Reading DNA with PNA: A Dynamic Chemical Approach to DNA Sequence Analysis

Frank R. Bowler

University of Edinburgh

Doctor of Philosophy

March 2011

ABSTRACT

Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) constitute important sources of genetic variation which provide insight into disease aetiology and idiosyncratic differences in drug response. The analysis of such genetic variation relies upon the generation of allele-specific products, typically by enzymatic extension or the hybridization of allele-specific DNA probes. Herein, a distinct enzyme-free, dynamic chemistry-based method of producing allele-specific products for genotyping was developed. The approach was initially demonstrated in model systems using synthetic DNA, which was used as a template in a base-filling reductive amination reaction on a PNA backbone. The templated dynamic reaction between a free secondary amine at a 'blank' position on the PNA strand and four aldehyde-modified nucleobases drove selective formation of the 'correct' iminium intermediate according to Watson-Crick base-pairing rules. In a blind trial, the method was extended to genotype twelve cystic fibrosis patients for two mutations (one SNP and one indel) linked to this disease. Enzyme-free dynamic chemistry thus permitted successful genotyping in both singleplex and duplex formats, demonstrating the application of dynamic chemistry as a distinct method of allele-discrimination with certain advantages over those reported previously. The application of this method as a tool for the discovery of non-natural nucleobases with improved properties for antisense and genotyping applications was also investigated. Furthermore, progress was made towards the use of dynamic chemistry as a means of full nucleic acid sequence analysis, through the templated sequence-selective extension of PNA probes by reductive amination.

DECLARATION

This thesis has been composed by the author, and describes research carried out by the author under the supervision of Professor Mark Bradley at the University of Edinburgh. Where work has been performed either jointly or wholly by others, this is clearly attributed. No part of this thesis has been previously submitted for any other degree or professional qualification.

Signed:

Date:

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor, Professor Mark Bradley, as it has been a great privilege to be part of the world-class research group which Mark has created at the University of Edinburgh. I am hugely grateful for the support and encouragement he has provided during my time here, and for the world of exciting opportunities which he has opened up to me.

I would also like to thank my friend and colleague Dr Juan Jose Diaz-Mochon for everything he has done for me during my time in Edinburgh. It has been one of the greatest pleasures in my life so far to work with someone of such boundless imagination and enthusiasm, and I could never have wished for a better workmate.

I would never have been able to complete my thesis without the herculean efforts of my colleagues in the Bradley Group. They have helped and supported me in my studies on countless occasions and I will be eternally grateful to them, especially my friends Adam and Juanma who have made me laugh on an almost daily basis during my time here in Edinburgh. I must also acknowledge the important contribution of all of the staff in the School of Chemistry who supported me throughout my studies. Although they are too many to name individually (and I would be sure to miss someone out), I particularly owe thanks to Dr Juraj Bella for the countless times he has helped me with my NMR queries. A special thank you is owed to Dr Andrew Cronshaw in the School of Biological Sciences for all of the expert help he has provided during my use of the MALDI-TOF MS facility. I would also like to thank Professor Duncan Graham at the University of Styrathclyde for allowing me to perform my T_m measurments in his labs, and for the kind and generous assistance of his group members, especially Dr Jennifer Dougan, Dr David Thompson and Dr Fiona McKenzie.

Finally, I thank my family; my parents, Jean and John, for their love and support and for teaching me the importance of asking questions, my brother Max for being such an ace human being, and my parents-in-law Jean and John for their kindness to me over the years. And it would all be pointless without my wife, Jennifer, whose love has kept the stars shining; this thesis is dedicated to you.

ABBREVIATIONS

A	adenine
Ac	acyl
AcOH	acetic acid
approx.	approximately
APS	adenosine 5'-phosphosulfate
aq	aqueous
ARMS	amplification refractory mutation system
ASO	allele-specific oligonucleotide
ATP	adenosine 5'-triphosphate
B	biotin
BAC	bacterial artificial chromosome
Bhoc	benzhydryloxycarbonyl
Boc	*tert*-butyloxycarbonyl
bp	base pair(s)
br s	broad singlet (NMR assignment)
C	cytosine
calcd	calculated
cat.	catalytic
Cbz	carboxybenzyl
CCD	charge-coupled device
CF	cystic fibrosis
CNP	copy number polymorphism
CNV	copy number variant
conc.	concentrated
CRT	cyclic reversible termination
CV	column volume(s)
D	2,6-diaminopurine
d	doublet (NMR assignment)
DCC	dynamic combinatorial chemistry *or* *N,N'*-dicyclohexylcarbodiimide

DCL	dynamic combinatorial libraries
DCM	dichloromethane
dd	doublet of doublets (NMR assignment)
Dde	1-(4,4-dimethyl-2,6-dioxacyclohexylidene)ethyl
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
ddATP	2',3'-dideoxyadenosine 5'-triphosphate
ddCTP	2',3'-dideoxycytidine 5'-triphosphate
ddGTP	2',3'-dideoxyguanosine 5'-triphosphate
ddNTP	2',3'-dideoxynucleoside 5'-triphosphate
ddTTP	2',3'-dideoxythymidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
DIBAL-H	diisobutylaluminium hydride
DIC	*N,N'*-diisopropylcarbodiimide
DiPEA	*N,N*-diisopropylethylamine
DMAP	4-dimethylaminopyridine
DMF	*N,N*-dimethylformamide
DMSO	dimethylsulfoxide
DNA	2'-deoxyribonucleic acid
dNTP	2'-deoxynucleoside 5'-triphosphate
dq	doublet of quartets (NMR assignment)
dsDNA	double-stranded 2'-deoxyribonucleic acid
dTTP	2'-deoxythymidine 5'- triphosphate
EDC	*N*-(3-dimethylaminopropyl)-*N'*-ethylcarbodiimide
EDTA	ethylenediaminetetraacetic acid
EI	electron impact
ELSD	evaporative light scattering device
eq	equivalent(s)
ES	electrospray
Et	ethyl
F	fluorophore
FA	formic acid

FEN	flap endonuclease
Fmoc	fluorenylmethoxycarbonyl
FRET	Förster resonance energy transfer
fwt	formula weight
G	guanine
Gb	Gigabases (i.e. billion bases)
GWA	genome wide associtaion
h	hour(s)
HGP	human genome project
HOBt	1-hydroxybenzotriazole
HPLC	high performance liquid chromatography
HRMS	high resolution mass spectrometry
I	relative peak intensity
IR	infra red
J	NMR coupling constant (preceding superscript number denotes the number of bonds separating coupled nuclei)
kb	kilobases (i.e. thousand bases)
LCMS	liquid chromatography mass spectrometry
LNA	locked nucleic acid
m	multiplet (NMR assignment) *or* medium (IR description)
MAF	minor allele frequency
MALDI	matrix-assisted laser desorption/ionization
Mb	megabases
Me	methyl
min	minute(s)
Mmt	monomethoxytrityl
mp	melting point
mRNA	messenger RNA
MS	mass spectrometry
m/z	mass-to-charge ratio
NEM	*N*-ethylmorpholine
NMP	*N*-methyl-2-pyrrolidone

NMR	nuclear magnetic resonance
nt	nucleotide
NUDGE	nucleotide depletion genotyping
oxyma	ethyl 2-cyano-2-(hydroxyimino)acetate
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PG	protecting group
PNA	peptide nucleic acid
PP_i	pyrophosphate
ppm	parts per million
PyBOP	(benzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate
PS	polystyrene
Q	quencher
q	quartet (NMR assignment)
RNA	ribonucleic acid
R_f	thin layer chromatography retention factor (R_f = (distance travelled by analyte) / (distance travelled by solvent))
rRNA	ribosomal RNA
RT	room temperature
s	singlet (NMR assignment) *or* strong (IR description)
SAMRS	self-avoiding molecular-recognition system
SAP	shrimp alkaline phosphatase
sat.	saturated
SBS	sequencing by synthesis
SISAR	serial invasive signal amplification reaction
SNP	single nucleotide polymorphism
SPC	solid-phase capture
SPE	solid-phase extraction
SPPS	solid-phase peptide synthesis

ssDNA	single-stranded 2'-deoxyribonucleic acid
STS	sequence tagged site
T	thymine
T*	2-pyridone
t	triplet (NMR assignment)
TAPS	*N*-tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid
Taq	*Thermus aquaticus*
TBTU	*O*-(benzotriazol-1-yl)-*N*,*N*,*N*′,*N*′-tetramethyluronium tetrafluoroborate
TFA	trifluoroacetic acid
THF	tetrahydrofuran
TIS	triisopropylsilane
TLC	thin layer chromatography
T_m	duplex melting temperature
TOF	time-of-flight
t_R	retention time
Tris	tris(hydroxymethyl)aminomethane
tRNA	transfer RNA
U	uracil
UV	ultraviolet
VSET	very short extension
v/v	volume/volume ratio
w	weak (IR description)
WGA	whole genome association
WGS	whole genome shotgun
w/v	weight/volume ratio
X	xanthine
δ	chemical shift
ν	frequency

TABLE OF CONTENTS

CHAPTER 1
Introduction

1.1 DNA Sequencing and SNP Analysis

1.1.1 Nucleic Acids

The nucleic acids DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) are linear heteropolymers of four different monomers, called nucleotides. Nucleotides are composed of a sugar (β-D-ribose in RNA, β-D-deoxyribose in DNA), one or more phosphate groups and a heterocyclic base (Figure 1.1). There are 5 commonly occurring bases in nature, two of which are derivatives of purine, namely adenine (A) and guanine (G), and three of pyrimidine, namely cytosine (C), thymine (T, present in DNA only), and uracil (U, present in RNA only).[1]

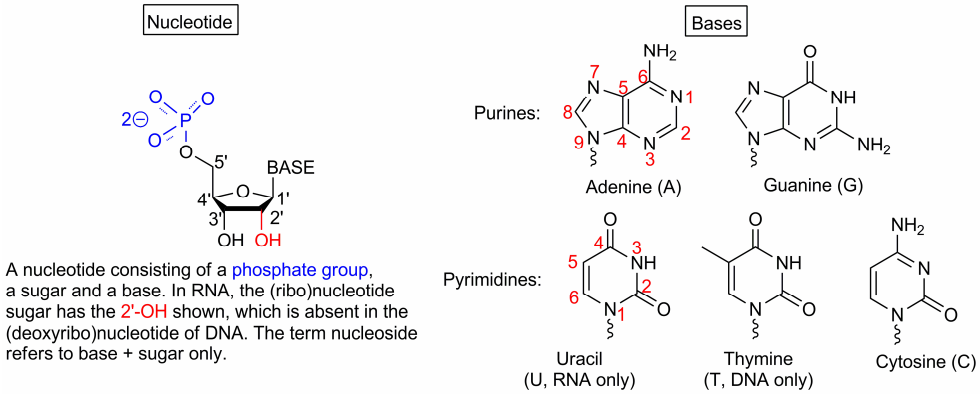


Figure 1.1 Nucleic acid building blocks. A nucleotide consists of between one (shown above) and three phosphate groups, a sugar and a base. In RNA, the (ribo)nucleotide sugar has the 2'-OH shown, which is absent in the (deoxyribo)nucleotide of DNA. A base bound to a sugar without a phosphate group is termed a 'nucleoside'.

The nucleotide monomers are linked by 3'→5' phosphodiester bonds to form the polymer. The famous double-helical structure of DNA, inferred by Watson and Crick in 1953 from (amongst other data) the X-ray diffraction patterns obtained by Franklin and Wilkins, arises from specific hydrogen bonding interactions between the base-pairs (Figure 1.2).[2] G forms three hydrogen-bonds to C, and A forms two hydrogen-bonds to T (in DNA) or U (in RNA). In the DNA double-helix, two anti-

parallel polynucleotide chains are coiled around a common axis, with the sugar-phosphate backbones on the outside and the bases on the inside.[1]

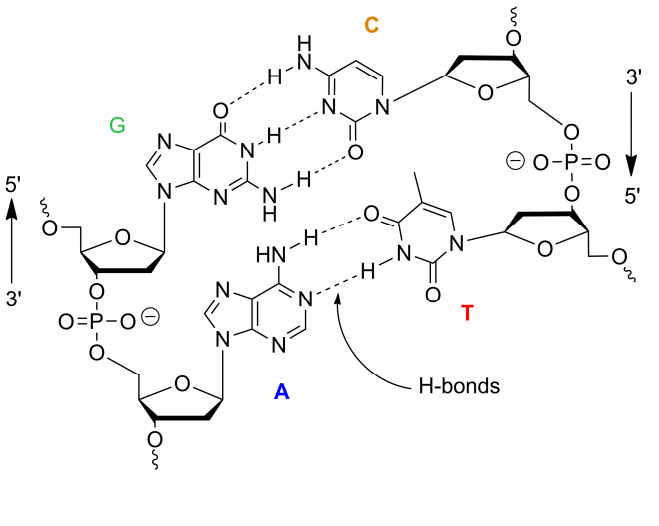


Figure 1.2 Base-pairing and hydrogen-bonding in the DNA double-helix.

The sequence of bases in a DNA or RNA strand uniquely characterizes the nucleic acid and represents a form of linear information. This heritable genetic information is stored as DNA in the nuclei of living cells and serves as a 'recipe' for protein production (a three base sequence or 'codon' codes for one amino acid in a protein). Expression of proteins by a cell involves transcription of the relevant sections of DNA into a class of RNA molecules called messenger RNA (mRNA), which then leave the nucleus and travel to the cell cytoplasm where they are translated into proteins. This translation process involves further RNA molecules called transfer RNA (tRNA) and ribosomal RNA (rRNA).[1]

During DNA replication, the helix is unwound and unzipped to expose the bases, and a DNA polymerase uses the original 'mother' strands as templates to prepare two new 'daughter' strands. The daughter strands are synthesized in a 5'→3' direction from an RNA or DNA primer (a short or 'oligonucleotide' strand complementary to the preceding section of DNA) using the four nucleoside triphosphates. The new DNA chain grows as the free 3'-OH undergoes nucleophilic attack on the innermost (α) phosphate of an incoming nucleotide with the concomitant release of pyrophosphate (PP_i). The result of such DNA replication is that two new DNA molecules are produced from the original, each containing one

mother strand and one daughter strand. The sequence of bases on each daughter strand is exactly determined by the sequence of its complementary mother strand, and it is this feature of DNA that enables the accurate transmission of hereditary information.[1]

1.1.2 Early Sequencing Methods

The elucidation of the structure of DNA in 1953 sparked a number of attempts to determine the sequence of bases in DNA molecules, but it wasn't until the 1970s that the forerunners of modern sequencing technologies were developed.[3] In 1975, the 'plus and minus' method of Sanger and Coulson was the first such method to be reported.[4] This technique relied upon polyacrylamide gel electrophoresis (PAGE) to fractionate DNA strands according to their size. A primer oligonucleotide (either synthetic or obtained from restriction enzyme digests) was hybridized to the DNA template to be sequenced then extended by DNA polymerase I (in the 5'→ 3' sense) in the presence of the four nucleoside triphosphates, one of which was radiolabelled with ^{32}P. Conditions were employed such that this chain extension was as non-synchronous and random as possible, thereby producing a population of all possible lengths of extended oligonucleotide. This mixture of oligonucleotides was then divided into eight aliquots for a further extension by DNA polymerase, but this time chain extension was terminated in each case either by supplying only three of the four nucleoside triphosphates (the 'minus' method) or only one of the four (the 'plus' method). Denaturing PAGE separation of the resulting mixtures in adjacent lanes of the same gel gave rise to bands which could be imaged by autoradiography (due to the presence of ^{32}P), and the relative positions of these bands enabled the base sequence to be determined.

The 'plus and minus' method allowed sequences of approximately 50 nucleotides to be deduced within a few days, and was applied in the sequencing of bacteriophage ϕX174 (a single-stranded circular DNA molecule of 5386 bases).[5] However, neither the 'plus' nor 'minus' method was entirely accurate which is why both had to be used in conjunction to obtain meaningful sequence data. The main problem encountered was that of sequencing homopolymer runs (i.e. multiple repeats

of a given nucleotide), as run lengths often had to be deduced (somewhat unreliably) from the distances between the gel bands.

An alternative, predominantly chemical method was reported by Maxam and Gilbert in 1977.[6] This involved radiolabelling a double-stranded DNA molecule at either the 3' or 5' ends with ^{32}P. The DNA could then be denatured and the two strands separated by PAGE and extracted for sequencing, or alternatively the molecule could be cut in two by a restriction enzyme and the two ends resolved by PAGE and isolated prior to sequencing. The DNA was then subjected to chemical reactions designed to first damage then remove a base from its sugar, with the result that the DNA backbone would cleave at that position. The 'damaging' reactions were limited in so far as they affected only one in every 50 to 100 bases along the DNA. This partial cleavage at each base gives rise to a range of radioactive fragments extending from the radiolabelled end to each of the cleaved positions. The DNA to be sequenced was subjected separately to four reactions: one cleaving at both purines but preferentially at adenine (A>G), one preferentially at guanine (G>A), one at both pyrimidines (C+T) and one at cytosine only (C). The purine-specific reactions involved methylation with dimethyl sulphate to weaken the glycosidic bonds prior to cleavage with alkali (the G>A reaction made use of the faster reaction rate of G methylation, and the A>G reaction utilized the fact that the glycosidic bond of methylated A is weaker than that of methylated G). The pyrimidine-specific reactions employed hydrazine to remove the bases before cleavage with piperdine (the C only reaction employed NaCl to preferentially suppress the removal of T). When the products of the four reactions were subjected to electrophoresis in adjacent lanes of a polyacrylamide gel and the gel imaged by autoradiography, a series of bands were produced, the pattern of which allowed determination of the DNA sequence.

The chemical method of Maxam and Gilbert permitted sequencing of at least 100 bases from the point of labelling, and was an improvement on Sanger's 'plus and minus' technique as it could be applied directly to double-stranded DNA. Furthermore, PAGE bands were produced for each base in a sequence (there are no problems with homopolymer runs), and the chemical treatment was readily controlled and optimized to yield an even distribution of labelled material across the

sequence. This chemical method found application in the sequencing of the simian virus SV40 in 1978.[7]

In 1977, Sanger reported an improved method of sequencing which would ultimately lead to him receiving a share in the 1980 Noble Prize in Chemistry (his second) together with Gilbert *"for their contributions concerning the determination of base sequences in nucleic acids"*.[8, 9] This technique utilized chain-terminating nucleotide analogues whose incorporation into a growing DNA strand by DNA polymerase would prevent further chain extension. Both arabinoside triphosphates and 2',3'-dideoxynucleoside triphosphates could be employed (Figure 1.3), with the latter dideoxy method proving the more useful.

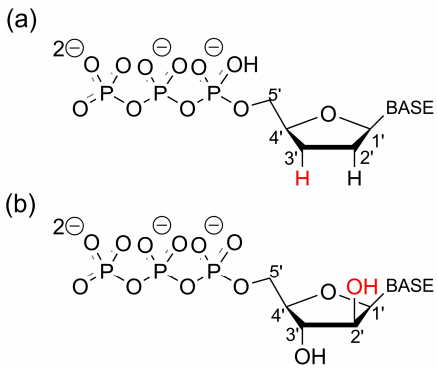


Figure 1.3 (a) Structure of 2',3'-dideoxynucleoside triphosphates, ddNTP. (b) Structure of arabinoside triphosphates. Arabinose is a stereoisomer of ribose which serves as a chain-terminating inhibitor of *Escherichia coli* DNA polymerase I.

In the original incarnation of dideoxy 'Sanger sequencing', the single-stranded DNA template to be sequenced was annealed to a primer, then divided into four aliquots. Each aliquot contained all of the ingredients necessary for DNA replication: the four 2'-deoxynucleoside triphosphates (dNTP), DNA polymerase and buffer. The dATP that was used was radiolabelled with ^{32}P, although the technique was later improved by incorporating ^{35}S (which emits lower energy β particles than ^{32}P and gives sharper bands in the resulting autoradiograph) into the primer instead.[3] Also added to each aliquot was one of the 2',3'-dideoxynucleoside triphosphates (ddNTP) at a concentration such that it was only incorporated a fraction of the time. Incorporation of a ddNTP prevented further extension of the growing DNA chain by the polymerase, as there was no longer a free 3'-OH group available for nucleophilic

attack on the innermost (α) phosphate of an incoming nucleotide. Taking the reaction with ddCTP as an example (Figure 1.4a), the resulting mixture of oligonucleotides contained a range of molecules, each of which terminated following incorporation of a dideoxy cytosine derivative. PAGE of the four reaction mixtures in adjacent lanes gave rise to a series of bands which could be imaged by autoradiography and the sequence read (Figure 1.4b). In this way, read lengths of up to 500 nucleotides were possible.

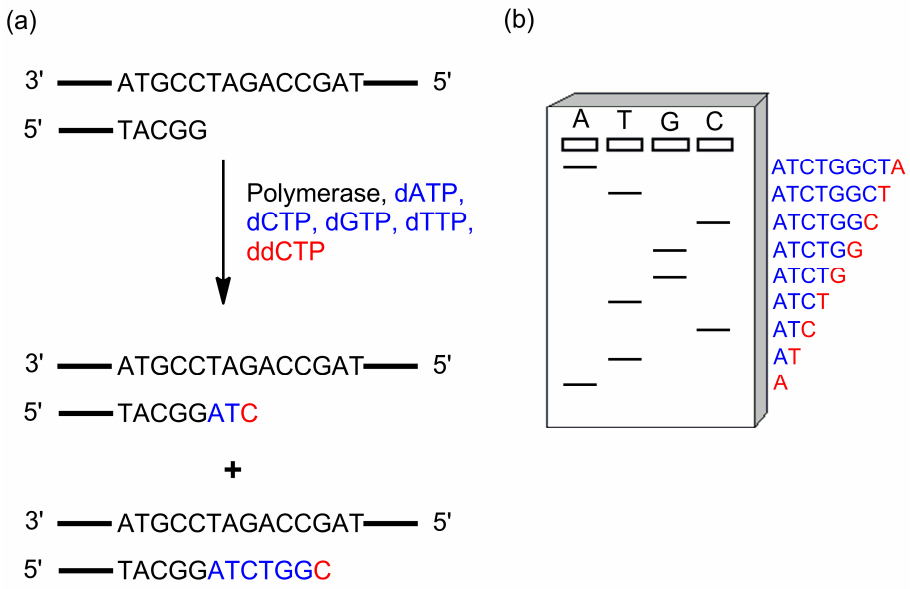


Figure 1.4 (a) Preparation of fragments chain-terminated at 'C'. (b) Analysis of the four dideoxy reactions by electrophoresis (adapted from Brückler, with permission).[10]

1.1.3 Sanger Sequencing and the Human Genome Project

Following its publication, the Sanger dideoxy method became the dominant DNA sequencing technique. Several improvements were made to the original method, such as the use of fluorescently labelled dideoxy nucleotides in place of radiolabelling (so-called 'four colour DNA sequencing' which uses lasers to read the fluorophore signals and facilitate single-lane analysis), the development of more efficient polymerases and the addition of capillary electrophoresis for the separation of labelled fragments. The pre-eminence of Sanger sequencing was boosted when Applied Biosystems (ABI) began to produce automated DNA sequencers which incorporated this approach.[3, 11] A drive towards high-throughput sequencing and

parallelism culminated in the launch of the ABI 3730*xl* DNA Analyzer. This 96-capillary machine is the state of the art technology for Sanger sequencing and is capable of generating up to 6 Mb (i.e. 6 million base pairs) of sequence data per day, with average read lengths of up to 1000 nucleotides (nt).[12]

Automated Sanger sequencers were employed in the sequencing of the human genome. The Human Genome Project (HGP) was established in 1990 and was a publicly funded international collaboration to sequence the ~ 3 billion bases of the human genome using samples of DNA taken from several individuals. The HGP employed a 'hierarchical shotgun sequencing' approach and began by digesting the genome with restriction enzymes (which recognize and cleave at specific sequences within a DNA molecule) into large fragments several hundred thousand base pairs in length. These fragments were subsequently cloned into 'bacterial artificial chromosomes' (BACs) and mapped onto the chromosomes of the human genome by identifying 'sequence tagged sites' (STSs), which are specific sequences whose location had already been identified. The BAC clones were then digested further ('shotgunned') to produce much smaller fragments of only a few hundred bases that were small enough to be subjected to automated Sanger sequencing. Computer algorithms then pieced together the sequence data using areas of fragment overlap, thereby elucidating the sequence of the DNA inserted into each BAC. As the location of each BAC in the genome had already been mapped, the sequence of the whole genome could be deduced.[13, 14]

The HGP published a working draft version of the human genome in February of 2001, but a second draft was published simultaneously by a company called Celera Genomics.[14, 15] Celera was founded in 1998 with the aim of sequencing the human genome, and its creation sparked a race with the HGP. This company employed a different strategy for sequencing called the 'whole genome shotgun' (WGS) approach, which skipped the BACs mapping stage and instead started by randomly fragmenting the genome directly into small fragments which could be cloned and sequenced (again with automated Sanger sequencers). The sequences obtained were then built up into larger 'scaffolds' by computer algorithms, and the positions of these scaffolds in the genome could be located through the identification of STSs. The entry of Celera onto the race to sequence the human genome brought

with it several methodological and technological innovations and ultimately speeded up the sequencing process.

1.1.4 Next Generation Sequencing Technologies

Sanger dideoxy sequencing has long been the dominant sequencing method, but despite numerous technological improvements to the original technique it remains costly, time-consuming and labour intensive to sequence large genomes like those of humans and other mammals. The sequencing of the diploid genome (i.e. both sets of chromosomes) of J. Craig Venter (co-founder of Celera) published in 2007 reputedly cost in the order of US $100 million.[16, 17] However, recent years have seen the emergence of so-called 'massively parallel' sequencers that are capable of generating many more sequence reads in a single experiment than the 96 generated by modern Sanger sequencers. These massively parallel sequencers typically employ a WGS approach to generate fragments of genomic DNA, but these are no longer cloned in *E. coli* or any other host cell prior to sequencing. One common limitation of these 'next generation' platforms is that the massively increased throughput comes at the expense of smaller individual read lengths and accuracies. This means that more reads are required, and that it is more difficult to piece together the reads into full genomic sequences. Thus Sanger sequencing may still be required for the more demanding applications, such as *de novo* sequencing of a large mammalian genome.[18]

The first such method to become commercially available was developed by 454 Life Sciences (now a subsidiary of Roche) and employed a pyrophosphate-based method termed 'pyrosequencing'. This approach begins by randomly fragmenting double-stranded DNA and ligating each fragment to adapter fragments at each end. The double stranded DNA is then separated into single strands, and these single-stranded fragments are bound to microbeads in an emulsion of water in oil. The concentration of beads and fragments is controlled such that only one fragment molecule is attached to each bead, and there is only one bead per water droplet. The water droplets then serve as microreactors in a polymerase chain reaction (PCR) step where the number of fragments per bead is amplified so that each bead carries millions of copies of a unique DNA template. These microbeads are subsequently

captured in picolitre-sized wells on a fibre-optic slide (one bead per well, ~ 1.6 million wells per slide), then many smaller microbeads carrying immobilized enzymes are added to each well. The actual sequencing of the fragments is achieved by adding a primer fragment and polymerase and washing the four nucleoside triphosphates in series over the plate of wells. Incorporation of the correct nucleotide by the DNA polymerase (Figure 1.5) releases pyrophosphate (PPi) which is used by a sulfurylase enzyme to generate adenosine 5'-triphosphate (ATP) from adenosine 5'-phosphosulfate (APS) present in the wash. This ATP is then used by a luciferase enzyme and luciferin to generate light which is detected by a charge-coupled device (CCD) sensor capable of capturing the photons emitted from the bottom of each individual well.[19] In this way the 454 pyrosequencer is able to sequence each DNA fragment by detecting the light emitted on incorporation of a base into a growing strand.

