKEEMT
281  N
```

## Δ118-125, Δ133-140/E111N/K147N (TBM Δ4)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTLTGDD TGKISIGHTL NYVQPDFKTI
141  LESPTDKKVG WKVIFNNMVN QNWGPYDRDS WNPVYGNQLF MKTRNGSMKA ADNFLDPNKA SSLLSSGFSP
211  DFATVITMDR KASKQQTNID VIYERVRDDY QLHWTSTNWK GTNTKDKWTD RSSERYKIDW EKEEMTN
```

## Δ118-125, Δ133-140/E111N/K147N (TBM Δ5)

```
  1  ADSDINIKTG TTDIGSNTTV KTGDLVTYDK ENGMHKKVFY SFIDDKNHNK KLLVIRTKGT IAGQYRVYSE
 71  EGANKSGLAW PSAFKVQLQL PDNEVAQISD YYPRNSIDTK NYMSTGDDTG KIGHTLNYVQ PDFKTILESP
141  TDKKVGWKVI FNNMVNQNWG PYDRDSWNPV YGNQLFMKTR NGSMKAADNF LDPNKASSLL SSGFSPDFAT
211 VITMDRKASK QQTNIDVIYE RVRDDYQLHW TSTNWKGTNT KDKWTDRSSE RYKIDWEKEE MTN
```

A.4. Oligonucleotide sequences

Table A.1. Sequences of the oligonucleotides used in this study. B represents the 3' biotin-TEG tag and linker (Figure 2.7).

| Oligonucleotide name | Oligonucleotide sequence (5'→3') |
|---|---|
| poly(dC)$_{60}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC-B |
| poly(dA)$_{60}$ | AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA-B |
| poly(dC)$_{40}$/poly(dC)$^{C9C14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC-B |
| poly(dC)$^{A1\text{-}5}$ (oligo i) | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAA-B |
| poly(dC)$^{A6\text{-}10}$ (oligo ii) | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAACCCCC-B |
| poly(dC)$^{A11\text{-}15}$ (oligo iii) | CCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAACCCCCCCCC-B |
| poly(dC)$^{A16\text{-}20}$ (oligo iv) | CCCCCCCCCCCCCCCCCCCCAAAAACCCCCCCCCCCCCCCC-B |
| poly(dC)$^{A21\text{-}25}$ (oligo v) | CCCCCCCCCCCCCCCCAAAAACCCCCCCCCCCCCCCCCCC-B |
| poly(dC)$^{A7}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCC-B |
| poly(dC)$^{A8}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCC-B |
| poly(dC)$^{A9}$/poly(dC)$^{A9C14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCC-B |
| poly(dC)$^{A10}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A11}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A12}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A13}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A14}$/poly(dC)$^{C9A14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A15}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A16}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A17}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A18}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A19}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{A20}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCC-B |
| poly(dC)$^{T9}$/poly(dC)$^{T9C14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCC-B |
| poly(dC)$^{G9}$/poly(dC)$^{G9C14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCCCCC-B |
| poly(dC)$^{T14}$/poly(dC)$^{C9T14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCC-B |
| poly(dC)$^{G14}$/poly(dC)$^{C9G14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCCCCCCCCC-B |
| poly(dC)$^{T16}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCC-B |
| poly(dC)$^{G16}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCCCCCCCCC-B |
| poly(dC)$^{T18}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCCCCCCCCC-B |
| poly(dC)$^{G18}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCCCCCCCCC-B |
| Heteropolymer-C | ACTACCTAGTTTACGTAATCCATCTGCACAATGCAGCATT-B |
| Heteropolymer-T | ACTACCTAGTTTACGTAATCCATCTGTACAATGCAGCATT-B |
| Heteropolymer-A | ACTACCTAGTTTACGTAATCCATCTGAACAATGCAGCATT-B |
| Heteropolymer-G | ACTACCTAGTTTACGTAATCCATCTGGACAATGCAGCATT-B |
| poly(dC)$^{T9A14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCTCCCCCCCC-B |
| poly(dC)$^{A9A14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCACCCCCCCC-B |
| poly(dC)$^{G9A14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCGCCCCCCCC-B |
| poly(dC)$^{T9G14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCTCCCCCCCC-B |
| poly(dC)$^{A9G14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCACCCCCCCC-B |
| poly(dC)$^{G9G14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCCCCGCCCCCCCC-B |
| poly(dC)$^{T9T14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCTCCCCCCCC-B |
| poly(dC)$^{A9T14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCACCCCCCCC-B |
| poly(dC)$^{G9T14}$ | CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCCCGCCCCCCCC-B |