The most advanced platform produced by 454 (the Genome Sequencer FLX Titanium) can reportedly generate around 750 Mb of sequence data per day, with an average read length of ~ 400 bases.[12] This compares favorably with the 6 Mb generated by a Sanger sequencer over the same period. However, this methodology does experience problems when sequencing homopolymer runs. Although there is a linear relationship between the intensity of light generated and the number of bases incorporated in a single wash, this linearity falls off at longer homopolymer runs with the result that errors can occur during the sequencing of such runs.

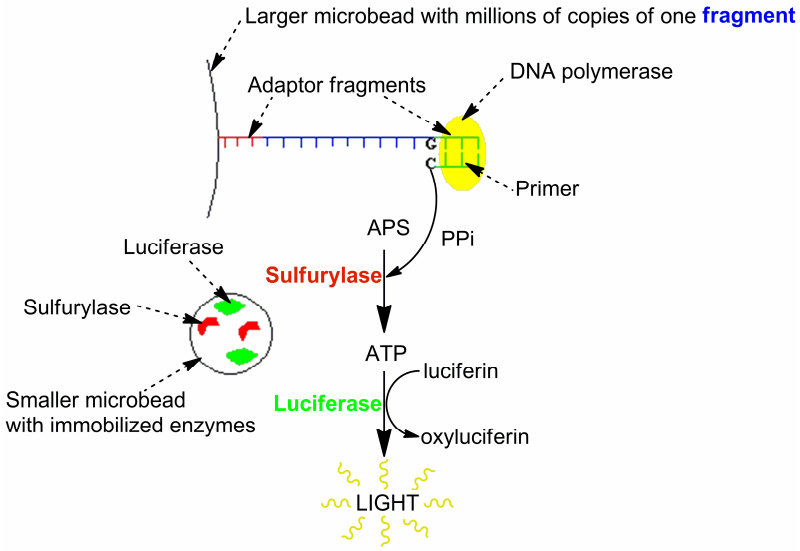


Figure 1.5 Generation of light during base incorporation in the pyrosequencer.

The second of the next generation platforms was launched by Solexa (now owned by Illumina) and takes a cyclic reversible termination (CRT) approach (Figure 1.6).[20] As for the pyrosequencer, the genomic DNA to be sequenced is isolated, fragmented, ligated to adaptor fragments and separated into single strands (Figure 1.7). The Illumina Genome Analyzer then requires that these single strands are bound to the surface of a glass flow cell which is coated with a dense lawn of adaptor fragments. The free end of a bound fragment can then bridge and hybridize to an adaptor fragment on the surface, which serves as a primer for a subsequent PCR amplification of the bound fragments. This amplification produces PCR colonies or 'polonies' of cloned fragments.[21] These clusters can then be sequenced by supplying a DNA polymerase and four differentially labelled fluorescent nucleotides which have been chemically blocked at the 3'-OH position such that only one base can be incorporated. This base incorporation step is followed by an imaging stage with lasers to identify the base incorporated at each polony, then a chemical deblocking step removes the fluorescent marker and liberates a free 3'-OH at each strand ready for the next sequencing cycle. Cycles of base incorporation, imaging and fluorescent marker removal/deblocking allow the sequence of the DNA at each cluster to be read.

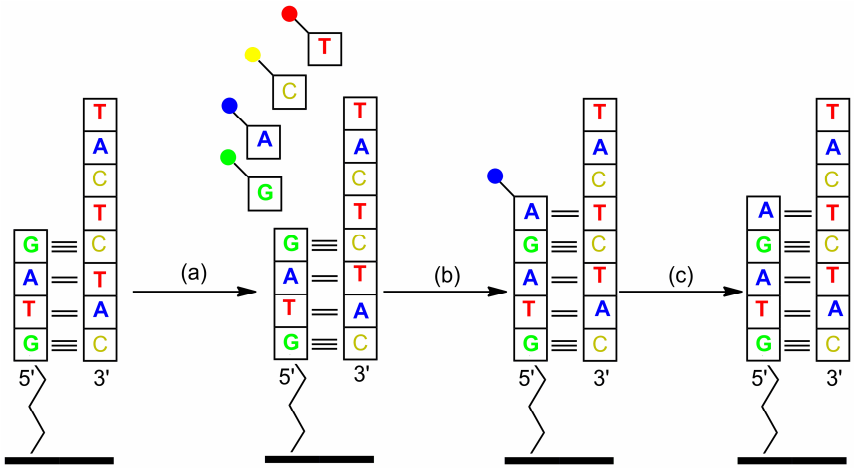


Figure 1.6 The CRT approach to sequencing. Either the primer (as shown) or template strand is attached to a surface. (a) Probing with reversibly terminating (i.e. blocked 3'-OH) fluorescently labelled nucleotides; (b) only the complementary nucleotide (according to Watson-Crick base-pairing) is incorporated into the growing strand by a DNA polymerase; (c) cleavage of fluorophore and 3'-OH block ready for next cycle (adapted from Brückler, with permission).[10]

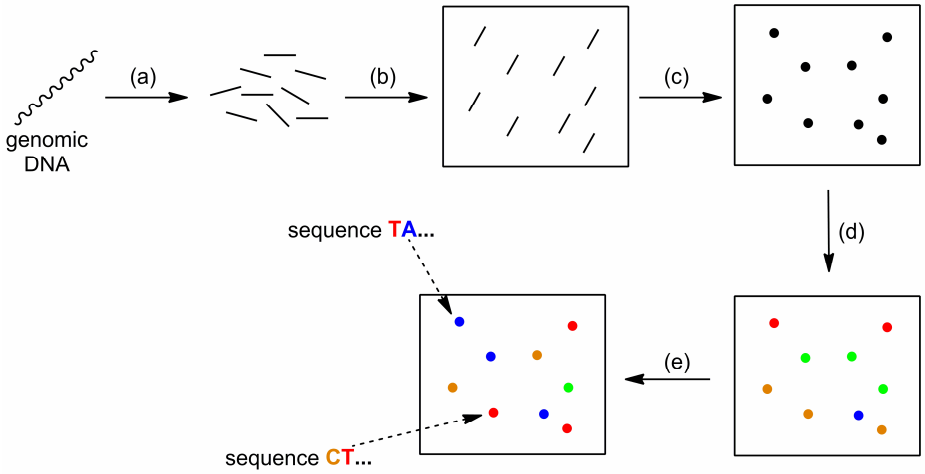


Figure 1.7 Sequencing with the Illumina Genome Analyzer. (a) Random fragmentation, ligation to adaptor fragments; (b) attachment to slide with lawn of adaptors/primers; (c) bridged PCR amplification to generate clusters or 'polonies' of identical templates; (d) chain extension with 3'-OH blocked, fluorescently labelled dATP, dCTP, dTTP, dGTP and imaging; (e) deblocking/removal of fluorophore and cycle repeat.

Both the 454 and Illumina approaches are often termed 'sequencing by synthesis' (SBS) as they involve the identification of a base immediately after its incorporation into a growing DNA strand. Unlike the 454 pyrosequencer system, however, Illumina's CRT method has no problems with homopolymer runs. Furthermore, higher throughput is possible as the platform can reportedly generate up to 5 Gb of sequence information per day, and an improved version (the 'HiSeq 2000' instrument) promising even greater throughput has recently been launched.[12] However, one drawback is that individual read lengths are much smaller (100 nt). This is because the reversibly terminating dye-labelled nucleotides are not incorporated efficiently by the modified polymerase that is employed. This enzyme also generates more base-substitution errors than is observed for the pyrosequencing approach.[3, 12]

A more recently developed massively parallel sequencing platform is Applied Biosystems' Supported Oligonucleotide Ligation and Detection (SOLiD™) system.[3, 12, 20] This technique applies the specificity of a DNA ligase to sequencing. Initially the SOLiD™ approach is much like that for 454 pyrosequencing, in so far as the genome is fragmented randomly, ligated to adaptor fragments then attached to

microbeads and subjected to emulsion PCR so that each bead has millions of copies of a unique DNA template bound to the surface. However, at 1 μm in diameter these beads are much smaller than those employed in the pyrosequencer (26 μm). The templates on the beads are chemically modified at their free 3'-OH for attachment to a glass slide. The smaller size of these beads and their subsequent random attachment to the slide mean that the SOLiD™ platform is capable of much higher throughput than the Illumina Genome Analyzer. For sequencing, universal primers are annealed to the adaptor at the 5'-end of the bound fragments and a set of semi-degenerate 8mer oligonucleotides are added along with a DNA ligase. The 8mer oligonucleotides are labelled with one of four fluorescent tags at their 5'-end which identifies two adjacent bases in the sequence (so-called 'two base encoding'). If these two bases correspond to the template sequence (according to Watson-Crick base-pairing), then the oligonucleotide hybridizes to the template and a DNA ligase seals the backbone of the growing strand. After imaging to determine the colour of the attached fluorophore, the newly incorporated oligonucleotide is cleaved to remove the fluorescent marker and the process is repeated several times. Finally, the new strand is removed by denaturation and washed away, and the whole process is repeated several times with primers of shorter length. This allows the sequence of the template to be built up.

The SOLiD™ sequencer holds an advantage over earlier next generation platforms in that its two base encoding approach facilitates the checking of base-calling errors.[18] The throughput of these instruments is comparable to the Illumina Genome Analyzer, at 5 Gb per day, although only shorter read lengths of 25-75 nt are possible.[12] It should be noted that a cheaper version of the SOLiD™ system is also available, namely the 'Polonator' developed by Dover Systems together with Church and colleagues of the Harvard Medical School. A similar chemistry has also been applied by Complete Genomics to the sequencing of arrays of so-called 'DNA Nanoballs' (DNB™) produced by rolling circle amplification of single-stranded DNA circles generated from genomic DNA.[22]

Most recently, a single-molecule DNA sequencer was launched by Helicos BioSciences (the HeliScope™) which applies an SBS approach to individual DNA molecules.[23] This technique removes the need for PCR amplification of DNA

templates (which can introduce errors, template biases and extra cost) because sequencing is performed on individual DNA molecules as opposed to clusters or bead colonies. The application of single-molecule imaging promised a step-change in sequencing, but the high cost (around $1 million) of the HeliScope™ has limited its uptake, and read lengths (32 nt) are shorter than is possible with other instruments.[24] Throughput is comparable to the Genome Analyzer and SOLiD™ platforms, although error rates are reportedly higher.[12]

When paired with advances in the field of bioinformatics and computing, the next or 'second' generation massively parallel technologies outlined above have greatly reduced the time, labour and cost associated with sequencing. In 2008, the diploid genome of James Watson was sequenced with the aid of the 454 pyrosequencer at a reputed cost of less than US $1.5 million in 4 months.[17, 25] This is substantially less than the US $100 million and 4 years required to sequence the diploid genome of J. Craig Venter using state of the art Sanger platforms, although with greater read lengths and lower error rates the Venter genome is arguably of better quality.[16] Three human genomes were recently reported to have been sequenced using the Complete Genomics system (which has yet to be made commercially available and is instead used to provide a sequencing service by this company) with average sequencing consumables costs of under $4400. However, no mention was made of the presumably substantial costs associated with the sequencing platform, data management (high-throughput sequencing generates terabytes of data files which must be handled, stored and analysed), and personnel required to perform the analysis.[12] The ultimate goal for sequencing has become the US $1000 genome, so despite the advances brought by these new technologies there is still some way to go. The ability to sequence a human genome for US $1000 or less would open the gateway to an era of genomic and personalized medicine, where patient treatment is informed and directed by the individual's genome or perhaps the genomes of disease-causing pathogens.[3]

However, further improvements in sequencing may not be long in the pipeline. A 'third' generation of platforms is currently in development, which promises to achieve ultrafast sequencing with greater read lengths by combining single-molecule detection with real-time analysis (as opposed to the step-wise

reagent addition associated with current technologies). Examples include an instrument from Pacific Biosciences which makes use of zero-mode waveguide nanostructure arrays,[26, 27] and platforms that make use of nanopores for sequencing.[28-30] More recently, an instrument (developed by Ion Torrent) has been described that employs semi-conductors to analyse the real-time enzymatic incorporation of unmodified nucleotides for sequencing, thereby removing the need for optical sensing and (expensive) fluorophore-labelling.[31, 32]

1.1.5 Single Nucleotide Polymorphisms

A single nucleotide polymorphism (SNP) is a one-base position in the genome for which two or more alternative alleles (i.e. forms of a gene located at a specific position on a chromosome) are present at appreciable frequency (traditionally, at least 1%) in the human population.[14] DNA sequencing has revealed approximately 10 million common SNPs, defined as those with a minor allele frequency, or 'MAF', of 5 % or greater in the population.[33] SNPs account for much of the genetic variation between individuals, and occur on average once in every 300 base pairs.[34] However, it should be noted that SNPs are not the only source of genetic variation; a range of structural variations exist, such as insertion-deletions (or 'indels') of one or more bases, and copy number variants (CNVs; common CNVs may be referred to as copy number polymorphisms, or CNPs) in which a section of the genome $\geq$ 1 kb in size is present in differing numbers of repeats between individuals.[34, 35] It is estimated that structural variations constitute around 20 % of the genetic variations between individuals, and 70 % of the variant nucleotides.[35] Further genetic variation between individuals arises from so-called epigenetic modification of DNA, such as the methylation and hydroxymethylation of cytosine.[36-38]

SNPs are not evenly distributed throughout the genome, and are more common in noncoding regions than in protein encoding regions. A SNP occurring in the regulatory site of a gene can alter the transcription rate of that gene, ultimately resulting in up- or down-regulation of the encoded protein. Similarly, a SNP in a protein encoding section can alter the amino acid sequence of the protein produced which can affect protein function and cause disease. So-called 'genome-wide association' (GWA) or 'whole genome association' (WGA) studies are performed

with the aim of linking certain SNPs (identified during sequencing and reported in public databases) with particular disease states, with a view to understanding disease aetiology and designing suitable therapies. Furthermore, in the field of pharmacogenomics attempts are made to link SNPs with an individual's response to particular drugs. In this way it is hoped that drug therapies can be tailored to individual patients, thereby ameliorating possible side-effects and heralding an era of personalized medicine.[39-42]

Of the myriad technologies available for SNP analysis, virtually all require a PCR amplification of the genomic DNA, incorporating, or prior to, some means of allele discrimination.[43] At present there exist four general approaches to allele discrimination; primer extension, hybridisation, ligation and cleavage.[44] Of these, methods based upon primer extension or the hybridisation of allele-specific oligonucleotide (ASO) probes have become the more widely adopted. Once allele-specific products have been generated, fluorescence detection and mass spectrometry are commonly used as read-out tools.

1.1.6 SNP Genotyping with Fluorescence Detection

The technologies applied to SNP analysis in large scale, genome-wide association (GWA) studies typically utilize fluorescence analysis on DNA microarrays (see Section 1.2.2).[40, 43, 45-47] A microarray is an ordered array of elements (i.e. collections of molecules in a microscopic spot) on a planar substrate such as glass, silicon or a polymer. The wide-ranging applications of microarrays stem from their facilitation of the automated and parallel testing and analysis of many samples in a single experiment.[48] Two of the most widely used microarrays for GWA studies are competing technologies developed by Affymetrix and Illumina.

The GeneChip system marketed by Affymetrix employs an allele-specific hybridisation approach (Figure 1.8) which makes use of differences in the thermal stability of matched and mismatched DNA strands to distinguish polymorphisms. Allele-specific 25mer oligonucleotide (ASO) probes are synthesized on a glass surface using photolithography to produce an ordered array. SNP-containing regions of an individual's genome are amplified, fragmented and tagged before they are hybridized on the probe array. Mismatched probes hybridize less strongly, so that

subsequent washing and fluorescent labelling steps reveal the SNP genotype based upon the probes that have hybridized to base-matched genomic DNA. The Affymetrix Genome-Wide Human SNP Array 6.0 has probes for 906,600 SNPs. One intrinsic limitation to this hybridization approach is that the relative stabilities of the matched and mismatched hybrids is dependent on the base sequences flanking the SNP position, which can place limits on the SNPs that can be readily identified. Optimization of each hybridization system is required, and reaction conditions such as temperature and ionic strength must be stringently controlled.

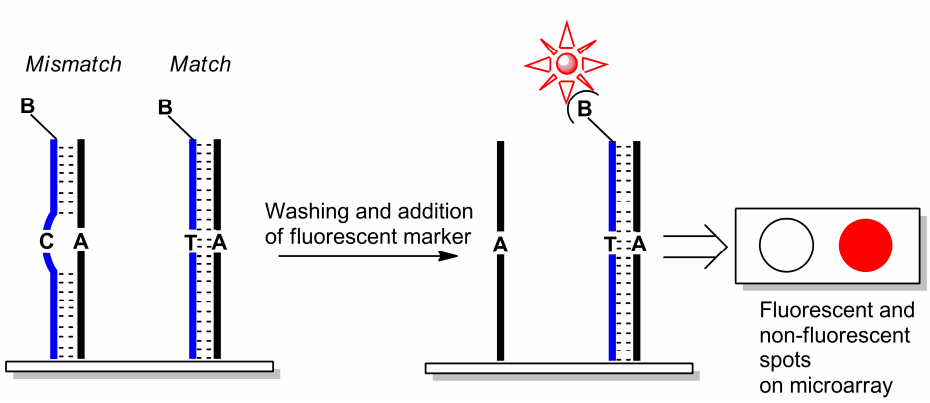


Figure 1.8 Hybridization approach to SNP genotyping, as employed in the Affymetrix GeneChip arrays. A single-base mismatch destabilizes the DNA duplex, so only fully matched target sequences will bind. In the GeneChip, the biotinylated ('B' = biotin) target sequences bind to fully matched probes on the array and are visualized using fluorophore-labelled streptavidin.

A different approach is employed by Illumina in their BeadArray platforms. In this case the microarray substrate is formed from dense bundles of 50,000 optical fibres that have wells etched into the tip to hold 3 μm microbeads. These microbeads each have hundreds of thousands of copies of a particular 50mer oligonucleotide probe covalently bound to their surface, providing specificity towards a particular SNP. The polymorphism detection can be based on either an allele-specific primer elongation as used in the Infinium I assay, or a single-base extension approach as employed in the Infinium II assay (Figure 1.9).[47, 49] In both cases the system relies upon the ability of a DNA polymerase enzyme to recognize base mismatches. An example of this technology is Illumina's Human1M-Duo BeadChip which interrogates more than 1.1 million SNPs per sample.

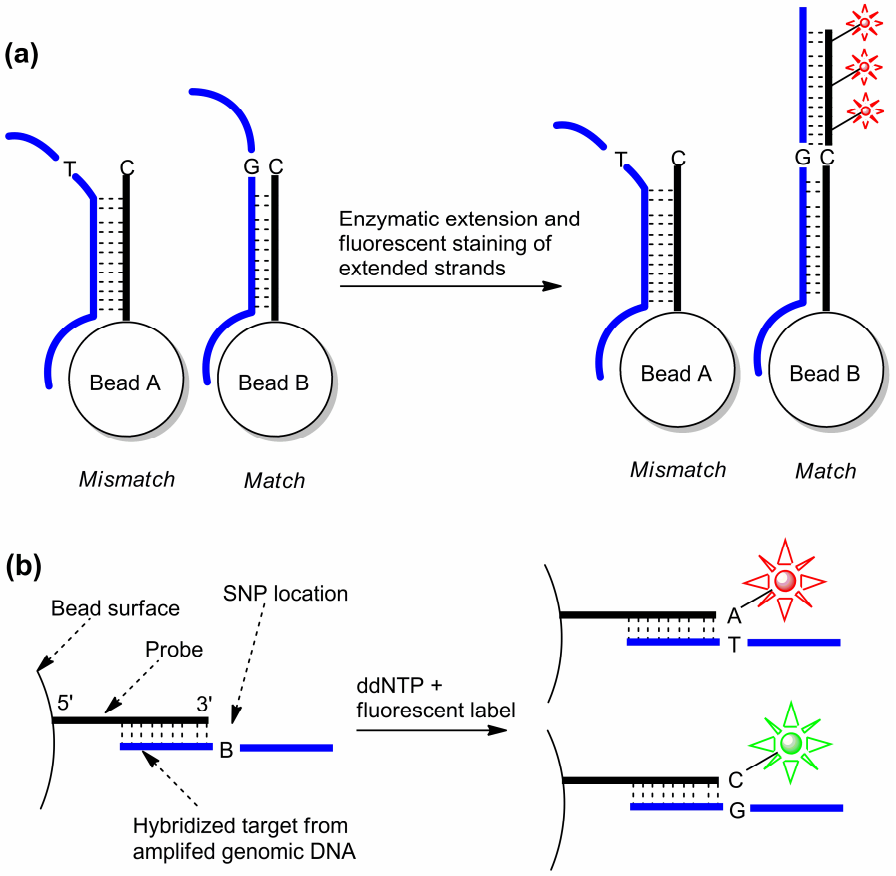


Figure 1.9 Genotyping by (a) allele-specific primer elongation, and (b) single-base extension, as employed in the Illumina Infinium I and II assays respectively.

Although these microarray-based platforms are useful for the high-throughput analysis of many thousands of SNPs simultaneously, a wide range of alternative fluorescence-based methods have been developed for more focused studies. These typically utilize real-time PCR, in which the increasing concentration of PCR amplicons is measured by a fluorescent read-out.[50] An advantage of such assays is that genotyping can be performed in a single closed tube, thereby minimizing sample handling and reducing the risk of contamination with exogenous DNA. A widely adopted example is the TaqMan® assay (Applied Biosystems), which makes use of the 5'-exonuclease activity of Taq DNA polymerase (Figure 1.10a).[51] As for the Affymetrix GeneChip, allele discrimination is achieved by hybridization. TaqMan® probes carry a fluorophore at one end and a quencher at the other, and the fluorescence is quenched by Förster Resonance Energy Transfer (FRET). Upon hybridization to a fully matched sequence, the probe is hydrolyzed by the polymerase

during PCR primer elongation, with the result that the fluorophore and quencher are separated. In this way FRET is eliminated and an increase in fluorescence is recorded. A single-base mismatch prevents stable hybridization with the amplified sequence and so reduces the efficiency of the hydrolytic probe cleavage.[52] A number of improvements to the design of the original TaqMan® probes have been made, including the use of LNA ('locked nucleic acid', Figure 1.10b; LNA probes are commercially available from Exiqon) which improves mismatch discrimination and allows the use of shorter probes as a result of the greater duplex stability associated with this RNA analogue (in which the 2'-O is connected to the 4'-C *via* a methylene bridge).[53, 54]

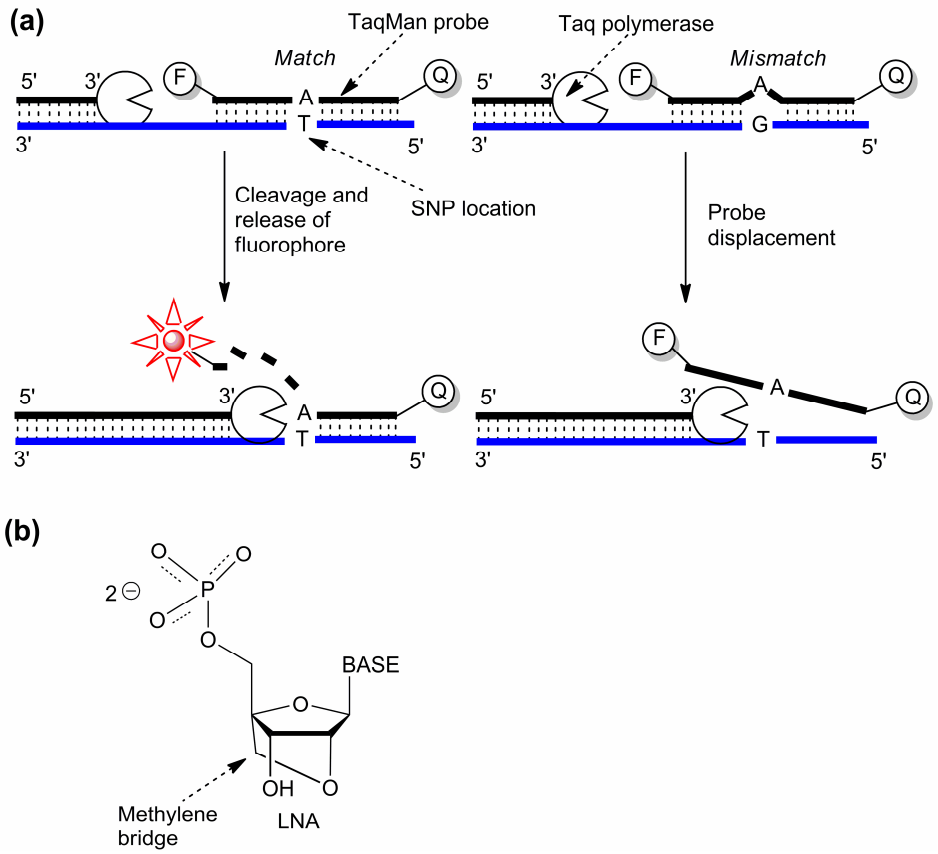


Figure 1.10 (a) Genotyping using TaqMan probes. A fully matched probe hybridizes and is hydrolyzed by the polymerase to generate a fluorescent signal, whilst a mismatched probe remains intact. (b) Structure of a locked nucleic acid nucleotide. 'F' = fluorophore, 'Q' = quencher.

Other examples of probe designs that rely upon hybridization for allele discrimination during real-time PCR have been developed, including so-called

Scorpion primers and HyBeacons. Like TaqMan probes, Scorpion primers incorporate a fluorophore and quencher.[55, 56] However, the probe sequence is tethered to a PCR primer (*via* a PCR-blocking linker to prevent copying of the probe sequence), such that enzymatic extension of the primer generates the target amplicon which then hybridizes to the probe. The distance between the donor and quencher is thus increased, thereby eliminating the collisional quenching and generating a fluorescent signal. The advantage of this approach is that the hybridization of the probe with the target is an intramolecular process, which gives a greater signal-to-noise ratio and rapid signal generation. Allele discrimination using Scorpion primers may be achieved by a so-called 'amplification refractory mutation system' (ARMS), using two Scorpions bearing different primer sequences (which cover the SNP location) and different fluorophores.[56, 57] Alternatively, a single Scorpion primer is employed wherein the SNP position is addressed thermodynamically using the probe sequence.[55] In the second case the presence of the mutant allele is indicated by a lower fluorescence signal recorded at a temperature close to the probe/target T_m (i.e. the duplex melting temperature; the temperature at which 50 % of the DNA molecules are hybridized in a double-helix). In their original incarnation, Scorpion primers utilized a stem-loop in the probe sequence for quenching, but this design was improved through the use of a separate oligonucleotide quencher (Figure 1.11a).[58] These 'duplex' Scorpions (supplied commercially by Qiagen) are less complex than the original stem-loop design, and additionally have the potential for improved fluorescence signals. This is because the distance between fluorophore and quencher is hugely increased upon target binding, and re-annealing with the quenching sequence is now a kinetically disfavoured intermolecular process as opposed to the intramolecular re-annealing that is possible with stem-loop architectures.

Unlike TaqMan probes and Scorpion primers, HyBeacons (Figure 1.11b and c; developed by LGC) do not possess secondary structure but instead make use of the enhanced fluorescence observed upon hybridization of linear probes functionalized with one or more fluorophores attached to internal nucleobases.[59-62] The intrinsic fluorescence-quenching properties of the nucleobases and fluorophores within the probes results in a low fluorescence in the unhybridized, randomly coiled state. This fluorescence is enhanced upon hybridization with a target sequence and formation of

a more rigid double-helix which holds the fluorophores further from the nucleobases and one another. Allele discrimination can be achieved by melting curve analysis, exploiting the greater T_m values for the fully matched *versus* single-base mismatched sequences.

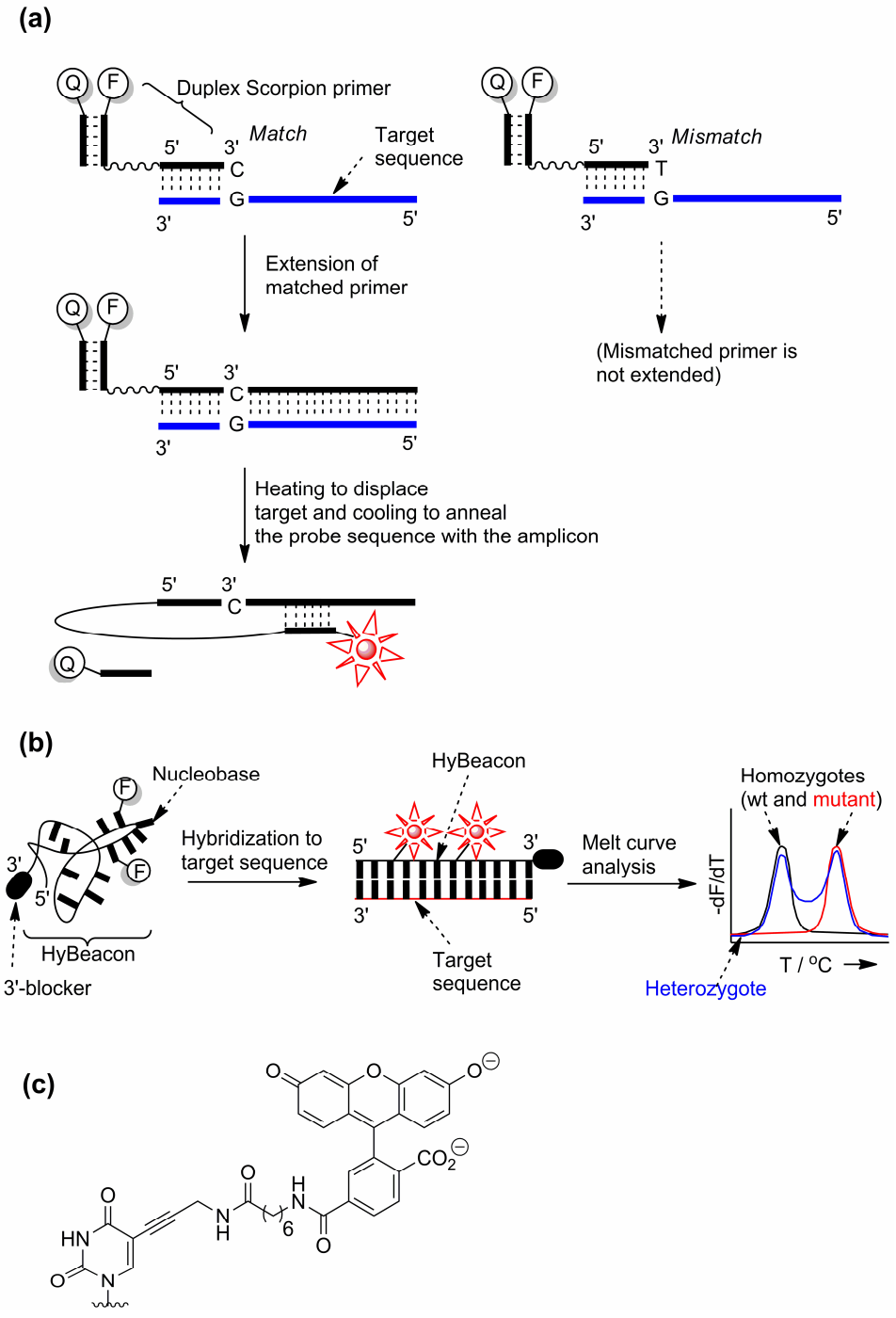


Figure 1.11 (a) Genotyping using duplex Scorpion probes, in this example incorporating the ARMS approach to allele discrimination. Only the fully matched primer is extended, and a fluorescent signal is generated following intramolecular hybridization of the probe with the

target amplicon and displacement of the quencher oligonucleotide. (b) HyBeacons generate a fluorescent signal upon hybridization to the target sequence. A 3'-blocker is present to prevent copying of the probe during PCR. (c) Structure of fluorescein (6-FAM)-modified thymine, as employed in HyBeacon probes. Eight different fluorophores have been demonstrated to be effective in HyBeacon probes, facilitating sample multiplexing.[62] 'F' = fluorophore, 'Q' = quencher.

An altogether different method of allele discrimination is employed in the Invader® assay.[63] This cleavage-based approach uses a flap endonuclease enzyme (FEN) together with a so-called upstream 'invader' and two allele-specific probes. The upstream invader is complementary to the region 3' of the SNP site, and terminates at its 3'-end with a nucleobase non-complementary to the SNP. The downstream allele-specific probes are complementary to the region 5' of the SNP, and differ in the nucleobase present at the SNP location. These downstream probes also bear a redundant, non-complementary 5'-arm. Upon hybridization of the invader and complementary probe, an overlap is produced that is recognized by the FEN, resulting in cleavage of the redundant 5'-arm of the probe. In the original design, this arm carried a fluorophore which formed part of a FRET system, such that cleavage eliminated FRET and generated a fluorescent signal. However, greater sensitivity is achieved if this cleaved product itself acts as an invader in a second FEN-mediated cleavage of a FRET incorporating probe (Figure 1.12).[64] In this (isothermal) serial invasive signal amplification reaction (SISAR), many thousands of fluorophores are generated by a single molecule of DNA target, enabling the direct detection of as few as 1000 target molecules (i.e. zeptomolar amounts of target) without prior (PCR) amplification. In practice, however, PCR-free analysis requires the input of a relatively large amount of genomic DNA and long reaction times, so a short PCR amplification is often employed to overcome these limitations. [65]

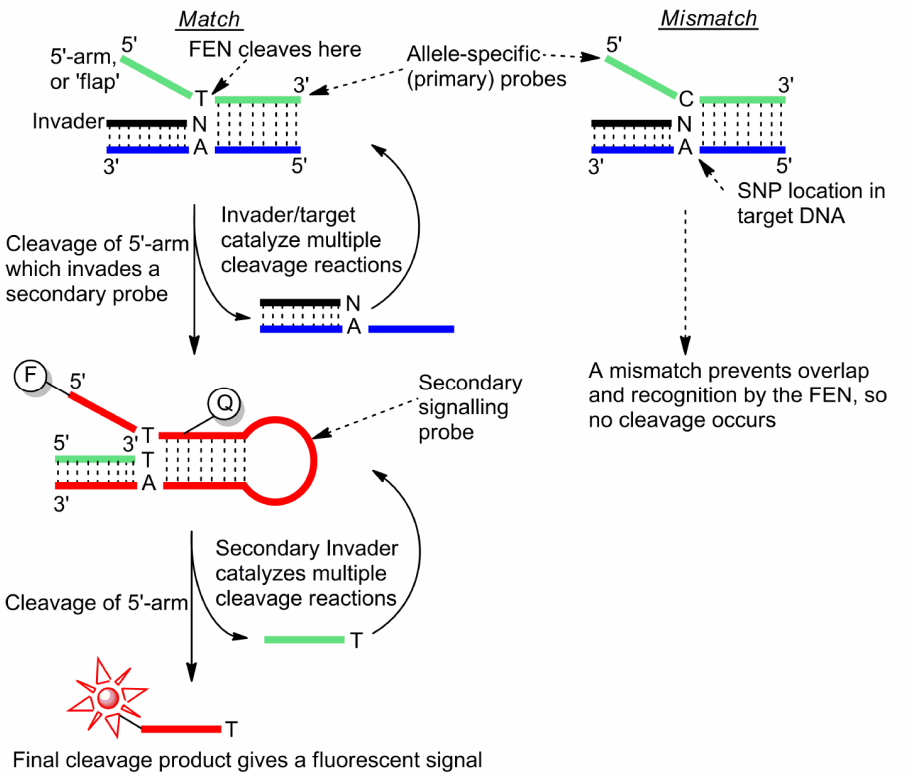


Figure 1.12 Genotyping using the Invader assay by serial signal amplification. 'F' = fluorophore, 'Q' = quencher.

1.1.7 SNP Genotyping with Mass Spectrometric Detection

A range of SNP genotyping methods has been developed that employ matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry, or MALDI-TOF MS, as a read-out tool.[66, 67] MALDI-TOF MS is routinely employed for the analysis of a range of biomolecules (e.g. peptides, proteins and nucleic acids).[68, 69] Briefly, laser energy is used to ionize the analyte which has been co-crystallized with a matrix of a small organic molecule, typically an acid. Different matrices are used depending upon the nature of the analyte. The matrix absorbs the energy of the laser and transfers some to the analyte, which is ionized. Since MALDI is a soft ionization technique it tends to produce little or no fragmentation. Although different analysis methods are possible, MALDI is usually coupled with time-of-flight (TOF) detectors that distinguish between ions of differing mass-to-charge (*m/z*) ratio by measuring the time it takes them to reach a detector after acceleration across a potential difference into a flight tube (ions with a larger *m/z* ratio take longer to reach the detector; the time-of-flight is directly proportional to the square root of *m/z*).

Following PCR amplification of genomic DNA, mass spectrometric SNP genotyping assays most commonly rely upon an enzymatic primer extension approach for allele-discrimination.[66] These methods routinely require an initial clean-up step of the amplified target DNA to remove unwanted dNTPs and PCR primers prior to single base extension. This can be done enzymatically using shrimp alkaline phosphatise (SAP) and exonuclease I to hydrolyze dNTPs and primers respectively, or alternatively by solid-phase capture of biotinylated amplicons on streptavidin functionalized beads followed by washing steps. Allele discrimination is performed by extension of a common primer which lies upstream of the SNP location, such that the 3'-end stops immediately before the SNP. Following allele discrimination, a further clean-up step is necessary to remove salt buffers and reaction components which can hamper analysis by MALDI-TOF MS. This is achieved by solid-phase purification on ion-exchange resins, streptavidin coated magnetic beads (if biotinylated primers are extended), or reverse-phase silica (as employed in ZipTip® pipette tips, supplied by Millipore), or alternatively by ethanol precipitation.

Allele discrimination by primer extension can involve a single base extension (SBE; Figure 1.13a) with dideoxynucleoside triphosphates (ddNTPs), as used in the PinPoint,[70] solid-phase capture (SPC)-SBE,[71] iPLEX® (Sequenom),[72] GenoSNIP™ [73] and GOOD assays.[74, 75] The GOOD assay is notable in that modified extension primers are used that possess a modified backbone at the 3'-end with alkylated methylphosphonate linkages, or alternatively phosphorothioate linkages which can be alkylated with methyl iodide in an additional step. Following extension, the rest of the primer can be selectively removed by enzymatic digestion with a phosphodiesterase. The remaining alkylated segment carries a charge tag and is readily detected by MALDI-TOF MS without further treatment. Furthermore, it has been reported that the initial SAP removal of dNTPs can be avoided through the use of an enzyme that selectively incorporates ddNTPs over dNTPs.[74, 76]

Unless charge-tagged ddNTPs are employed,[77] the mass discrimination for allele-specific SBE products can be difficult, especially for A/T polymorphisms that give rise to a mass difference of just 9 Da. One way to improve mass discrimination is through the use of mixtures of ddNTPs and dNTPs. Examples of such systems include the MassEXTEND® (Sequenom)[78] and very short extension (VSET)[79]

assays. A variation on this theme is the nucleotide depletion genotyping (NUDGE) assay that uses just 3 of the 4 dNTPs to obtain allele-specific products varying in length by more than one nucleotide.[80]

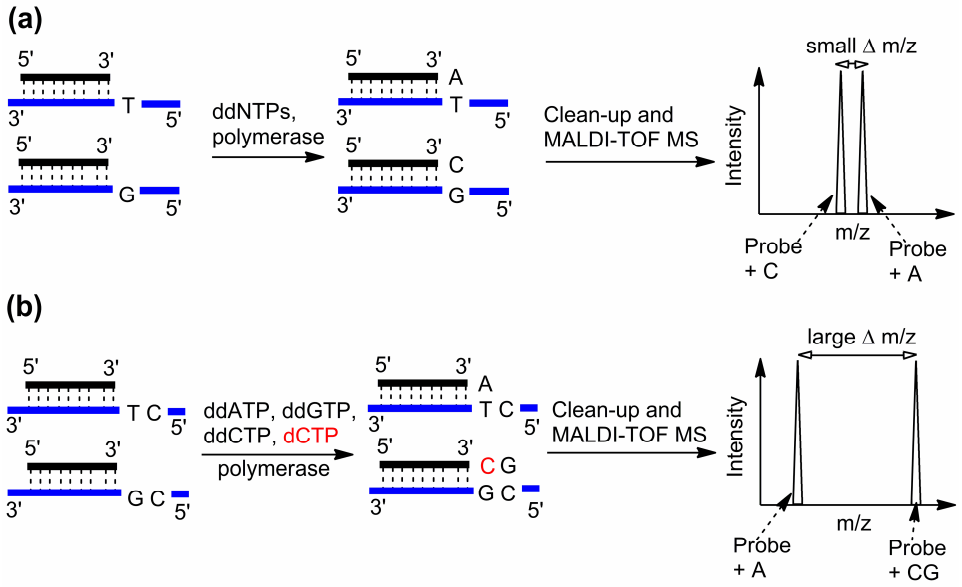


Figure 1.13 Approaches to allele discrimination for genotyping by MALDI-TOF MS. (a) Single base extension of a primer with ddNTPs generates products differing in mass by between 9 and 40 Da. (b) Greater mass discrimination is achieved by multiple base extension using a mixture of ddNTPs and dNTPs, giving products differing in mass by approximately 300 Da or more.

1.2 A Novel Chemical Approach to DNA Sequence Analysis

The research described in this thesis draws inspiration from the field of dynamic (combinatorial) chemistry and applies it to an artificial DNA mimic. These concepts will be outlined in this section.

1.2.1 Peptide Nucleic Acid

Peptide nucleic acid (PNA) is a structural mimic of DNA possessing an acyclic, achiral and uncharged pseudopeptide backbone in which the sugar-phosphate groups of DNA have been replaced with repeating *N*-(2-aminoethyl) glycine units.[81, 82] Purine and pyrimidine bases are attached to this backbone through methylene carbon linkages. PNA can hybridize efficiently to complementary PNA, DNA and RNA to form very stable duplexes according to Watson-Crick base pairing rules (Figure

1.14). PNA can hybridize in an antiparallel (the 3'-end of DNA aligned with the *N*-terminus of PNA) or parallel (3'-end aligned with the *C*-terminus) fashion, but the antiparallel orientation is strongly preferred.

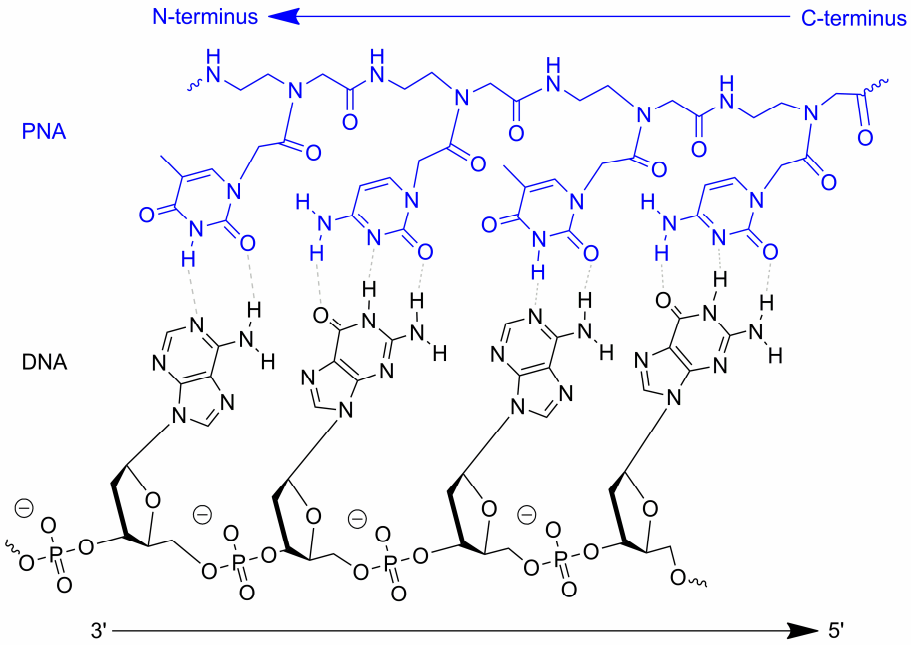


Figure 1.14 Structure of PNA and DNA showing antiparallel PNA/DNA hybridization.[83]

Under physiological conditions, the binding affinity and selectivity of PNA towards DNA, RNA and PNA is higher than for the analogous DNA duplexes. This has been attributed to the neutral character of PNA and the resulting lack of electrostatic repulsion with a negatively charged sugar-phosphate backbone. Furthermore, PNA has better discriminating power than DNA in that a base mismatch in a PNA/DNA duplex is more destabilizing than a mismatch in a DNA/DNA duplex. PNA oligomers are also resistant to enzymatic degradation by proteases and nucleases, which extends their lifetime *in vitro* and *in vivo*.

The pseudopeptidic nature of PNA means that straightforward solid-phase peptide synthesis (SPPS) techniques can be applied to their preparation.[83] The monomeric building blocks for PNA (Figure 1.15, analogous to the nucleotide monomers of DNA) must employ orthogonal protecting groups for the backbone and the exocyclic amines of the nucleobases (A, G and C).

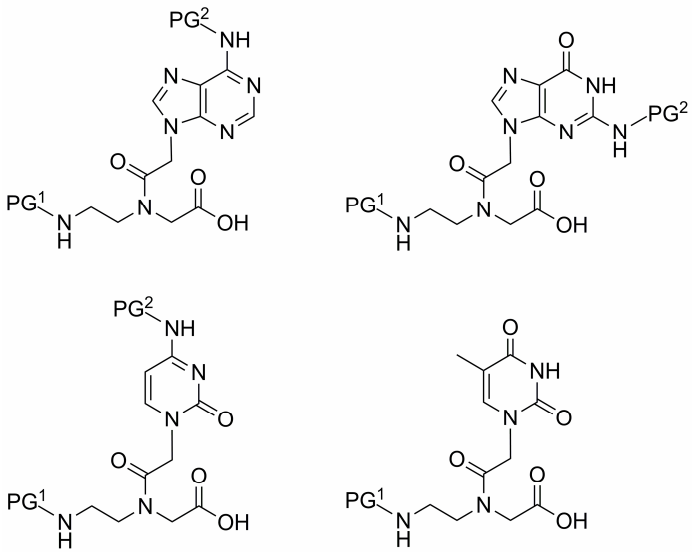


Figure 1.15 Monomers for PNA synthesis with orthogonal protecting groups, PG^1 and PG^2. A range of PG^1/PG^2 combinations have been successfully employed in the solid-phase synthesis of PNA oligomers, including Fmoc/Bhoc,[84] Dde/Mmt,[83] Fmoc/Mmt,[85] Boc/Cbz[86] and Fmoc/Cbz[87].

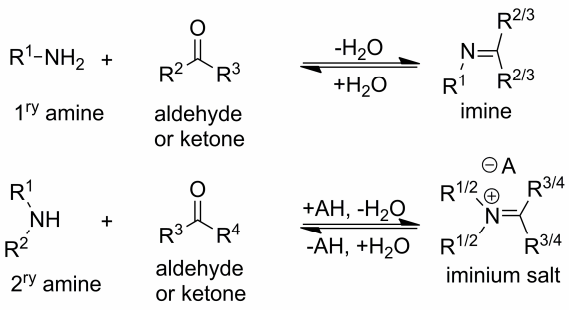
The unique chemical, physical and biological properties of PNA have stimulated attention at the interface of chemistry and biology, and a number of diagnostic and pharmaceutical applications have been sought such as gene therapy and mRNA profiling.[81, 88] Forays have even been made into the use of PNA for SNP analysis; a PNA mediated PCR clamping technique has been employed to prevent amplification of DNA sequences exactly complementary to a PNA probe,[89, 90] and allele-specific PNA probes have also been used in conjunction with MALDI-TOF MS as a tool for SNP genotyping.[91-98]

1.2.2 Dynamic Combinatorial Chemistry

Dynamic combinatorial chemistry (DCC) refers to the study of libraries (or 'dynamic combinatorial libraries', DCLs) of compounds in which all the constituent species are in thermodynamic equilibrium.[99] Library members are interconvertable through the formation of reversible covalent or noncovalent interactions, such that the library composition is determined by the thermodynamic stability of each of the library members. Those structures possessing the most favourable *inter* or *intra*molecular interactions will be more stable and will dominate over other library members. It is

therefore possible to add a template molecule to a DCL that will bind to a specific constituent and remove it from the 'pool', with the result that there is an 'equilibrium shift' to increase or 'replace' the concentration of that library member. As DCLs are under thermodynamic control, library composition can also be altered through variation of other factors such as temperature and pH.[100] DCC has found applications in areas such as drug discovery, catalyst screening and self-assembly of inorganic architectures.[101-103]

Numerous types of reversible reaction have been used to mediate the exchange of building blocks and interconversion of DCL members. These reversible reactions may involve noncovalent (e.g. hydrogen bonds), coordinative or covalent bonds. A pertinent example involving covalent bond formation is the reversible formation of imines (or 'Schiff bases') and iminium salts through the reaction of aldehydes or ketones with amines (Scheme 1.1).[99]



Scheme 1.1 Reversible formation of imines from primary amines, and iminium salts from secondary amines, by reaction with aldehydes or ketones.

1.2.3 A Novel Chemical Approach to SNP Analysis and DNA Sequencing

The strategy underpinning the research described in this thesis applies dynamic chemistry to peptide nucleic acid for the development of an entirely novel method of DNA sequence analysis. It was hypothesized that Watson-Crick base pairing could be harnessed to template a dynamic reaction on a strand of PNA, by creating a nucleobase-free position on the PNA (a so-called 'blank' position) situated opposite to a nucleotide under interrogation on a complementary DNA strand. A reversible reaction, between an aldehyde-modified nucleobases (Figure 1.16a) and the free secondary amine at the blank position of the PNA probe would generate an iminium

intermediate that could be reduced and analyzed (Figure 1.16b). In the presence of all four aldehyde modified nucleobases (i.e. adenine, thymine guanine and cytosine derivatives) the DNA template would stabilize the iminium species with the correct hydrogen-bonding motif (obeying Watson-Crick base-pairing), thereby amplifying it over the other possible iminium products. This would be reflected in the final product distribution after reduction to the corresponding amine, and would provide a means of allele-discrimination for SNP analysis.

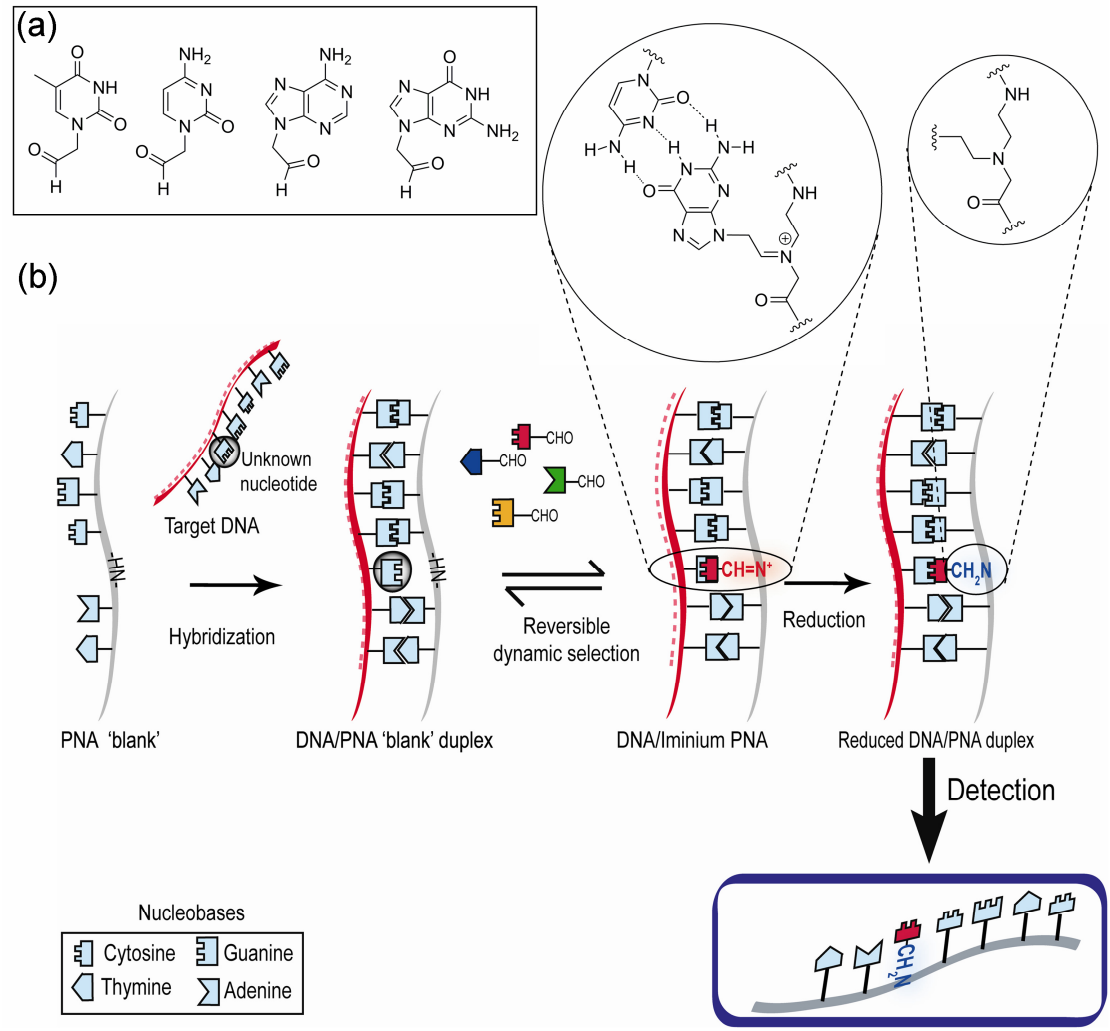


Figure 1.16 (a) Aldehyde modified nucleobases. These molecules could conceivably be labelled with fluorophores to facilitate fluorescent detection. (b) Sequence specific DNA-templated reductive amination on a PNA backbone possessing a 'blank' position constituting a free secondary amine.

The same principle could conceivably be applied to the DNA-templated extension of a PNA oligomer with a free (primary) amine terminus, using aldehyde-

modified PNA monomers (Figure 1.17a). In this way, SNP analysis could be performed by a single base extension (Figure 1.17b). The method could also be extended further to full DNA sequencing by cyclic reversible termination, if the free secondary amine in the backbone of the extended PNA probe was capped (e.g. with an acyl group) to prevent side-reaction, the terminal protecting group was removed to reveal a free primary amine, and the process of extension by templated reductive amination was repeated.

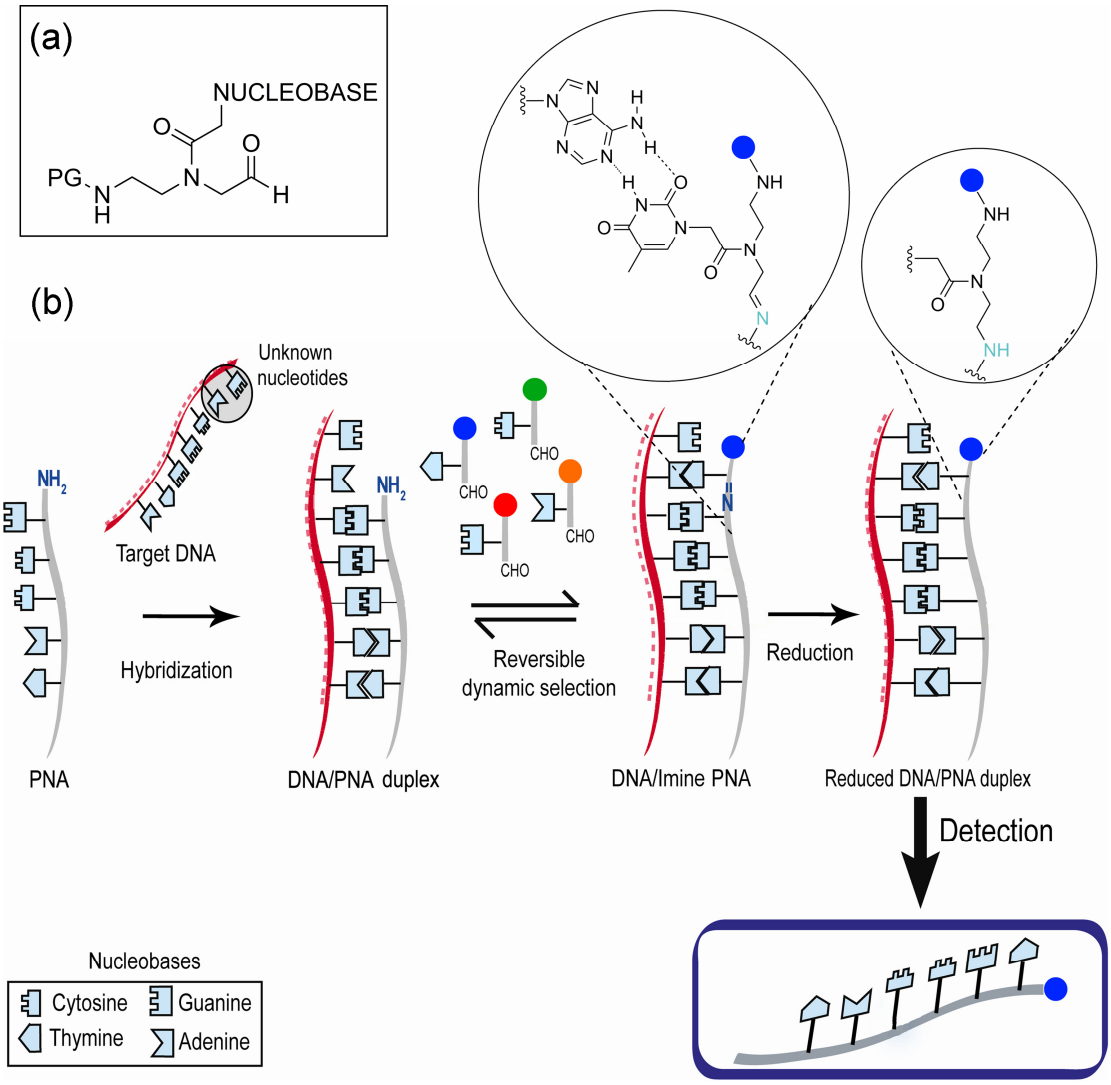


Figure 1.17 (a) General structure of an aldehyde-modified PNA monomer. The protecting group, 'PG', could contain a fluorophore to enable fluorescent read-out. (b) Sequence specific DNA-templated extension of a PNA probe by reductive amination.

The global aim of this thesis was to investigate the scope of templated dynamic chemistry on PNA as a tool for DNA sequence analysis.

CHAPTER 2

DNA-Templated Base-Filling Reactions on a Peptide Nucleic Acid Backbone

"DNA analysis by Dynamic Chemistry", F. R. Bowler, J. J. Diaz-Mochon, M. D. Swift and M. Bradley, *Angew. Chem., Int. Ed.*, 2010, **49**, 1809-1812.

2.1 Introduction

The question at the forefront of early investigations related to the degree of selectivity that could be obtained by DNA-templated reductive aminations at 'blank' positions on a PNA backbone. Model studies were designed to probe the scope of this dynamic approach, which necessitated the synthesis of aldehyde-modified nucleobases and PNA probes containing one or more free secondary amines ('blanks') at defined locations.

Mass spectrometry was selected as the detection method which would allow direct read-out of the reaction products without the need for additional probe or nucleobase labelling. Specifically, it was envisaged that MALDI-TOF MS would be used. PNA is well-suited to detection by MALDI-TOF MS, as the neutral backbone renders it less prone to adduct formation and its ions (which tend to be singly charged) are more stable than those of DNA.[92] Indeed, PNA has already been used in a number of hybridization-based genotyping assays that employ MALDI-TOF, as described previously (see Chapter 1.2.1).

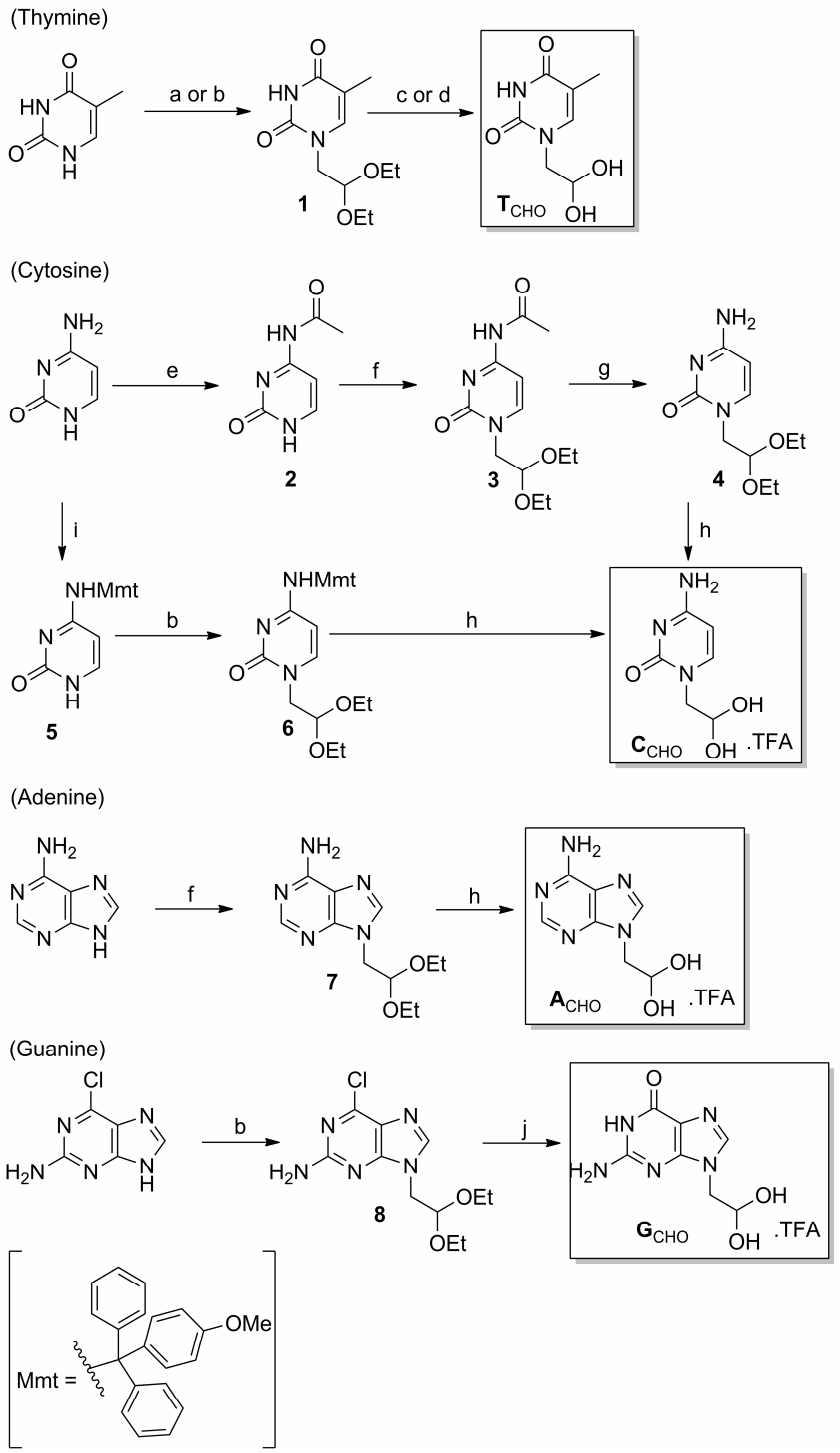
2.2 Synthesis of Aldehyde-Modified Nucleobases

The four target aldehydes ($\mathbf{T}_{CHO}$, $\mathbf{C}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$; Scheme 2.1) were prepared by microwave-assisted acid hydrolysis of the corresponding diethyl acetals. Thymine was thus alkylated at the *N*1 position with bromoacetaldehyde diethyl acetal to afford **1**, which was hydrolyzed by refluxing in aqueous HCl to yield aldehyde $\mathbf{T}_{CHO}$ as the hydrate according to literature methods.[104-106] This route was later accelerated through the use of microwave heating. The protocol was extended to the direct

alkylation of adenine at *N*9 to yield acetal **7**, which was hydrolyzed in aqueous TFA to afford aldehyde $\mathbf{A}_{CHO}$ as the hydrate trifluoroacetate.

To derivatize cytosine, it was necessary to protect the exocyclic amine before alkylation at *N*1. This was achieved through the use of the monomethoxytrityl (Mmt) protecting group[85] (see intermediates **5** and **6** in Scheme 2.1), allowing a subsequent acetal and amine deprotection to be performed in a single step by treatment with aqueous TFA. However, the target aldehyde $\mathbf{C}_{CHO}$ required purification, and column chromatography (eluting with methanol:dichloromethane) afforded a portion of the target as the methanolic hemiacetal. An alternative route employed the acyl (Ac) protecting group which could be removed with methanolic ammonia after alkylation (see intermediates **2**, **3** and **4** in Scheme 2.1). Acetal deprotection of **4** then afforded $\mathbf{C}_{CHO}$ as the hydrate trifluoroacetate without the need for purification.

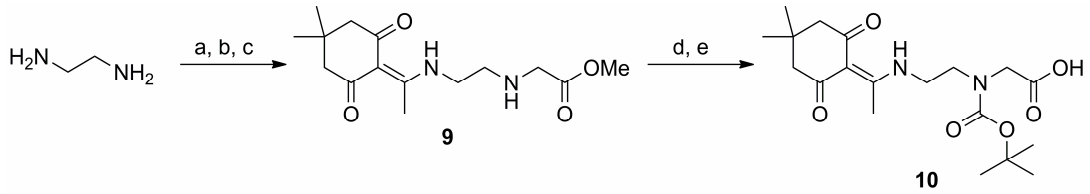
In the case of guanine, 2-amino-6-chloropurine was employed as the starting material to circumvent the poor solubility of the natural base and difficulty in achieving selective alkylation at the *N*9 position.[83] Thus, acetal **8** was obtained by direct alkylation of 2-amino-6-chloropurine with bromoacetaldehyde diethyl acetal. Nucleophilic aromatic substitution of the chloride and acid hydrolysis of **8** was performed in a single step to yield $\mathbf{G}_{CHO}$ as the hydrate trifluoroacetate. An interesting point to note is that $\mathbf{G}_{CHO}$ was occasionally isolated as a blue solid, and the ^{1}H NMR complicated by multiple peaks when recorded immediately after dissolution in H_2O/D_2O, possibly due to self-condensation to form (poly)imines or hydrogen bonded (e.g. G-quartet) species that were long-lived on the NMR timescale.[107] However, LCMS analysis always showed the target mass, and upon standing overnight at room temperature in H_2O or D_2O the ^{1}H NMR showed a single species, consistent with the target structure.



Scheme 2.1 Synthesis of modified nucleobases: (a) bromoacetaldehyde diethyl acetal, K_2CO_3, DMF, 130 °C, 20 h; (b) bromoacetaldehyde diethyl acetal, Cs_2CO_3, DMF, microwave, 100 °C, 30 min; (c) 1 M HCl aq, reflux, 70 min; (d) 1 M HCl aq, microwave, 30 min; (e) 5:1 v/v acetic anhydride:glacial acetic acid, reflux, 17 h; (f) bromoacetaldehyde diethyl acetal, K_2CO_3, DMF, microwave, 130 °C, 30 min; (g) 2 M NH_3 in MeOH, 92 h; (h) 1:1 v/v TFA:H_2O, microwave, 100 °C, 30 min; (i) 4-methoxytrityl chloride, 40 °C, 30 min, then RT, 16 h; (j) 1:2 v/v TFA:H_2O, microwave, 100 °C, 30 min.

2.3 Design and Synthesis of PNA Oligomers and Determination of Duplex Melting Temperatures

PNA oligomers were prepared by solid-phase synthesis (SPS) using Fmoc/Bhoc protected monomers. The free secondary amine ('blank') functionality was built into the probes through the use of monomer **10** (Scheme 2.2) which enabled the blank to be liberated from a *tert*-butyloxycarbonyl (Boc) protecting group upon acidic cleavage from the solid support.



Scheme 2.2 Synthesis of the Dde and Boc protected monomer used to assemble blank positions into PNA probes: (a) chloroacetic acid, 11 °C → RT, 16 h; (b) $SOCl_2$, MeOH, 0 °C → reflux, 15 h; (c) DiPEA, Dde-OH, 1:1 v/v DCM:EtOH, 16h; (d) Boc_2O, Et_3N, 5 h; (e) Cs_2CO_3, 1:1 v/v MeOH:H_2O, RT, 1.5 h.

The 15-mer probe **P1** (Table 2.1) was designed to allow optimization of reaction conditions for the dynamic incorporation and address the question of reaction selectivity (see Chapter 2.4) using DNA 21-mer templates **I-IV** (Table 2.2) The *N*-terminus of the PNA was protected with an acyl group, but triphenylphosphonium charge tags (Figure 2.1) were used as *N*-terminal caps in later probes to enhance the MALDI-TOF detection limit (see Chapter 3.2).[97]

Table 2.1 PNA 'blank' probes for templated base-filling reactions.

PNA Oligomer	Sequence (N – C)[a]	Mass (Da)[c]
P1	Ac-TAC TAC ATC _CT TCC	3824.6
P2[b]	Phosphonium-PEG1-GTG GAG _TC AAC GA	4356.8
P3[b]	Phosphonium-PEG2-GTG GAG _ _C AAC GA	4118.7
P4[b]	Phosphonium-PEG1-GTG GAG _ _ _ AAC GA	4039.7
P5	Phosphonium-TCG TTG A _C TCC AC	3916.6

[a]'_' Represents a blank site (see Chapter 1, Figure 1.16b). PNA oligomers were synthesized by solid phase synthesis and had a *C*-terminal primary amide. [b]See Figure 2.1 for structures of Phosphonium and Phosphonium-PEG groups. [c]Calculated for the most common isotope.

PNA probes **P2-4** (Table 2.1) were prepared to study the templated incorporation of multiple contiguous nucleobases (see Chapter 2.5) by DNA **V** (Table 2.2). These probes were connected to a 'charge-tag' through polyethylene glycol (PEG) spacers (Figure 2.1). Such spacers improve the aqueous solubility of PNA, and would also permit multiplexed analysis of templated reactions by differentiating the probe masses to prevent overlap in the mass spectrum. **P5** (Table 2.1) was used to determine the effect that an abasic site in the templating position of the complementary DNA strand **VI** (Table 2.2) would have on the outcome of the reductive amination reaction (see Chapter 2.6).

To investigate whether blank positions in the PNA backbone would prevent hybridization with their complementary DNA templates, the duplex melting temperatures (*T*m) were measured for probes **P1-5** (Figure 2.2 and Table 2.3) using the hyperchromicty method (i.e. by measuring the increase in absorbance at 260 nm when the duplex unzips). Sigmoidal melting curves indicating duplex formation were observed for all probes with one exception; a melting temperature could not be obtained for **P4**, suggesting that the presence of three contiguous blanks prevented formation of a stable duplex with DNA **V**. The melting temperature of **P2** was substantially higher than that for **P1**, possibly because the triphenylphosphonium charge tag provides some stabilization through electrostatic interaction with the negatively charged DNA backbone. Surprisingly, the melting temperature for **P3** was approximately equal (within experimental uncertainty) to that for **P2**, despite the presence of an extra blank position. However, it should be noted that these probes are not entirely comparable as the charge tag of **P2** is connected *via* a longer PEG spacer than that used for **P3**. The broader melting transitions observed for probes **P2, P3** and **P5** as compared to **P1** indicate that their hybridization is less cooperative.

Table 2.2 DNA templates.

DNA Oligomer	Sequence (5' – 3')[a]
I	TTT TTT GGA AG**G** GAT GTA GTA
II	TTT TTT GGA AG**A** GAT GTA GTA
III	TTT TTT GGA AG**T** GAT GTA GTA
IV	TTT TTT GGA AG**C** GAT GTA GTA
V	TCG TT**G AC**C TCC AC
VI	GTG GAG **Z**TC AAC GA

[a]Nucleobases which lie opposite a blank position on the complementary PNA probe(s) are shown in bold . Z = abasic site.

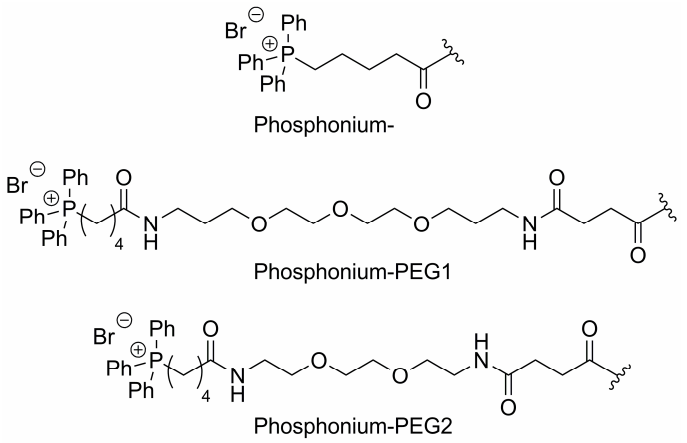


Figure 2.1 Structure of the charge tags (with and without PEG linkers) appended to probes **P2-5**.

Table 2.3 Melting temperatures of PNA probes (see Table 2.1 and 2.2 for oligomer sequences).

PNA Oligomer	DNA Oligomer	Melting temperature, T_m (± 1)/ ^oC[a]
P1	**I**	53
P2	**V**	66
P3	**V**	68
P4	**V**	NA
P5	**VI**	59

[a] Determined using CaryWin UV software, from the maximum of the first derivative of a plot of T *vs* A_{260nm}. NA, no sigmoidal melting curve was discernible. Uncertainty based upon the limited precision associated with A_{260nm} measurements at 1 ^oC intervals.

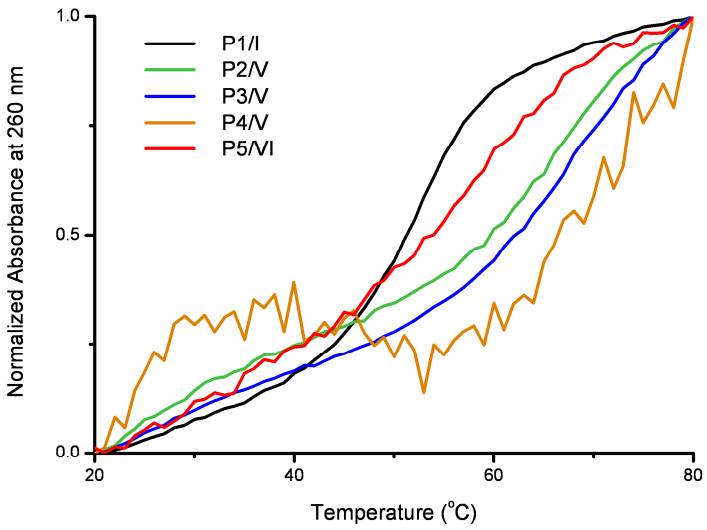
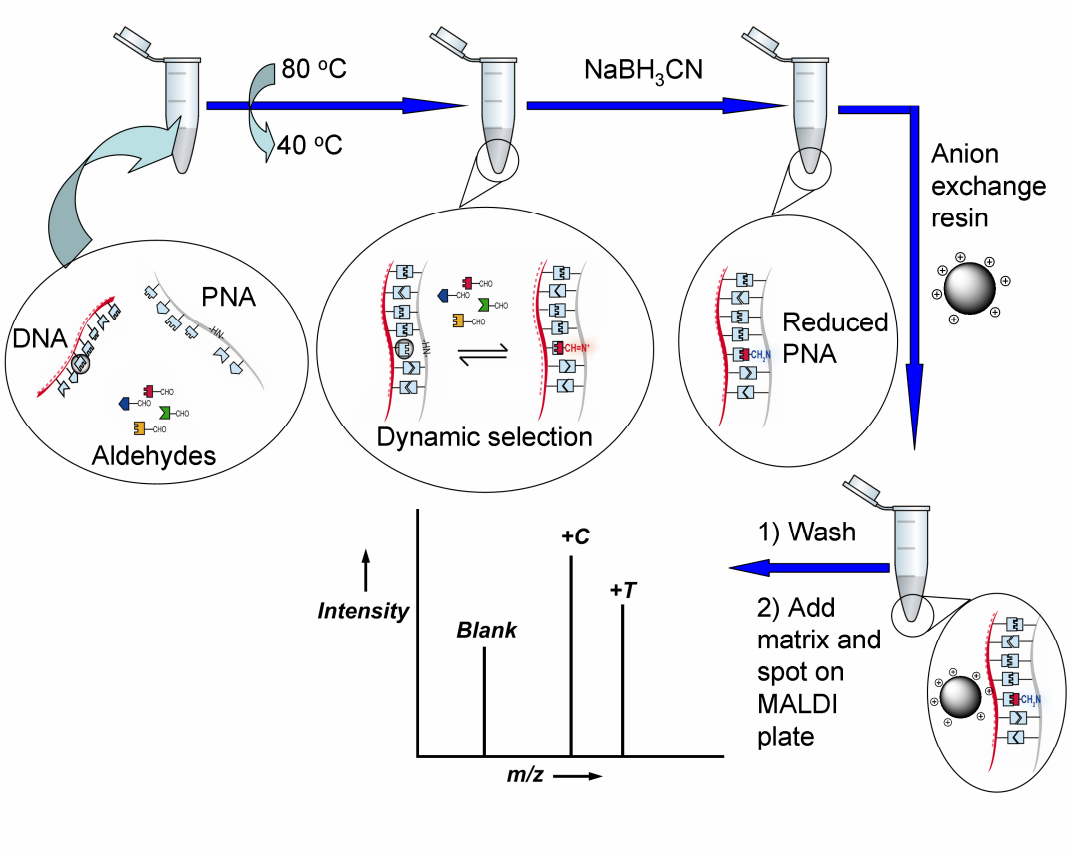


Figure 2.2 Melting curves for probes **P1** and **P2-4**. The absorbance values have been normalized to allow direct comparison.

2.4 DNA-Templated Single Base Incorporation

Early studies involved the development of a robust protocol for the MALDI-TOF analysis of DNA-hybridized PNA following templated reductive amination (Scheme 2.3). Thus, equimolar aqueous solutions of the four nucleobase aldehydes were prepared and added in excess to a 1:1 mixture of probe **P1** and one of the complementary DNA templates **I-IV**. To allow hybridization and the dynamic formation of iminium intermediates, this mixture was heated to 80 °C for 5 minutes and then cooled to 40 °C. Reduction was performed using sodium cyanoborohydride for 1 hour before Q Sepharose™ (GE Healthcare) was added. This anion exchange resin carries quaternary ammonium functionality and binds the negatively charged sugar phosphate backbone of the DNA template. In doing so, any hybridized PNA probe is also bound, permitting a washing step to remove any inorganic salts (which constitute the pH buffer and reducing reagent) that may give rise to adducts in the mass spectrum, together with any unbound PNA.[97] Finally, the resin beads were mixed with a sinapic acid matrix solution and spotted directly onto a stainless steel MALDI plate for MS analysis.



Scheme 2.3 Steps involved in DNA analysis by dynamic chemistry and MALDI-TOF MS.

The resulting mass spectra (Figure 2.3) demonstrated selective incorporation of the nucleobase complementary to the templating base on the DNA strand. The PNA signals suggested complete dehybridization from the DNA template, which may occur either upon addition of the acidic matrix or during desorption/ionization in the mass spectrometer. In each reaction, some unreacted **P1** was observed together with a peak of + 14 Da. This is hypothesized to be the result of a borane adduct due to the available free amine at the blank position of the probe (i.e. **P1** + BH_3).

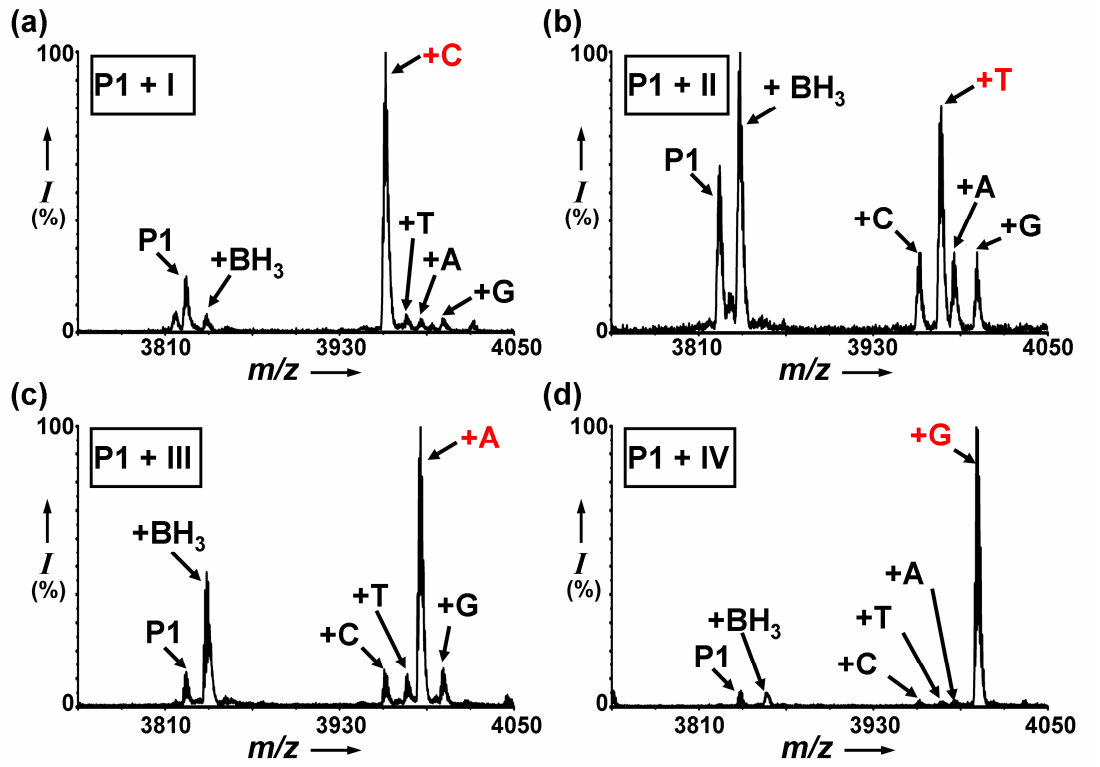
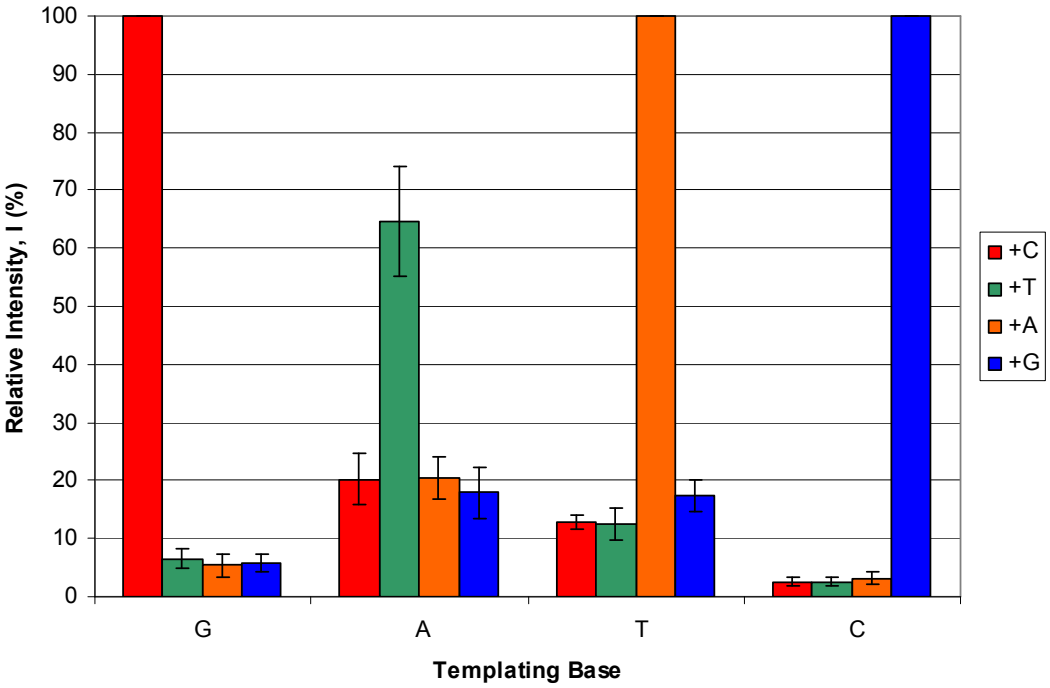


Figure 2.3 Representative mass spectra recorded after DNA-templated reductive aminations using an equimolar ratio of the four aldehydes. Peaks shown in red are due to products with the 'correct' base incorporated according to Watson-Crick base-pairing. (a) DNA template **I** directs incorporation of C; (b) **II** directs incorporation of T; (c) **III** directs incorporation of A; (d) **IV** directs incorporation of G. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.14 mM in each of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$; 5 μM in DNA template and **P1**; 100 mM in $NaBH_3CN$. *I* = relative intensity (as a percentage of the most intense peak), *m/z* = mass-to charge ratio.

Each reaction was performed in duplicate, and five mass spectra were obtained for each. Relative peak intensities (of the most common isotope) were recorded and averaged across the ten spectra. Based upon these data (Graph 2.1), it can be concluded that (under these conditions) guanine and cytosine are incorporated in greater yield and more selectively than either adenine or thymine (attributed to the greater number of templating hydrogen-bonds). Furthermore, purine bases were incorporated with greater yield and selectivity than pyrimidines (A>T, G>C), which may be due to greater π-stacking interactions associated with the bicyclic pyrimidine rings (Table 2.4).



Graph 2.1 Mean peak intensities (of the most common isotope, relative to the most intense peak) resulting from the templated reaction of an equimolar mixture of the four nucleobase aldehydes with **P1**. Error bars indicate the standard deviation across ten mass spectra (five from each duplicate analysis).

Table 2.4 MALDI signal ratios resulting from an equimolar mixture of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$.

DNA Oligomer	Templating Base	MALDI Signal Ratios[a]	
		C:T:A:G[b]	(C+T+A+G):SM[c]
I	G	**19**:1:1:1	4:1
II	A	1:**4**:1:1	1:1
III	T	1:1:**8**:1	2:1
IV	C	1:1:1:**39**	9:1

[a]Based upon the mean relative intensities of the most common isotope and reported to the nearest integer. [b]The value for the nucleobase complementary to the position under interrogation on the DNA template is in bold. [c]Ratio of starting material (sum of the peaks associated with **P1** and (**P1**+ BH_3)) to base incorporated product.

The reversibility of the nucleobase incorporation (prior to reduction) was investigated by analyzing the reaction of PNA/DNA (**1**/**IV**). In the absence of $\mathbf{G}_{CHO}$ small levels of mis-primed incorporation were detected after reduction. However,

when $\mathbf{G}_{CHO}$ was added to the reaction mixture immediately before reduction, the removal of virtually all mis-primed binding resulted, showing the reversibility of the selection process (see Figure 2.4).

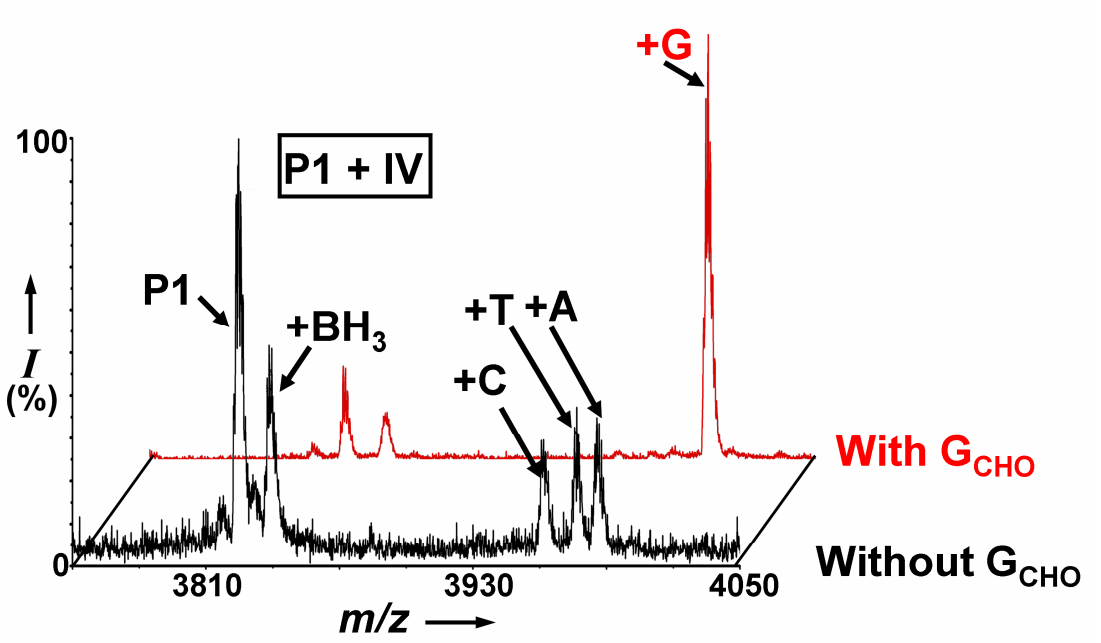


Figure 2.4 Reaction in the absence of $\mathbf{G}_{CHO}$ results in the incorporation of C, T and A (resulting mass spectrum shown in black), but the addition of $\mathbf{G}_{CHO}$ just before reduction (and after reaction with $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$ and $\mathbf{A}_{CHO}$ for 1h) prevents formation of essentially all of these 'incorrect' products (resulting mass spectrum shown in red).

It was found that altering the starting concentrations of the nucleobase aldehydes gave different product ratios (Figure 2.5). Increasing the concentrations of those bases that showed poor selectivity and yield under equimolar conditions (such that $[\mathbf{T}_{CHO}] > [\mathbf{A}_{CHO}] > [\mathbf{C}_{CHO}] > [\mathbf{G}_{CHO}]$) resulted in improved yields and peak ratios for the 'correct' templated product (Graph 2.2). In particular, an improvement in selectivity was observed for T and A at the expense of C and G incorporation (Table 2.5).

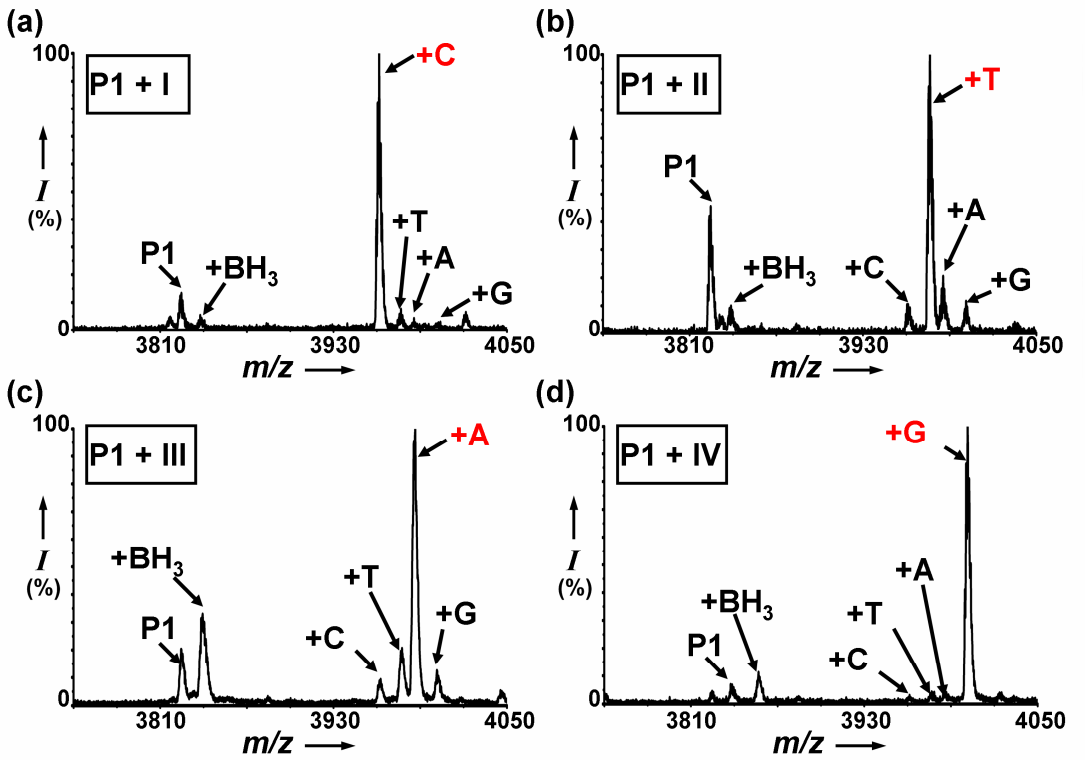
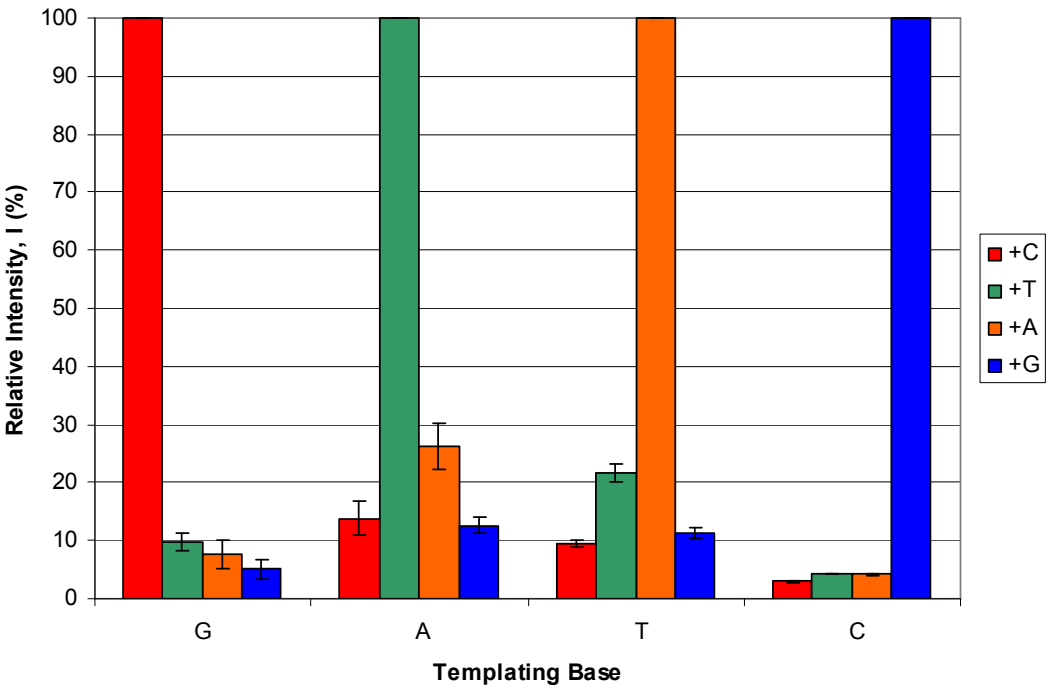


Figure 2.5 Representative mass spectra recorded after DNA-templated reductive aminations using a non-equimolar ratio of the four aldehydes. (a) DNA template **I** directs incorporation of C; (b) **II** directs incorporation of T; (c) **III** directs incorporation of A; (d) **IV** directs incorporation of G. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.18 mM in $\mathbf{C}_{CHO}$, 0.45 mM in $\mathbf{T}_{CHO}$, 0.26 mM in $\mathbf{A}_{CHO}$ and 0.14 mM in $\mathbf{G}_{CHO}$ (ratio ~ 1:3:2:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 5 μM in DNA template and **P1**; 100 mM in $NaBH_3CN$.

Table 2.5 MALDI signal ratios resulting from a non-equimolar mixture of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$.

DNA Oligomer	Templating Base	MALDI Signal Ratios[a]	
		C:T:A:G[b]	(C+T+A+G):**P1**[c]
I	G	**20**:2:1:1	5:1
II	A	1:**8**:2:1	2:1
III	T	1:2:**10**:1	2:1
IV	C	1:1:1:**34**	9:1

[a]Based upon relative intensities of most common isotope and reported to the nearest integer. [b]The value for the nucleobase complementary to the position under interrogation on the DNA template is in bold. [c]Ratio of starting material (sum of the peaks associated with **P1** and (**P1**+ BH_3)) to base incorporated product.



Graph 2.2 Mean relative peak intensities (of the most common isotope, relative to the most intense peak) resulting from the templated reaction of a non-equimolar mixture the four nucleobase aldehydes with **P1**. Error bars indicate the standard deviation across three mass spectra recorded for a single analysis.

To study the effect of pH on reaction outcome, DNA **II** was used to template reductive aminations on **P1** using buffers to vary the pH over the range 5.0-8.5 (Figure 2.6). As anticipated for iminium ion formation, conversions were better at mildly acidic pH; pH 6.0 was optimal, giving the best yields as judged by the relative intensities of product and starting material peaks. A mildly acidic pH strikes the balance between being high enough to provide sufficient free amine to attack the carbonyl group of the aldehyde, but low enough for protonation of the carbonyl oxygen prior to nucleophilic attack, and also protonation of the resulting tetrahedral intermediate for elimination of water.

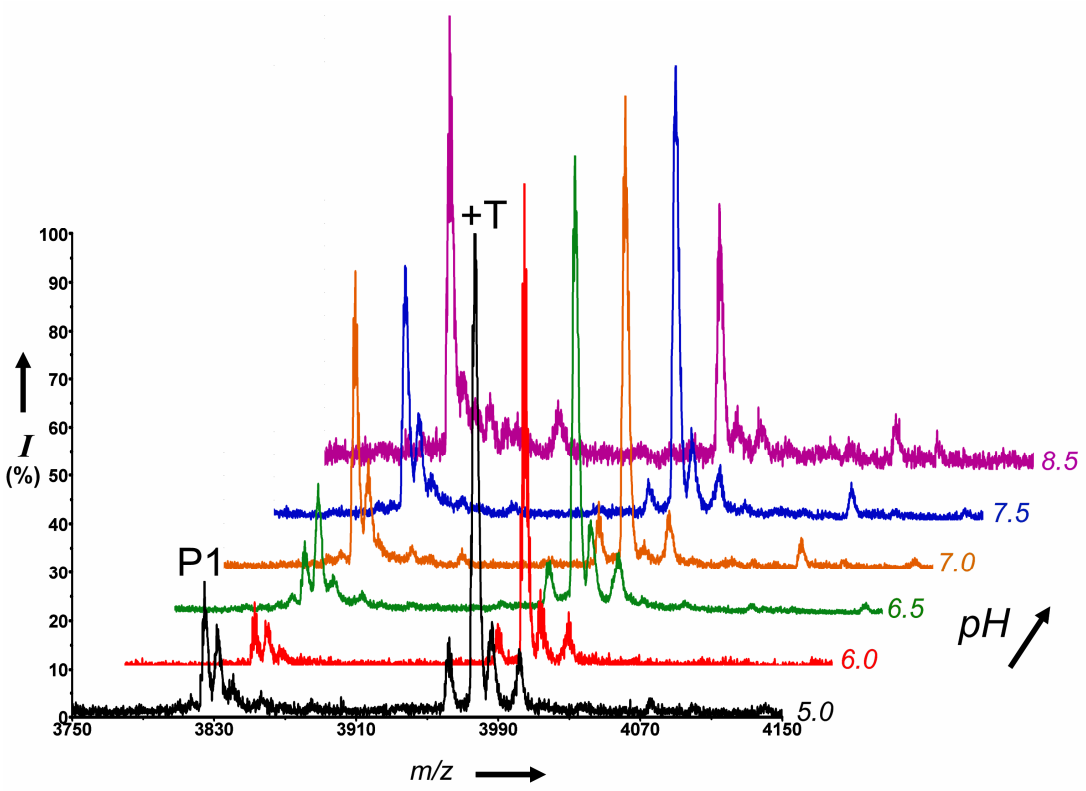


Figure 2.6 Effect of pH on templated incorporation of **T** on **P1**. TAPS was used to buffer pH 8.5, sodium acetate to buffer pH 5.0, and sodium phosphate to buffer the intermediate pH values. Final concentrations in a 20 μL reaction volume: 4 mM buffer; 0.18 mM in $\mathbf{C}_{CHO}$, 0.45 mM in $\mathbf{T}_{CHO}$, 0.26 mM in $\mathbf{A}_{CHO}$ and 0.14 mM in $\mathbf{G}_{CHO}$; 5 μM in DNA template and **P1**; 100 mM in $NaBH_3CN$.

Work next turned to the analysis of mixtures of DNA templates. This was a prelude to SNP genotyping, as any method for allele discrimination must permit the genotyping of heterozygous individuals who possess two different alleles of a particular gene. Heterozygotes present a greater challenge than homozygotes, as the ratio of signals due to each allele (or the 'allelic ratio') should be close to unity. To explore the analysis of 'heterozygotes' by dynamic chemistry, DNA oligomers **I-IV** were mixed to generate each of the six possible combinations of templating base which may be present for a single biallelic variant. Thus a ratio of 1:3:2:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$ gave approximately equal ratios of the correct incorporation products for the templating combinations A/T and G/C ('allelic ratios' of 0.83 and 0.70 respectively, Figure 2.7b and e). Poorer 'allelic ratios' were observed for templates G/A and G/T (0.25 and 0.53 respectively, Figure 2.7a and c), whilst

selectivity was worse still for T/C and A/C ('allelic ratios' of 0.13 and 0.15 respectively, Figure 2.7d and f; results summarized in Table 2.6). However, by further increasing the concentrations of those nucleobase aldehydes ($\mathbf{T}_{CHO}$ and $\mathbf{A}_{CHO}$) incorporated in lower yield and poorer selectivity, it was possible to bring the ratio of peaks representing 'correct' products closer to unity. For example, an improvement in the 'allelic ratio' from 0.13 to 0.86 was observed for T/C templated incorporation in this way (Figure 2.8a and last entry in Table 2.6). The mass spectrum for a control reaction (with no DNA template, Figure 2.8b) with this unequal ratio of aldehydes gave an unselective and low yield of incorporation products, further demonstrating the importance of the template in the base-filling reaction.

Table 2.6 MALDI signal ratios observed for mixtures of two DNA templates.

DNA Oligomers	Templating Bases	MALDI Signal Ratios[a]	
		C:T:A:G[b]	'Allelic' Ratio[c]
I+II	G+A	**29**:**7**:1:1	0.25
I+III	G+T	**11**:2:**6**:1	0.53
I+IV	G+C	**8**:1:1:**11**	0.70
II+III	A+T	1:**7**:**8**:1	0.83
II+IV	A+C	1:**3**:1:**20**	0.15
III+IV	T+C	1:2:**3**:**20**	0.13
III+IV[d]	T+C	1:5:**27**:**31**	0.86

[a]Based upon mean relative intensities of most common isotope. In each case values have been averaged between two spectra recorded for a single analysis using a ratio of 1:3:2:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$ unless noted otherwise. [b]The values for the nucleobases complementary to the position under interrogation on the DNA templates are in bold and are reported to the nearest integer. [c]Calculated by dividing the intensity of the larger 'correct' peak by that for the smaller 'correct' product. [d]Using a ratio of 1:14:8:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$.

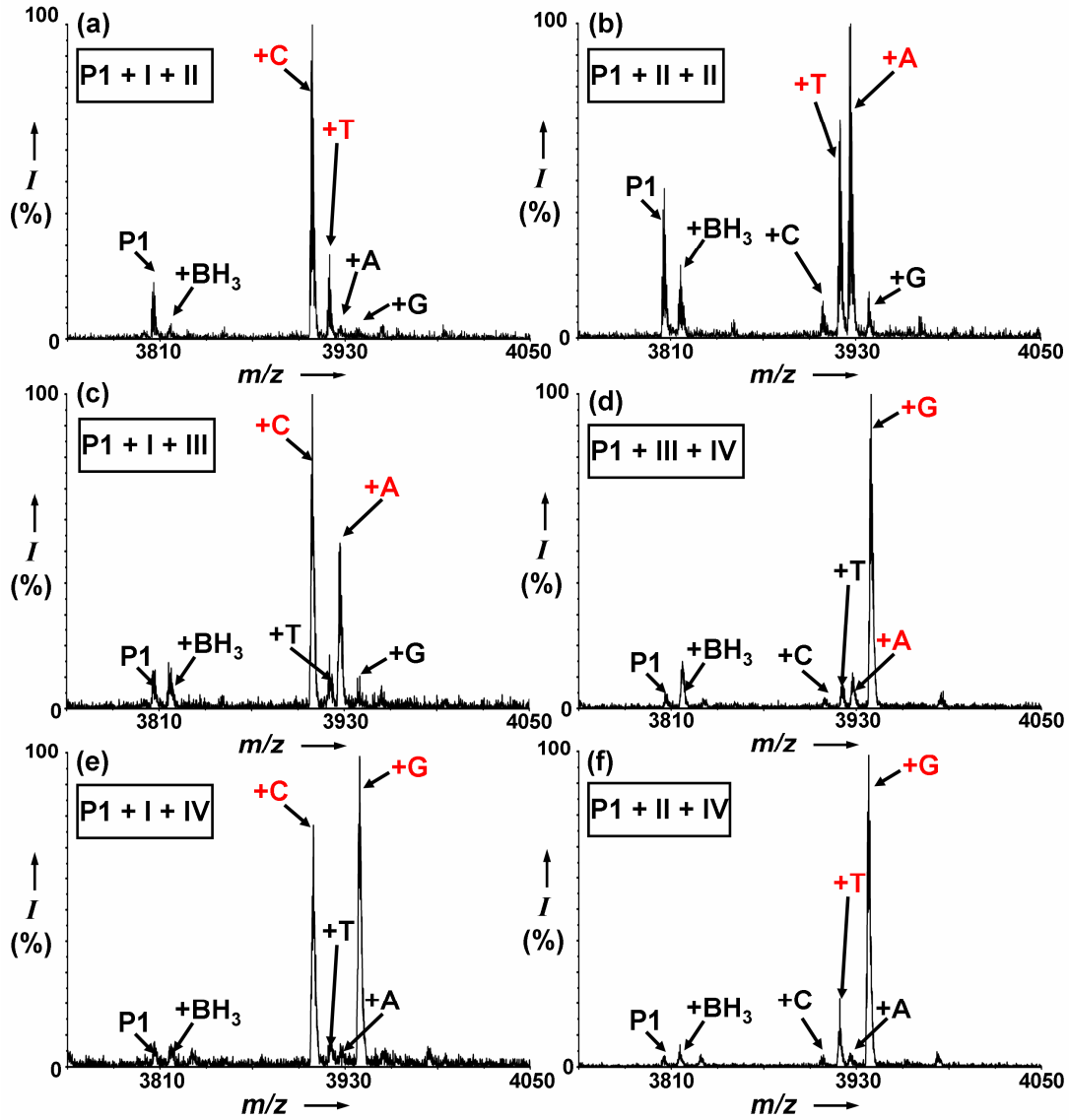


Figure 2.7 Dynamic incorporation templated by a 1:1 mixture of two DNA oligomers. Templating bases were: (a) G and A; (b) T and A; (c) G and T; (d) T and C; (e) G and C; (f) A and C. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.18 mM in $\mathbf{C}_{CHO}$, 0.45 mM in $\mathbf{T}_{CHO}$, 0.26 mM in $\mathbf{A}_{CHO}$ and 0.14 mM in $\mathbf{G}_{CHO}$ (ratio 1:3:2:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 2.5 μM in each of two DNA templates (5.0 μM total [DNA]); 5.0 μM in **P1**; 100 mM in $NaBH_3CN$.

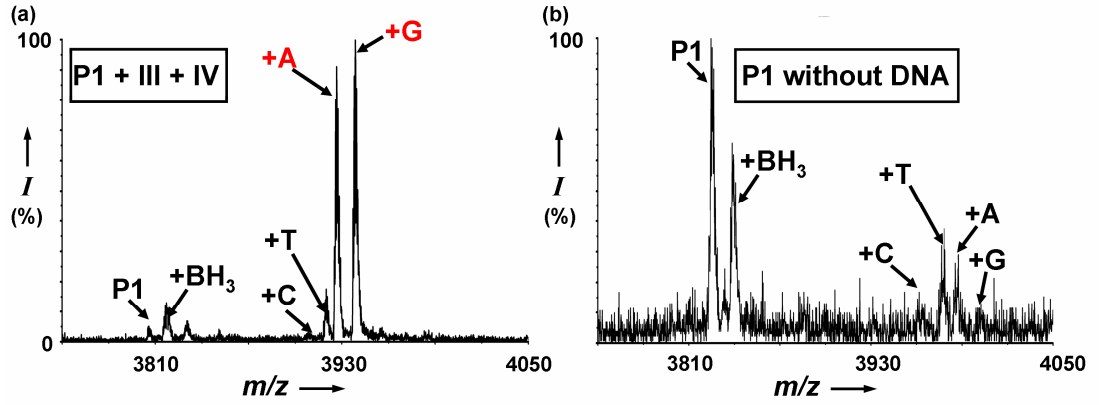
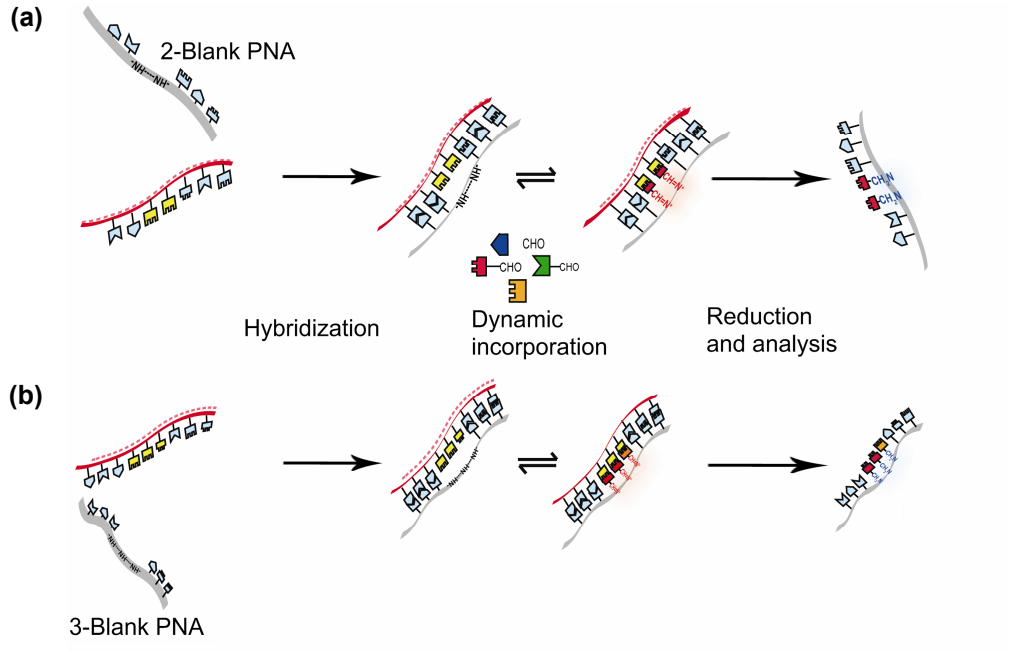


Figure 2.8 (a) Selective incorporation of A and G in approximately equal ratio. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.07 mM in $\mathbf{C}_{CHO}$, 0.73 mM in $\mathbf{T}_{CHO}$, 0.43 mM in $\mathbf{A}_{CHO}$ and 0.05 mM in $\mathbf{G}_{CHO}$ (ratio 1:14:8:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 2.5 μM in each of two DNA templates (5.0 μM total [DNA]); 5.0 μM in **P1**; 100 mM in $NaBH_3CN$. (b) Control reaction in the absence of DNA (using water to maintain a constant reaction volume and concentration of other reaction components and buffer). The lower signal-to-noise ratio arises because there is no DNA to aid binding of PNA to the anion exchange resin prior to the washing step.

2.5 DNA-Templated Incorporation of Multiple Bases

It was anticipated that probes with two or more contiguous blanks could permit the genotyping of mutations involving the insertion or deletion of multiple bases (indels, see Chapter 1.1.5). To investigate incorporation at multiple contiguous blank sites (Scheme 2.4), DNA **V** (Table 2.2) was used to template reactions on probes **P2**, **P3** and **P4** which possess one, two and three abasic positions respectively (Table 2.1). Full and selective reaction on these probes would yield '+G', '+GT' and '+GTC' incorporation products. As expected, the selective addition of $\mathbf{G}_{CHO}$ to **P2** proceeded smoothly (Figure 2.9a). However, although peaks attributed to fully base-filled products were observed for **P3** and **P4** (Figure 2.9b and 2.9c) more intense peaks were observed with masses corresponding to incomplete incorporation, and this was not improved by increasing the times for either the dynamic selection or reduction steps. Furthermore, the mass spectra recorded for **P3** and **P4** were complicated by the presence of multiple adducts attributed to BH_3. An explanation for this may be that the additional free amine functionality in these probes provides greater scope for complexation and adduct formation with borane.

Despite the fact that T_m analysis for **P4**/**V** did not indicate the formation of a stable duplex, the observation of base incorporation products for **P4** would suggest that some hybridization with **V** has occurred. Indeed, the dynamic incorporation would 'drive' formation of double stranded product.



Scheme 2.4 Illustration of the templated incorporation of (a) two and (b) three nucleobases at contiguous blank positions.

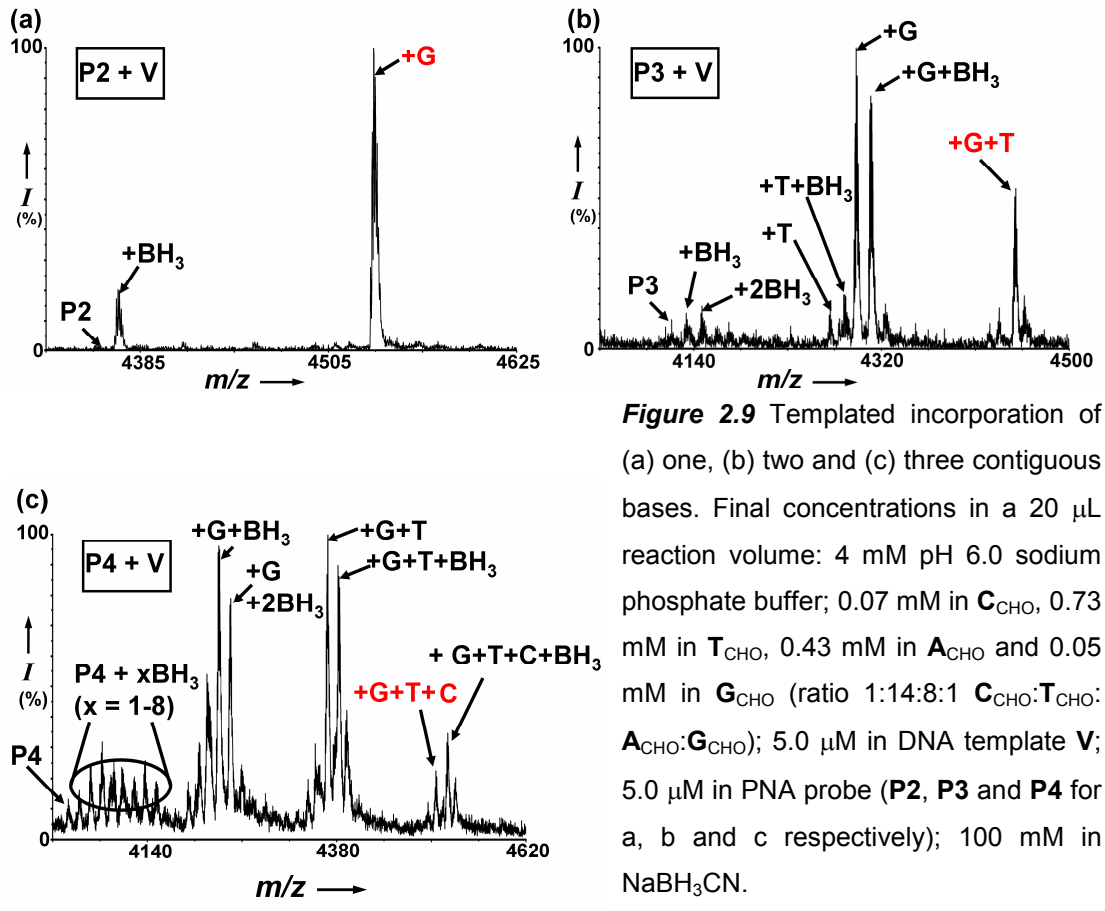
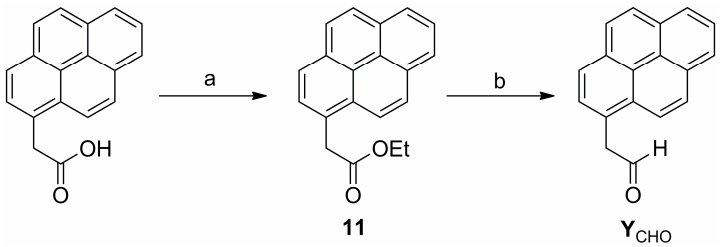


Figure 2.9 Templated incorporation of (a) one, (b) two and (c) three contiguous bases. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.07 mM in $\mathbf{C}_{CHO}$, 0.73 mM in $\mathbf{T}_{CHO}$, 0.43 mM in $\mathbf{A}_{CHO}$ and 0.05 mM in $\mathbf{G}_{CHO}$ (ratio 1:14:8:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 5.0 μM in DNA template **V**; 5.0 μM in PNA probe (**P2**, **P3** and **P4** for a, b and c respectively); 100 mM in $NaBH_3CN$.

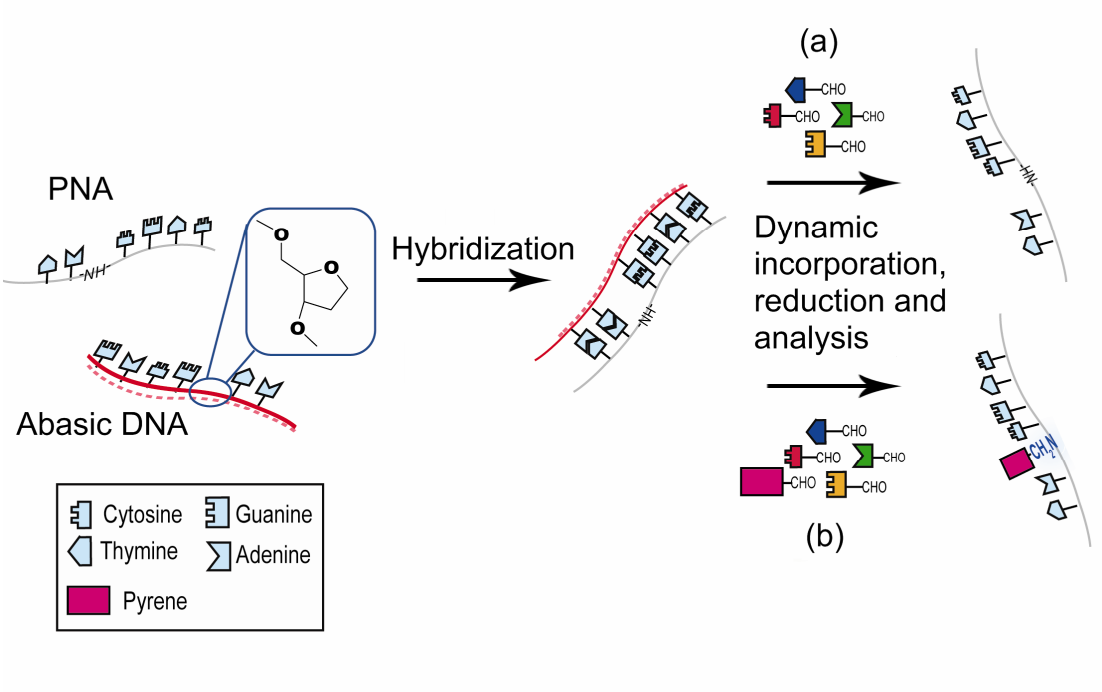
2.6 Abasic Sites and Templated Incorporation

Abasic sugars are found naturally in the genome as a result of spontaneous lesions, or chemical or physical damage.[108, 109] To investigate the effect of an abasic site in the templating position of the DNA strand, oligomer **VI** was obtained which possessed a single abasic ribose moiety. It has been reported that pyrene deoxynucleoside triphosphate (dPTP) can be enzymatically incorporated opposite a templating abasic site in DNA.[110] The flat, fused aromatic rings of pyrene mimic the shape of a purine/pyrimidine base pair, enabling it to the span space between the backbones in the double helix. With this in mind, 1-pyreneacetaldehyde ($\mathbf{Y}_{CHO}$) was synthesized according to a literature method (Scheme 2.5) by reduction of ethyl ester **11**, prepared by esterification of commercially available 1-pyreneacetic acid.

Interrogation of the DNA abasic site with probe **P5** and the four nucleobase aldehydes resulted in virtually no base-incorporation products (Scheme 2.6a and Figure 2.10a), again demonstrating the role of the complementary base of the DNA template in promoting selective incorporation. Analysis of **VI** with **P5** was repeated, but this time $\mathbf{Y}_{CHO}$ was included in the reaction mixture. The resulting mass spectrum showed that $\mathbf{Y}_{CHO}$ had been selectively incorporated as predicted (Scheme 2.6b and Figure 2.10b).



Scheme 2.5 Synthesis of 1-pyreneacetaldehyde: (a) ethanol, H_2SO_4, reflux, 4 h; (b) DIBAL-H, toluene, - 78 ^{o}C, 1.5 h.



Scheme 2.6 Dynamic analysis of abasic DNA using the four canonical nucleobase aldehydes in the absence (a) and presence (b) of 1-pyreneacetaldehyde, $\mathbf{Y}_{CHO}$.

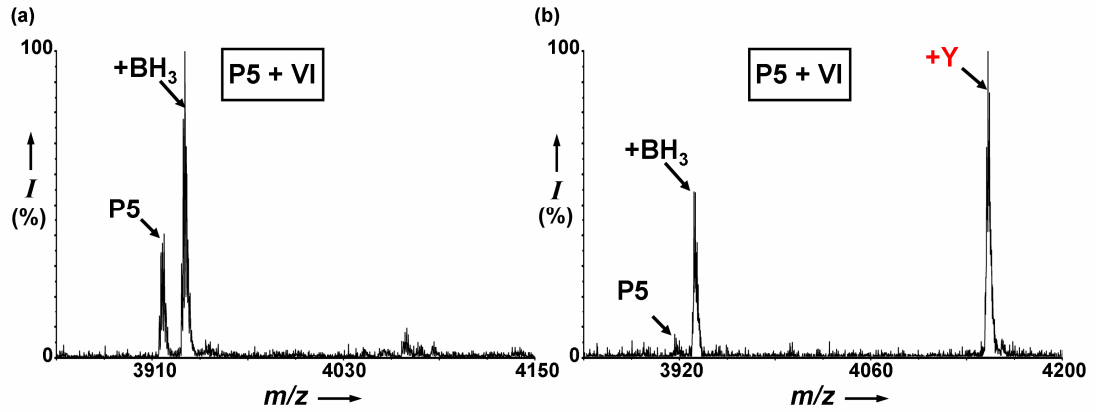


Figure 2.10 Mass spectra resulting from the reaction templated by an abasic site in the absence (a) and presence (b) of $\mathbf{Y}_{CHO}$. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.14 mM in each of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$; 0.14 mM in $\mathbf{Y}_{CHO}$ (for b only); 5 μM in DNA template **VI** and **P5**; 100 mM in $NaBH_3CN$.

2.7 RNA

PNA/RNA hybridization has been reported widely, and it was therefore hypothesized that RNA would also be able to template base-filling reactions on PNA.[111] In view of this, a 15mer RNA complementary to **P1** was obtained (**RI**; sequence 5'-GGA AGU

GAU GUA GUA-3'). The templating base on this RNA oligomer was uracil (U), which would be expected to template incorporation of $\mathbf{A}_{CHO}$. Indeed, MALDI-TOF analysis of a base-filling reaction in the presence of equimolar amounts of the four nucleobase aldehydes showed selective addition of A (Figure 2.11).

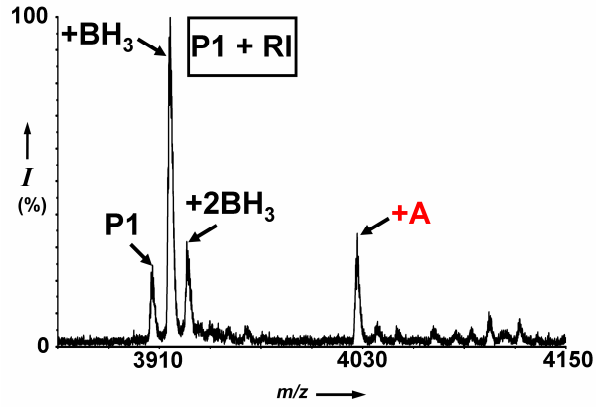


Figure 2.11 RNA-templated incorporation. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.14 mM in each of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$; 5 μM in **R1** template and **P1**; 100 mM in $NaBH_3CN$.

2.8 Discussion and Conclusions

The work presented in this Chapter has established that DNA can be used to template selective base-filling reductive amination reactions on a PNA strand possessing a blank position. Four iminium species (one for each base) are created, but the one with the correct shape and hydrogen-bonding motif (obeying Watson-Crick base-pairing) is the most stable and this is reflected in the product ratios measured using MALDI-TOF mass spectrometry. Differences in selectivity and yield between the four nucleobases have been attributed firstly to differing numbers of templating hydrogen bonds, and secondly to differences in π-stacking interactions between the purines and pyrimidines. As anticipated for iminium ion formation, reaction yields were found to be optimal at a slightly acidic pH (6.0).

In accordance with Le Chatelier's principle for a system in dynamic equilibrium, increasing the starting concentrations of the aldehyde-modified nucleobases that gave lower reaction selectivities and yields increases the representation of those nucleobases in the product distribution. This result has implications for the application of the system to the genetic analysis of individuals heterozygous for a particular SNP, as approximately equal peak intensities resulting from each allele (or an 'allelic ratio' close to unity) would permit easier genotyping.

The mass spectra obtained for multiple contiguous blanks showed incomplete base-filling, and were complicated by multiple adducts (attributed to BH_3). The use of such probes would therefore seem unlikely to afford a useful means of genotyping indels using the conditions described. It is interesting to note that the most intense peaks observed for incorporation on the probe (**P4**) possessing three contiguous blanks correspond to base-filling in the *N*→*C* direction. The reason for this apparent directional preference for base-filling is unclear, but may stem from a greater duplex melting temperature for the (6 base) *N*-terminal stem of **P4** relative to the (5 base) *C*-terminal stem.

A reaction in which the templating DNA base was replaced with an abasic site, and a control reaction in the absence of DNA, have further demonstrated the role of the DNA template in promoting selective nucleobase incorporation. In both cases the absence of template afforded a low and unselective yield of base-filled products. However, 1-pyreneacetaldehyde ($\mathbf{Y}_{CHO}$) was found to selectively incorporate opposite an abasic site on the DNA template, which provides scope for the application of this system to the analysis of abasic sites within genomic DNA.

Finally, an experiment using an RNA template demonstrated the selective incorporation of the 'correct' base. This result suggests that the approach to allele-discrimination reported herein could be extended to direct RNA sequence analysis.

During the course of this work, a similar study undertaken by Heemstra and Liu was published in which sequence-selective base-filling reactions on PNA oligomers were shown to be templated by complementary PNA strands.[112] This provides further confirmation of the results presented herein, in particular the relative reaction selectivities observed for the four nucleobases. Base-filling reactions were performed at the middle and end of a PNA strand, but middle-of-strand incorporation was more efficient, emphasizing the role of base-stacking interactions in the reaction. Both reductive amination and acylation reactions were reported to give sequence-selectivity, but better yields and selectivities were obtained by the former method, which supports the hypothesis that dynamic iminium formation provides a route to reverse the generation of mis-templated products. However, the use of PNA as a template limits the utility of this study, and no mention is made of a potential application to genetic analysis.

The utility of this method of DNA analysis would be greatly enhanced if the incorporation could be made to be catalytic, such that selective incorporation on a PNA 'blank' probe would be driven by a catalytic amount of DNA template. This would result in a substantial lowering of the detection limit by this method, and could perhaps be achieved if the dynamic incorporation and reduction were performed at, or slightly above, the duplex melting temperature (T_m) of the blank PNA/DNA. This may be difficult to control, however, given that the base-filled product is likely to be more stable than the original blank PNA/DNA duplex.[112] Furthermore, the copying fidelity observed in the model systems described above is insufficient (particularly in the case of thymine incorporation) for the transmission of sequence information through repeated rounds of dynamic incorporation and reduction.

Although iminium formation has been used as a dynamic reaction requiring minimal alteration of the canonical PNA structure, a wealth of alternative strategies compatible with aqueous conditions are available for reversible covalent bond formation.[99] Recently, Ghadiri and co-workers demonstrated the DNA-, RNA- and self-templated dynamic self-assembly of thioester-modified nucleobases on oligo-dipeptide backbones containing alternating cysteine residues (thioester PNA, or 'tPNA').[113] This approach employs thioester exchange as the reversible reaction, which has an advantage over iminium formation in that an additional reduction step is not needed to trap out the final products. However, tPNA hybridizes much less strongly than PNA with DNA and RNA, and thioester exchange is possible at all positions on the tPNA strand which would make sequence analysis more problematic. Nonetheless, this work further demonstrates the utility of dynamic chemistry on a pre-formed peptide backbone for the formation of informational polymers. Given the initial complexities observed in the ^{1}H NMR spectrum of $\mathbf{G}_{CHO}$ (see Chapter 2.2 above), it is interesting to note that these workers employed 7-deazaguanine in place of guanine with the explicit aim of avoiding G-quadruplex structures.

CHAPTER 3

Genotyping Cystic Fibrosis-Linked Mutations by Dynamic Chemistry

"Dynamic Chemistry for Enzyme-Free Allele Discrimination in Genotyping by MALDI-TOF Mass Spectrometry", F. R. Bowler, P. A. Reid, C. A. Boyd, J. J. Diaz-Mochon and M. Bradley, *Anal. Methods*, accepted for publication.

3.1 Introduction

Although dynamic chemistry had been validated as an approach to the analysis of single-base changes in synthetic DNA (Chapter 2), the question remained as to whether dynamic chemistry could be reliably applied as a means of allele discrimination for genotyping. To answer this, a study was planned in which individuals diagnosed with cystic fibrosis (CF), whose genotypes had been previously determined using established methods, would be genotyped for CF-linked mutations using dynamic chemistry.

CF is the most prevalent lethal recessive genetic disease among Caucasians, and arises from mutations in the gene coding for the cystic fibrosis transmembrane regulator.[114] This 1480-amino-acid protein is responsible for the transport of chloride ions across epithelial cell membranes. Over 800 disease-associated mutations (mostly SNPs) have been found in the *CFTR* gene, but around two-thirds of CF cases are linked to a single mutation, namely ΔF508, which constitutes a 3 base-pair deletion.[115, 116] This mutation was selected as a suitably challenging target to validate the new method, since the loss of 3 bases from the templating DNA necessitates a modification of the original probe design. As a more straightforward test, the G551D CF-linked SNP was also targeted for analysis. This SNP accounts for around 1-2 % of CF cases worldwide, and is one of just four mutations other than ΔF508 which account for > 1 % of CF alleles.[114] Moreover, G551D is particularly prevalent in the Scottish population, accounting for around 5 % of mutant alleles.[117] A second SNP, W1282X, which accounts for approximately 1-2 % of CF mutant alleles globally,

was targeted for model studies using synthetic DNA. W1282X is linked primarily with people of Ashkenazi Jewish descent, and as such has a relatively low incidence in the Scottish population.[117, 118] Clinical samples (sourced from the Western General Hospital in Edinburgh) were therefore not tested for this SNP.

Following initial model studies for ΔF508, W1282X and G551D on synthetic DNA, it was envisaged that a protocol for analysis of human genomic DNA could be developed using material sourced 'in-house' from the author of this thesis and another scientist working on the project (both of unknown genotype), and two samples from a commercial source (of known genotype). CF patients would then be genotyped for G551D and ΔF508 in a blind trial of the technology.

3.2 Design and Synthesis of PNA Oligomers

PNA probes were designed to 'clamp' the region around the target DNA (Table 3.1), selected to represent the sense sequence (the PNA probes were thus antisense, such that the PNA *N*-terminus lined up with the 3' end of the sense DNA). For the chosen SNPs (G551D and W1282X), this was straightforward as the sequence either side of the point mutation remains constant (see probes **P5-6**, Table 3.2). However, in the case of the indel (ΔF508), the removal of 3 bases causes a frameshift, which prohibits analysis with a single probe. Given the observation that multiple contiguous bases on a PNA blank leads to complicated mass spectra (see Chapter 2.5), the most simple way to address this was to employ two PNA probes for the analysis of ΔF508 which differed in their *N*-terminal sequence (see probes **P7-8**, Table 3.2); one to clamp the wild-type sequence, and one the mutant (Figure 3.1). To investigate the potential for multiplex genotyping, PNA probes were designed to prevent overlap in the mass spectrum.

Table 3.1 CF-linked mutations and the corresponding PNA probes used to test for them.

Mutation[a]	Reference dbSNP[b]	Allele Change[c]	PNA Probes
G551D	rs75527207	G → A	**P5**
W1282X	rs77010898	G → A	**P6**
ΔF508	rs113993960	del CTT	**P7, P8**

[a]Numbers refer to the amino acid position in the CFTR protein, and letters to the amino acid change (G is glycine, D aspartic acid, W tryptophan, X stop codon, ΔF phenylalanine deletion). For example, the W1282X mutation results in premature termination of the protein at a position that would normally contain tryptophan. [b]dbSNP is an open access archive of genetic variation hosted by the National Center for Biotechnology Information (NCBI, a US government-funded resource). It can be accessed at http://www.ncbi.nlm.nih.gov/projects/SNP/. [c]Sequence change of the mutation. 'G → A' means a replacement of guanine by adenine in the mutant allele, 'del CTT' refers to a deletion of these three bases from the code.

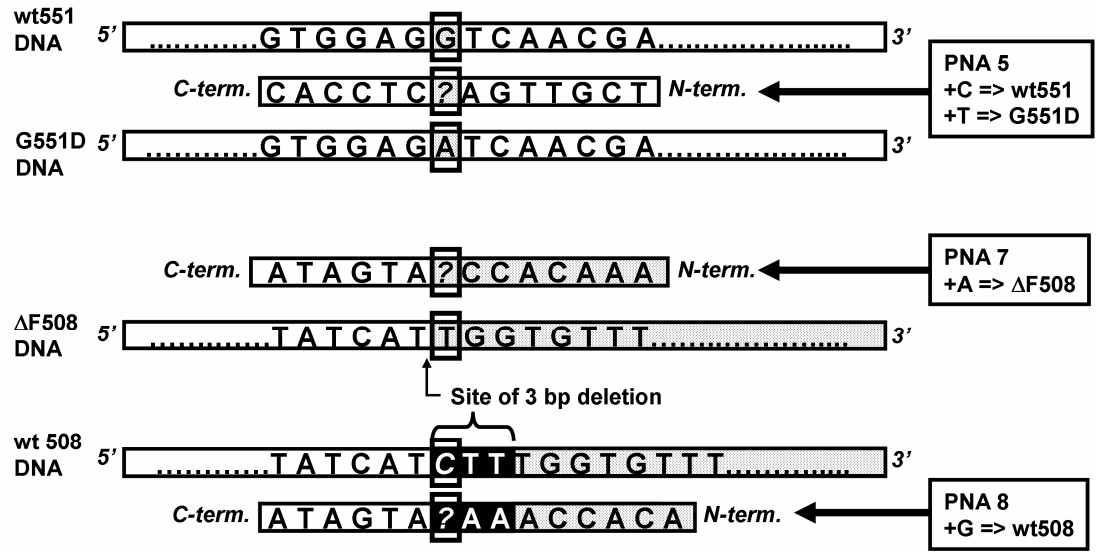


Figure 3.1 Illustration of the different approaches to SNP and indel allele discrimination. SNPs (G551D is shown but the principle is the same for W1282X) can be distinguished using a single PNA probe; incorporation of C and/or T will indicate the presence of the wild-type and/or mutant alleles respectively. The frameshift caused by a 3 base-pair deletion means that two probes are required to differentiate ΔF508 and wt508 alleles (**P7** and **P8** respectively). Incorporation of A into **P7** reports ΔF508, whilst incorporation of G into **P8** reports wt508. '*?*' denotes the blank position on a PNA probe.

Table 3.2 PNA 'blank' probes for CF genotyping.

PNA Oligomer	Length (number of nucleotides)	Sequence (N – C)[a]	Mass (Da)[b]
P5	14	TCG TTG A _C TCC AC	3916.6
P6	14	CTT TCC T _C ACT GT	3882.6
P7	14	AAA CAC C_A TGA TA	3966.6
P8	15	ACA CCA AA_ ATG ATA	4241.8

[a]'_' Represents a blank site. PNA oligomers synthesized by SPS with a *C*-terminal primary amide and *N*-terminal triphenylphosphonium charge tag (see Chapter 2, Figure 2.1). [b]Calculated mass of the most common isotope.

3.3 Model Studies using Synthetic DNA

Synthetic single-stranded DNA representative of G551D, W1282X and ΔF508 mutant and wild-type sequences (Table 3.3) was purchased. The standard protocol developed previously was applied to the analysis of these synthetic oligomers, using concentrations of $\mathbf{C}_{CHO}$, $\mathbf{T}_{CHO}$, $\mathbf{A}_{CHO}$, and $\mathbf{G}_{CHO}$ aldehydes that had been shown to normalise their differing incorporation selectivities (see Chapter 2.4). Thus, the synthetic DNA oligomers were analyzed in reactions designed to model the homozygous mutant, homozygous recessive and heterozygous cases for each mutation. The resulting MALDI spectra (Figures 3.2 and 3.3) allowed clear calling of the homo- and heterozygous cases, demonstrating that PNA oligomers **P5-8** were suitable for the analysis of the DNA regions of interest.

Table 3.3 DNA oligomers used in model studies for genotyping.

PNA Oligomer	Corresponding Allele [a]	Sequence (5'-3')[a]
VII	wt551	GTG GAG **G**TC AAC GA
VIII	G551D	GTG GAG **A**TC AAC GA
IX	wt1282	ACA GTG **G**AG GAA AG
X	W1282X	ACA GTG **A**AG GAA AG
XI	wt508	TAT CAT **C**TT TGG TGT
XII	ΔF508	TAT CAT **T**GG TGT TTC

[a]'wt' = wild type allele. [b]Nucleobases that lie opposite a blank position on the complementary PNA probe are shown in bold.

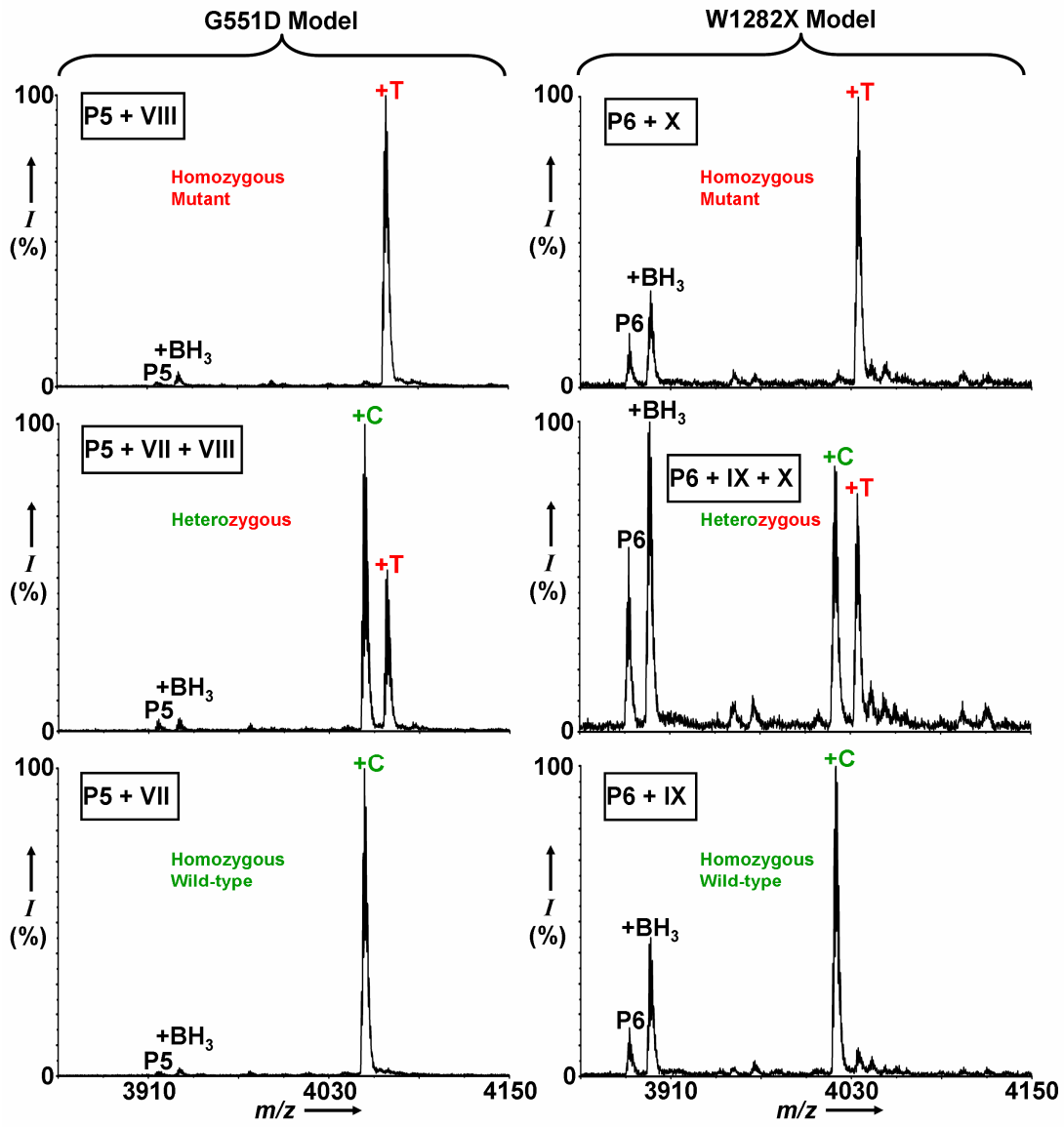


Figure 3.2 Model systems for genotyping G551D and W1282X. For both SNP models (G551D and W1282X), templated incorporation of C and/or T indicates the presence of the wild-type and/or mutant allele respectively. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.07 mM in $\mathbf{C}_{CHO}$, 0.73 mM in $\mathbf{T}_{CHO}$, 0.43 mM in $\mathbf{A}_{CHO}$ and 0.05 mM in $\mathbf{G}_{CHO}$ (ratio 1:14:8:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 5.0 μM in DNA (for the heterozygous models, 2.5 μM in each of two templates); 5.0 μM in the PNA probe; 100 mM in $NaBH_3CN$.

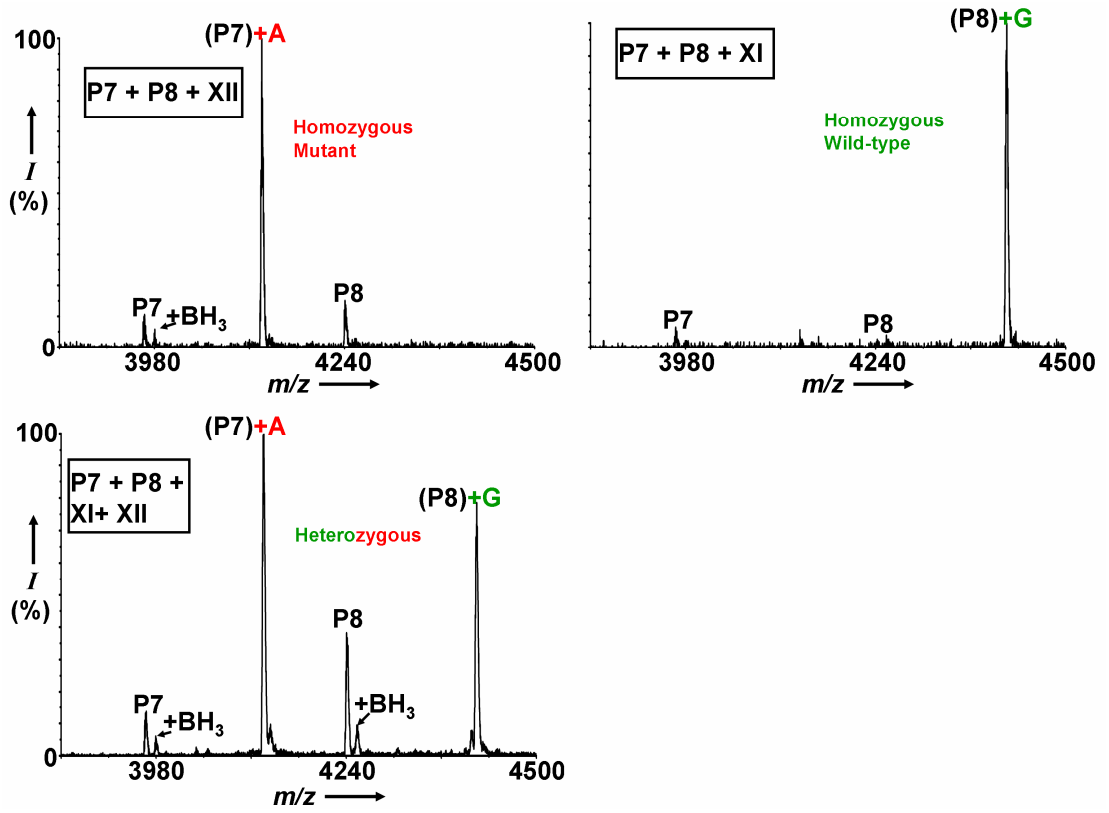


Figure 3.3 Model systems for genotyping ΔF508. Selective incorporation of A into **P7** shows the presence of the mutant allele, and G into **P8** the presence of the wild-type allele. Final concentrations in a 20 μL reaction volume: 2.8 mM pH 6.0 sodium phosphate buffer; 0.07 mM in $\mathbf{C}_{CHO}$, 0.73 mM in $\mathbf{T}_{CHO}$, 0.43 mM in $\mathbf{A}_{CHO}$ and 0.05 mM in $\mathbf{G}_{CHO}$ (ratio 1:14:8:1 $\mathbf{C}_{CHO}$:$\mathbf{T}_{CHO}$:$\mathbf{A}_{CHO}$:$\mathbf{G}_{CHO}$); 5.0 μM in DNA (for the heterozygous models, 2.5 μM in each of two templates); 5.0 μM in each PNA probe; 100 mM in $NaBH_3CN$.

To model a duplex genotyping of an individual heterozygous for G551D and W1282X, DNA oligomers **VII-X** were mixed in equal ratio and analyzed as before using an equimolar ratio of **P5** and **P6**. The resulting mass spectrum showed the selective incorporation of the expected nucleobases, although the signals for G551D analysis were stronger than for W1282X (Figure 3.4a). These could be normalized by lowering the amount of **P5** (used for G551D analysis) relative to **P6**, although importantly the concentrations of the four DNA templates were kept the same (Figure 3.4b).

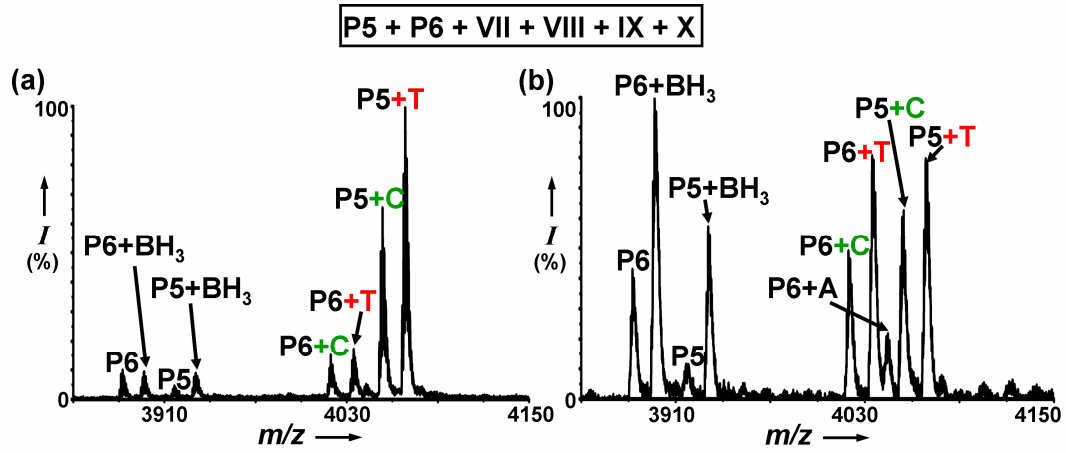


Figure 3.4 Model duplex analyses for G551D and W1282X. (a) Using an equimolar ratio of the two PNA probes, signals arising from **P6** are lower than those for **P5**. (b) Signal strengths were normalized by using a 5:3 molar ratio of **P6**:**P5**.

Dynamic chemistry was thus successfully applied to the analysis of mixtures of synthetic DNA. However, three features of the protocol developed in these studies presented potential hurdles to the analysis of genomic DNA:

1. The PNA probe concentration must always be approximately equal to or less than that of the DNA template. If a large excess of PNA was used, then any (now small) signals for the templated products were masked by the nonspecific incorporation products observed in the absence of template (see Chapter 2, Figure 2.8b). Poor yield and no selectivity were observed for reactions in which the PNA was in a 10^2-10^3 fold excess relative to DNA template (Figure 3.5). The sensitivity of this method of DNA analysis is therefore limited by the detection limit for the PNA probe(s).
2. The detection limit for the PNA following dynamic incorporation was established to be on the order of 100 femtomoles (using the triphenylphosphonium charge-tagged **P5** and DNA **VII**, by serial dilutions of the PNA/DNA and maintaining all other reaction components and conditions *as per* the model study outlined in Figure 3.2). In the absence of a charge tag the limit of detection was 1 pmole (established using acyl-capped **P1** and DNA **II**, see Chapter 2.3). The ultimate goal of the work presented in this Chapter was the analysis (by dynamic chemistry) of human genomic DNA which had been isolated from buccal (cheek) swabs using a commercial kit (Isohelix). According

to the manufacturer's instructions, 2 - 10 μg of DNA is typically obtained from a swab which (neglecting any contribution from mitochondrial and bacterial DNA) equates to approximately 1 - 5 attomoles (10^{-18} moles) of template sequence (see Appendix 1 for the derivation of this value). Given point '1' above and the detection limit established for the PNA probes of around 100 femtomoles (10^{-13} moles), then the amount of DNA obtainable from buccal swabs is a factor of 10^5 too low for direct analysis.

3. Only single stranded DNA was analyzed in model studies, because canonical PNA (containing all four bases) is unable to invade linear (B-form) double-stranded (ds)DNA.[119] However, genomic DNA is double stranded, which provides a further barrier to direct analysis.

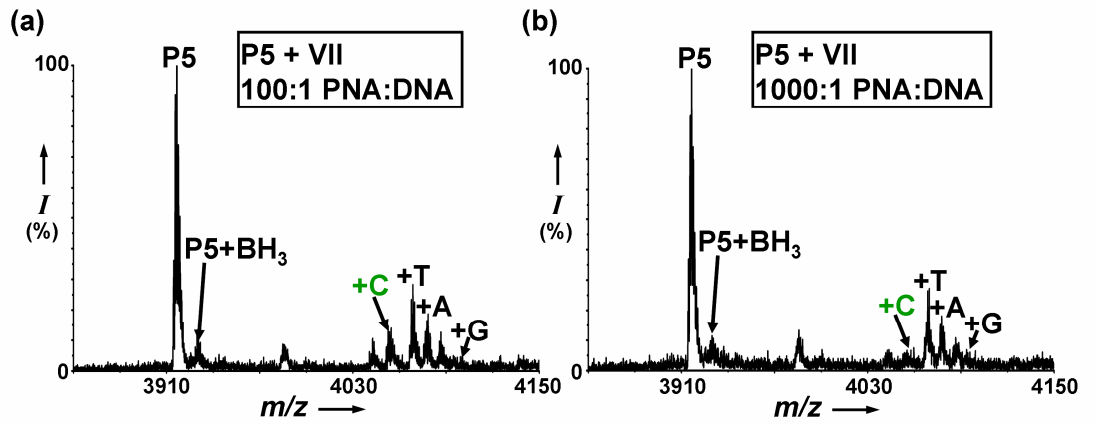


Figure 3.5 Analysis of DNA **VII** with an excess of **P5**. Conditions as per Figure 3.2, but with final concentrations of (a) 0.5 μM DNA and 50 μM PNA (a 10^2 fold excess of PNA), and (b) 0.05 μM DNA and 50 μM PNA (a 10^3 fold excess of PNA).

To analyze genomic DNA, it was therefore concluded that an amplification step would be required. PCR would be used to amplify double-stranded DNA around the region of interest, and a second asymmetric PCR step would then generate sufficient single-stranded (ss)DNA for analysis. Asymmetric PCR would be achieved by amplifying a sample of the original (symmetric) PCR mixture in a second round using just one primer, thereby generating an excess of a single strand which would serve as the template for analysis.

3.4 Analysis of 'In-House' and Commercial Genomic DNA Samples for G551D

DNA was isolated from the two scientists working on this project (samples 'FB' and 'JJ') using commercially available buccal swabs and isolation kits (Isohelix). This involved treatment of the swabs consecutively with three proprietary solutions; one to disrupt the phospholipid membranes of the cells (such solutions often contain detergent and guanidinium hydrochloride), a second containing an enzyme (protease) to break-down cellular proteins, and a third (alcohol-based) solution to precipitate the genomic DNA. The final, isolated precipitates were resuspended in a (proprietary, though possibly tris-EDTA) basic storage buffer ready for amplification.

PCR primers were chosen from a previous publication and used in a symmetric PCR to amplify a 92 bp section of double-stranded DNA in the region around the G551D mutation.[55] Gel electrophoresis of the PCR products for samples FB and JJ confirmed the presence of an amplicon of the correct length (Figure 3.6). An aliquot of each sample was then subjected to asymmetric PCR with a single primer. Attempted analysis of this PCR mix directly by dynamic chemistry was unsuccessful, probably because the mixture contains components and buffer which could hamper the reductive amination (for example if they contained amines). The products of this second PCR were therefore purified using commercially available spin-columns (Qiagen). These silica-based columns require an initial acidification of the crude PCR mix with high-salt buffer before it is applied to the column and bound. A washing step is then performed to remove enzymes, buffer, nucleotides and primers (< 40 bp in length) before the purified DNA amplicon is eluted with a buffer of $7.0 \geq pH \leq 8.5$. These columns are supplied with a pH 8.5 elution buffer, but this was substituted with pH 7.4 phosphate buffered saline (PBS) owing to the previous observation that the reductive amination is not efficient at pH 8.5 (see also Chapter 2.4, Figure 2.6). Although not optimal for reductive amination, pH 7.4 PBS allowed elution of the amplified DNA whilst still permitting analysis by dynamic chemistry without the need for further acidification (*vide infra*).

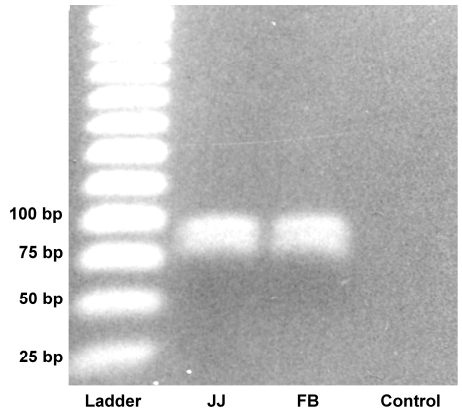


Figure 3.6 Agarose gel showing PCR amplification of 92 bp double-stranded DNA in the region around the G551D mutation for samples FB and JJ.

An aliquot of both eluted samples was used to estimate the final DNA concentrations by UV absorbance (Table 3.4). Based upon these values, it was calculated that each contained approximately 40-50 picomoles of target DNA (see Appendix 2). Subsequent analyses of these DNA samples by dynamic chemistry were performed with 40 picomoles of probe **P5**. Only $\mathbf{T}_{CHO}$ and $\mathbf{C}_{CHO}$ (using concentrations reflective of the relative incorporation selectivities of these bases) were used in order to simplify the analyses, given that only C or T incorporation would be expected for the wt551 or G551D alleles respectively. For both samples (i.e. for FB and JJ) selective incorporation of C only was observed (Figure 3.7). This was consistent with both individuals being homozygous for the wt551 allele. It should also be recorded that spin-column purification and analysis was performed directly after the initial (symmetric) PCR stage but without success, thereby supporting the assertion that single-stranded DNA is required for this method of DNA analysis.

Although the results for samples FB and JJ had been expected on the basis that neither individual has a family history of cystic fibrosis, these genotypes were not validated by established technologies, and so samples of known genotype for the G551D mutation were sought. Thus, two isolated genomic DNA samples were purchased (Coriell Cell Repositories) that had been extracted from cell cultures sourced from individuals genotyped for CF-linked mutations (Table 3.4). Both samples (‘Cor1’ and ‘Cor2’) were from individuals heterozygous for the G551D allele. Analyses of these commercial samples were performed after a single asymmetric PCR step using an unequal ratio of forward and reverse primers (i.e.

skipping the preliminary symmetric PCR step). The resulting MALDI-TOF spectra showed selective incorporation of both C and T, consistent with the known genotypes (Figure 3.8).

Table 3.4 Genomic DNA samples analyzed for the G551D mutation.

Sample	Source	Extract DNA Conc. (ng/μL)[a]	DNA Conc. after Asymmetric PCR (ng/μL)[a]	Genotype (by Dynamic Chemistry)[b]	In Agreement with Known Genotype?[c]
FB[d]	Buccal swab	58	97	N/N	n/a
JJ[d]	Buccal swab	144	108	N/N	n/a
Cor1[e]	Cell culture	371	44	G551D/N	Yes
Cor2[e]	Cell culture	360	46	G551D/N	Yes

[a]Determined from the UV absorbance. [b]'N/N' = homozygous wt551, 'G551D/N' = heterozygous for the G551D mutation. [c]'n/a' = genotype unknown. [d]Analysis performed after two step PCR (symmetric then asymmetric). [e]Analysis performed after one step (asymmetric) PCR.

It was even possible to analyze (with probes **P5**, **P7**, **P8** and all four nucleobase aldehydes) for both G551D and ΔF508 (the Coriell samples were homozygous wild-type for ΔF508) after a single duplexed asymmetric PCR in one reaction (Figure 3.9). Amplification of a 173 bp region around the ΔF508 mutation was achieved using primers again selected from the literature.[55] However, this single-step asymmetric amplification gave weaker signals than were observed using the two step procedure employed for the 'in-house' samples and in later work (see Chapter 3.5 and 3.6). Furthermore, it could not be reliably reproduced using samples sourced from buccal swabs, most likely due to differences in the purity and concentration of the input genomic DNA.

Gel electrophoresis after the duplexed asymmetric PCR for the Coriell samples confirmed the presence of the double-stranded ΔF508 and G551D amplicons. Other faint bands were observed, but they could not be unambiguously assigned to single-stranded (ss)DNA (Figure 3.10). However, this may be due to the relatively low affinity of ethidium bromide for single-stranded DNA *versus* double-stranded DNA.

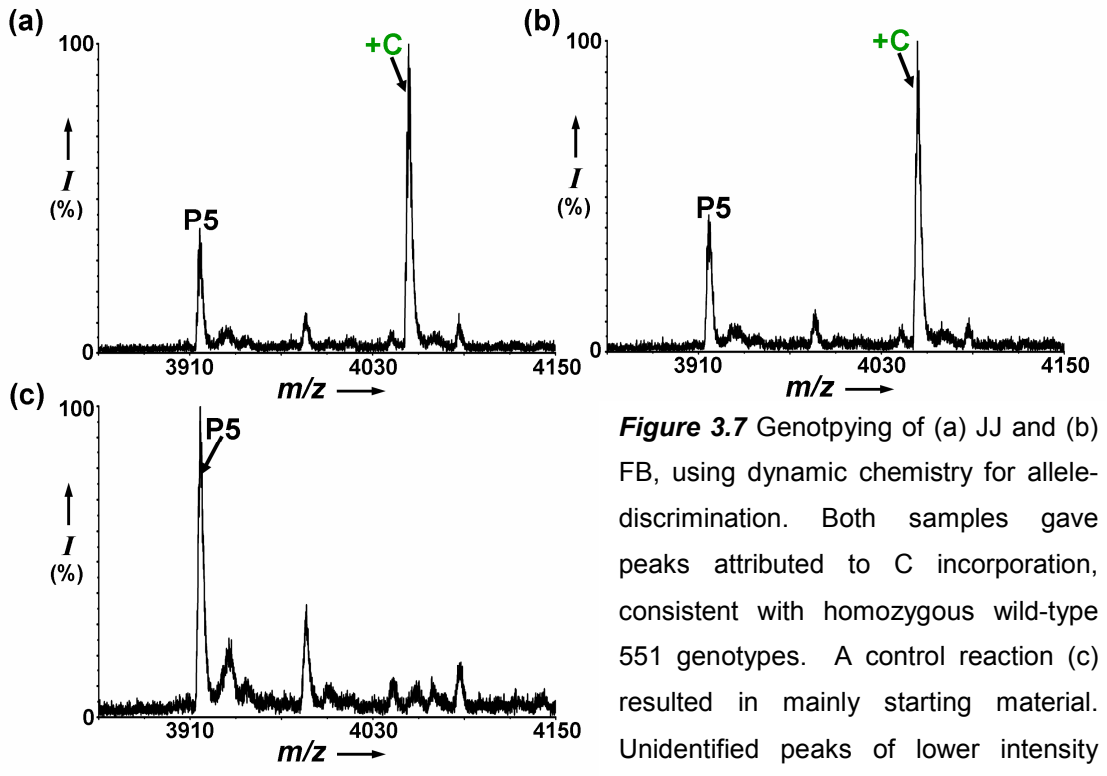


Figure 3.7 Genotpying of (a) JJ and (b) FB, using dynamic chemistry for allele-discrimination. Both samples gave peaks attributed to C incorporation, consistent with homozygous wild-type 551 genotypes. A control reaction (c) resulted in mainly starting material. Unidentified peaks of lower intensity were present in each case but did not impede genotyping.

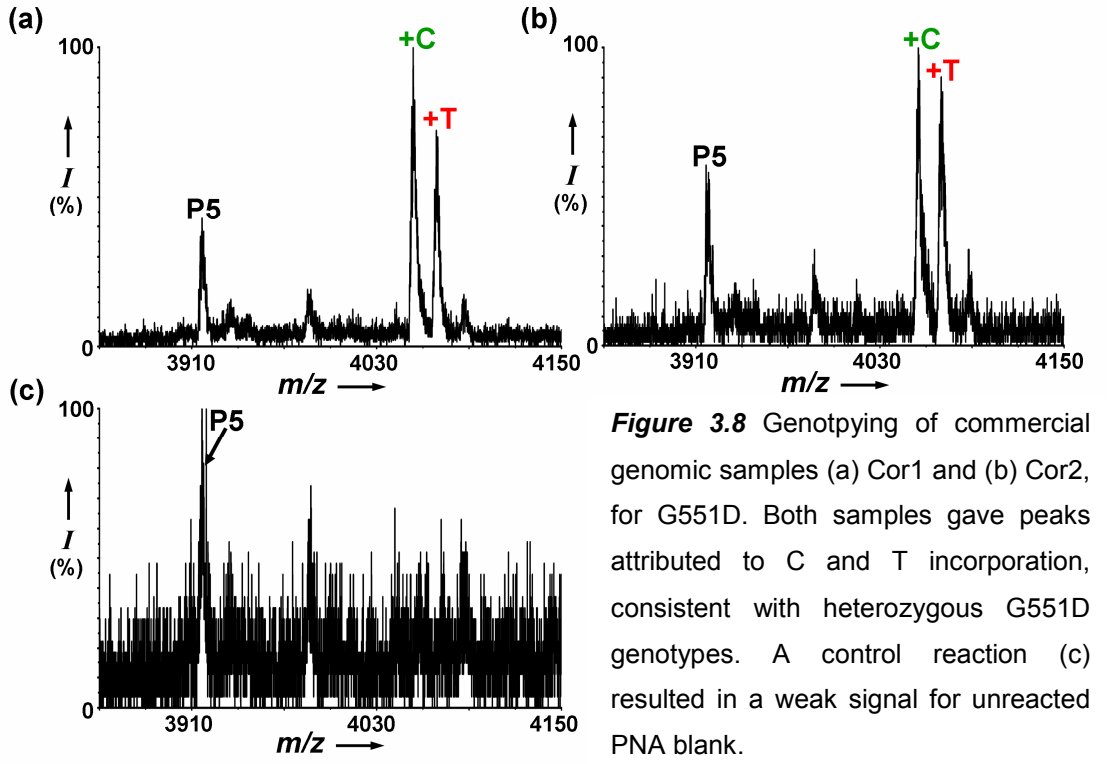


Figure 3.8 Genotpying of commercial genomic samples (a) Cor1 and (b) Cor2, for G551D. Both samples gave peaks attributed to C and T incorporation, consistent with heterozygous G551D genotypes. A control reaction (c) resulted in a weak signal for unreacted PNA blank.

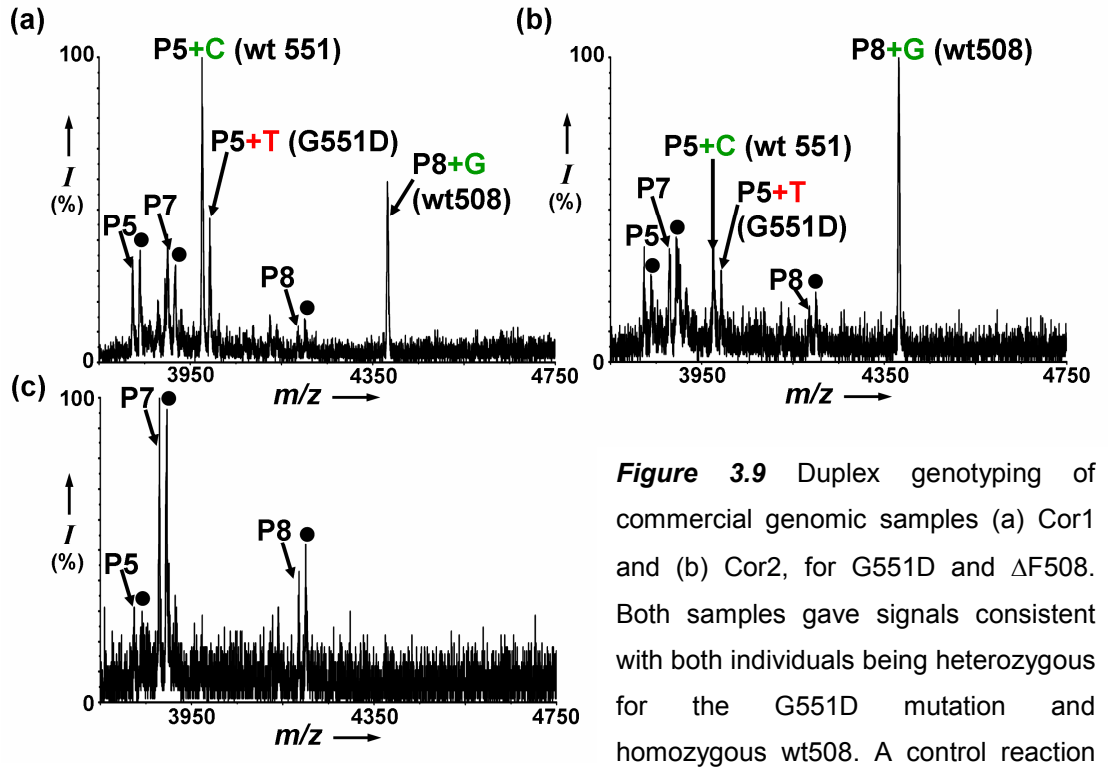


Figure 3.9 Duplex genotyping of commercial genomic samples (a) Cor1 and (b) Cor2, for G551D and ΔF508. Both samples gave signals consistent with both individuals being heterozygous for the G551D mutation and homozygous wt508. A control reaction (c) resulted in weak signals for unreacted PNA blank. ● = BH_3 adduct.

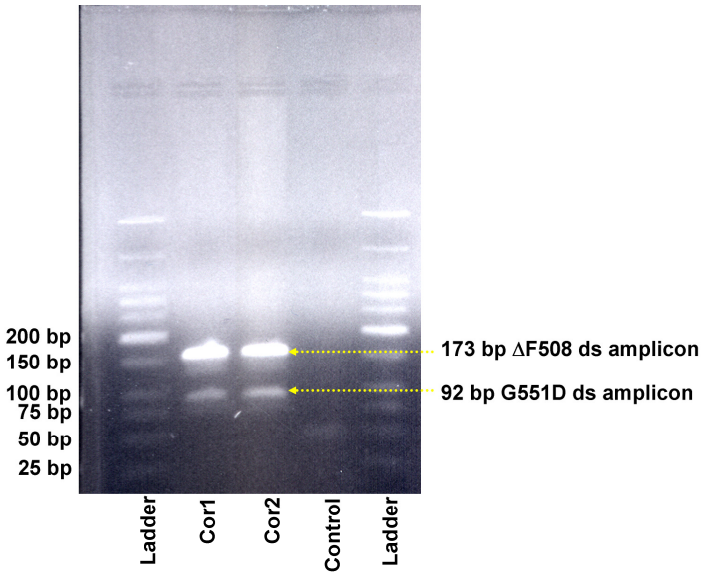


Figure 3.10 Agarose gel after duplex asymmetric PCR for commercial samples Cor1 and Cor2. 92 bp and 173 bp double-stranded DNA in the regions around the G551D and ΔF508 mutations respectively are observed, but single-stranded DNA cannot be readily distinguished.

3.5 'Singleplex' Analysis of Clinical Genomic DNA Samples for G551D and ΔF508

Genomic DNA was obtained from twelve individuals with cystic fibrosis who had been genotyped for CF-linked mutations using established methods. The technology currently used by the South East Scotland Genetic Service (Western General Hospital, Edinburgh) to test for CF mutations is an enzymatic amplification-refractory mutation system (ARMS, supplied by Gen-Probe as the Elucigene™ CF-EU2 kit; see Chapter 1.1.6 for a description of the ARMS method as applied to Scorpion primers) that uses capillary electrophoresis and fluorescence detection (on an Applied Biosystems Genetic Analyzer) for read-out. The CF patients provided buccal (cheek) swabs from which the genomic DNA was isolated as before. Regions surrounding the G551D and ΔF508 mutations were then amplified in separate ('singleplex') PCR reactions using the two-step protocol (i.e. symmetric then asymmetric, Figure 3.11). PCR primers were as above for amplification of the G551D and ΔF508 templates (Chapter 3.4). For one of the samples, gel-shift electrophoretic analysis using **P5** after the asymmetric PCR stage for G551D also suggested the presence of the target single-stranded DNA, together with some unidentified amplicons possibly formed as a result of primer-dimer formation or other nonspecific amplification (Figure 3.12).

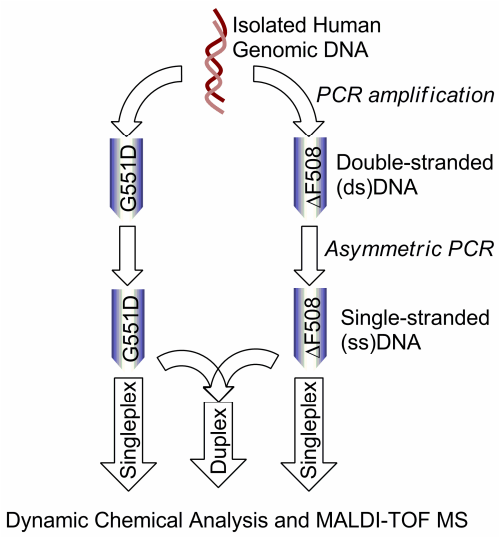


Figure 3.11 PCR steps used to generate single-stranded DNA for analysis of clinical CF samples by dynamic chemistry.

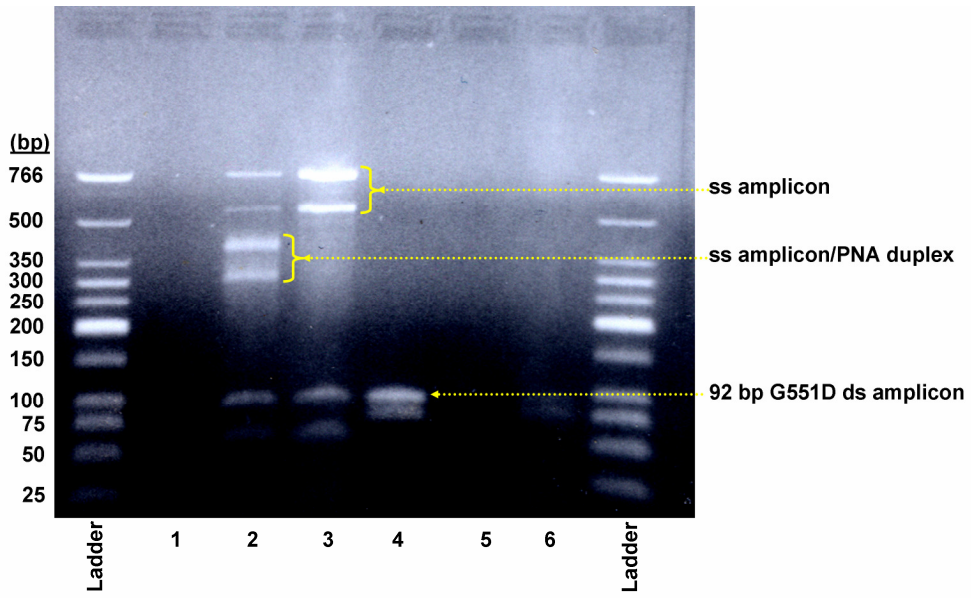


Figure 3.12 Gel electrophoresis of PCR products during the analysis of sample CF2 for G551D: Lane 1 = **P5**; 2 = **P5** + product of second, asymmetric PCR; 3 = product of second asymmetric PCR; 4 = product of first, symmetric PCR (dsDNA); 5 = control for second, asymmetric PCR; 6 = control for first, symmetric PCR. The addition of **P5** to the product of the asymmetric PCR results in a shift in two of the bands, suggesting that these may be represent single-stranded DNA amplicons.

In a blind trial (such that only the clinician who collected the samples was aware of the previously determined genotypes), analysis for the G551D and ΔF508 mutations were performed in singleplex fashion (Figure 3.11) following the two-step asymmetric PCR protocol. Of the twelve samples analyzed, five were found to be ΔF508 homozygotes, four were G551D/ΔF508 heterozygotes, two were G551D heterozygotes who did not possess the ΔF508 allele, and one was a G551D homozygote (see Table 3.5 and Figure 3.13a-c, e-g). The results were in complete agreement with the known genotypes determined using established technologies. Peaks were always observed for unreacted PNA probes which could be used as internal calibrants. For control reactions in the absence of any DNA template (Figure 3.13d, h), no nucleobase incorporation products were detected.

Table 3.5 Analysis of clinical samples for G551D and ΔF508 CF-linked mutations.

CF Sample	Extract DNA Conc. (ng/μL)[a]	Result of G551D Genotyping[b]	Result of ΔF508 Genotyping[c]	Known Genotype[d]
1	112.0	G551D/N	ΔF508/N	ΔF508/G551D
2	18.5	G551D/N	ΔF508/N	ΔF508/G551D
3	6.4	G551D/N	ΔF508/N	ΔF508/G551D
4	5.5	N/N	ΔF508/ΔF508	ΔF508/ΔF508
5	17.0	G551D/G551D	N/N	G551D/G551D
6	19.0	G551D/N	N/N	P67L/G551D
7	142.0	G551D/N	ΔF508/N	ΔF508/G551D
8	110.5	G551D/N	N/N	G542X/G551D
9	157.5	N/N	ΔF508/ ΔF508	ΔF508/ΔF508
10	377.5	N/N	ΔF508/ ΔF508	ΔF508/ΔF508
11	145.5	N/N	ΔF508/ ΔF508	ΔF508/ΔF508
12	107.0	N/N	ΔF508/ ΔF508	ΔF508/ΔF508

[a]Determined from the UV absorbance. [b]'N/N' = homozygous wt551, 'G551D/N' = heterozygous for the G551D mutation, 'G551D/G551D' = homozygous for the G551D mutation. [c]N/N' = homozygous wt508, 'ΔF508/N' = heterozygous for the ΔF508 mutation, 'ΔF508/ ΔF508' = heterozygous for the ΔF508 mutation. [d]Mutations P67L and G542X (present in samples CF6 and CF8) were not tested for in this study.

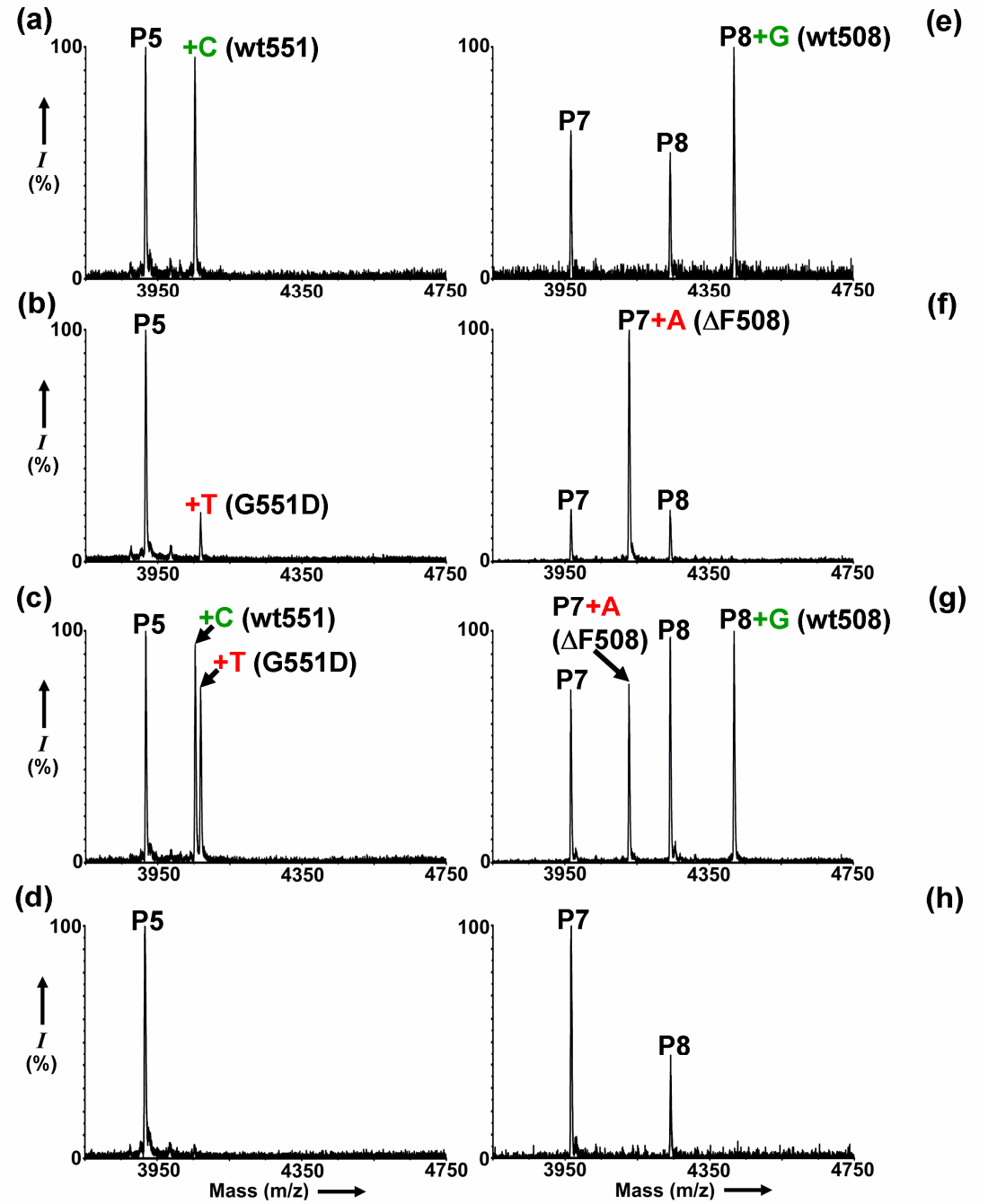


Figure 3.13 Representative MALDI-TOF spectra for genotyping: (a) C incorporation into **P5** showing an individual homozygous for the wild-type 551 allele (wt551); (b) **P5**+T showing an individual homozygous for the G551D mutant allele; (c) **P5**+C and **P5**+T together showing an individual heterozygous for the G551D mutation; (d) a control reaction for G551D analysis showing unreacted **P5** only; (e) **P7**+G showing an individual homozygous for the wt508 allele; (f) **P8**+A showing a genotype of homozygous ΔF508; (g) **P7**+A and **P8**+G together showing an individual heterozygous for the ΔF508 mutation; (h) a control reaction for ΔF508 analysis showing unreacted **P7** and **P8**.

Mean allelic ratios for heterozygotes (averaged across the 5 spectra recorded for each singleplex analysis) were 0.46 (SD = 0.12; n = 30) for G551D (A/G) and 0.79 (SD = 0.25; n = 20) for ΔF508 (C/T) analysis. Peaks resulting from DNA-templated incorporation were always > 5 % relative intensity. No peaks corresponding to nucleobase incorporation were observed with lower intensities, with one exception. In this case, when analyzing a sample for ΔF508, a peak of mass corresponding to '2 + A' (indicating presence of the ΔF508 mutant allele) was observed at < 5 % relative intensity in 3 of the 5 recorded spectra. However, given the low signal strength, the peak was disregarded and the genotype was called as homozygous wild-type. Furthermore, repeat analysis of this sample for ΔF508 firmly corroborated this assertion (see Figure 3.13e)

The genomic DNA used in this study was isolated from buccal swabs with Isohelix kits, but no further purification was performed, and the concentration of DNA was not normalized prior to input into the first PCR cycle. However, this made no discernable difference to the quality of the final mass spectra, despite the concentration of genomic DNA ranging from 5.5 ng/ μL to 377 ng/ μL across the samples (Table 3.5).

3.6 Duplex Analysis of Clinical Genomic DNA Samples for G551D and ΔF508

The duplex analysis of one of the G551D/ΔF508 heterozygous samples was performed by repeating the singleplex amplification of ssDNA for G551D and ΔF508, then combining the final crude PCR mixtures and purifying them into a single batch for analysis with all three PNA probes. The resulting mass spectrum permitted clear determination of the genotype for both mutations (Figure 3.14a). Allelic ratios for this duplex analysis (averaging peak intensities across the 5 spectra) were 0.28 (SD = 0.03; n = 5) for G551D (G/A) and 0.86 (SD = 0.09; n = 5) for ΔF508 (C/T). Again, a control reaction showed no incorporation in the absence of DNA template (Figure 3.14b).

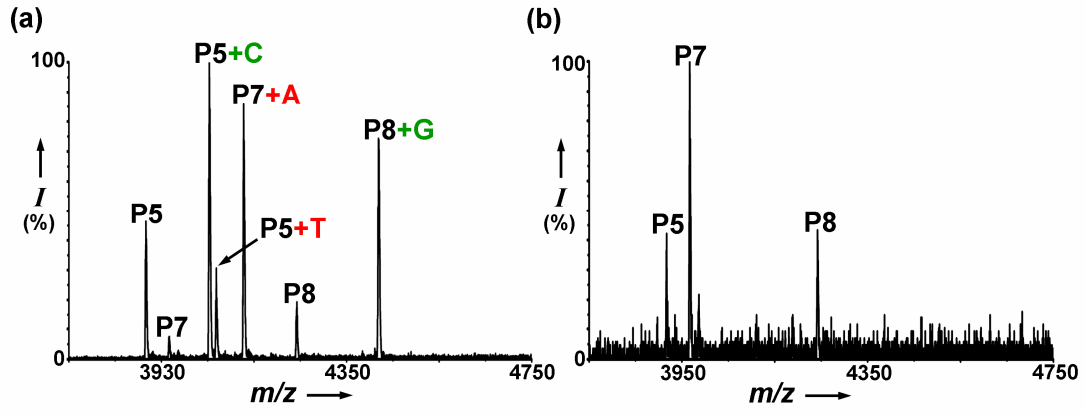


Figure 3.14 (a) Representative MALDI-TOF spectrum for duplex genotyping of G551D and ΔF508 showing an individual heterozygous for both mutations. Peaks are observed for all four possible incorporation products (**P5**+C and **P5**+T report wt551 and G551D alleles respectively; **P7**+A and **P8**+G report ΔF508 and wt508 alleles respectively). (b) A control duplex analysis showing unreacted **P5**, **P7** and **P8**.

3.7 Discussion and Conclusions

Dynamic chemistry can now be added to the lexicon of methods to generate allele-specific products for genotyping (outlined in Chapter 1.1.5-7). Twelve individuals were successfully genotyped for the G551D and ΔF508 CF-linked mutations in a blind trial of this approach with no false negatives or positives. Dynamic chemistry was thus shown to be applicable to the genotyping of SNPs and indels, both in a singleplex and duplex format. This distinct method holds several advantages over the established allele discrimination assays. Specifically, there is no need for the stringent optimization and control of conditions associated with the hybridization of ASO probes and subsequent washing steps,[120] because the PNA 'blank' probes clamp either side of the polymorphic nucleotide position irrespective of the allele present. When compared to methods that rely upon enzymatic extension for allele-discrimination, dynamic chemistry removes the need for an enzyme. This is of particular advantage in assays that use ddNTPs, as these unnatural triphosphates are not incorporated as selectively as dNTPs by natural enzymes, and modified enzymes are often required to overcome this.[121] Furthermore, assays like iPLEX[72] and SPC-SBE[71] require an additional enzymatic step with shrimp alkaline phosphatase (SAP) and exonuclease I to degrade dNTPs and primers before a final extension with a mixture of ddNTPs; dynamic chemistry obviates these enzymatic

reactions, with concomitant advantages in terms of cost and analysis time. The ARMS technology used currently by the NHS in Edinburgh for genotyping CF-linked mutations relies upon the enzymatic extension of allele-specific primers.[57]

When compared to other methods relying upon MALDI-TOF for read-out, dynamic chemical allele-discrimination with PNA has an additional advantage over DNA-based assays, in that PNA is particularly well suited to detection by MALDI using matrices suitable for peptide analysis.[122] DNA and RNA are more difficult to analyze; indeed, the GOOD assay uses an additional alkylation step and a permanent charge tag to address this. The application of PNA to genotyping by MALDI-TOF MS has been reported by others, but these assays use ASO probes for discrimination, with the associated disadvantages touched on above.[97]

The American College of Medical Genetics recommends testing for a panel of 23 mutations for CF carrier screening, which between them account for the majority of cases.[123] The assay reported herein analyzes for two of the more important mutations, which is evidently insufficient for full CF carrier screening. Indeed, a recent test was reported which screened for 108 CF-linked mutations, using enzymatic primer extension and MALDI-TOF MS.[124] It is not proposed therefore that this assay *per se* be utilized for this purpose. Instead, it has been demonstrated that dynamic chemistry can be applied to the generation of allele-specific products for genotyping, which argues that the scope of this method of allele discrimination merits further investigation.

Given a successful duplex analysis, then it is reasonable to expect that the multiplex analysis of several mutations would be possible with probes designed to prevent overlap in the mass spectrum. In this respect, peaks due to the presence of unreacted PNA probes can be regarded as an advantage in the sense that they serve as useful internal calibrants. However, they also 'clutter' the spectrum, which would limit the number of mutations which could be multiplexed. It is interesting that analysis of the clinical DNA samples by dynamic chemistry generated MALDI spectra containing virtually no borane adducts (Figure 3.13) which had complicated spectra during model studies with synthetic DNA. The reason for this is unclear, although perhaps it is related to differences in PNA/DNA concentration or the lengths of the templating DNA strands. It should also be noted however that borane

adducts were still observed during the fully duplexed analysis of the commercial genomic DNA samples (Figure 3.9).

It has been shown that indels can be analyzed by dynamic chemistry if two PNA probes are used. This is a disadvantage when compared to primer extension assays which could achieve this with a single probe. However, indels are much rarer than SNPs in the genome, and the use of an additional probe for each is unlikely to limit the utility of dynamic chemistry for allele-discrimination. Furthermore, it may be that indel analysis would be possible with the use of a single probe if universal bases (e.g. hypoxanthine) were built in to the *N*-terminal side of the 'blank' position of the PNA.[125] A second alternative could involve incorporation of the 'blank' site directly at the *N*-terminus, but this has been shown to give poorer dynamic selection by others investigating dynamic chemistry on PNA/PNA duplexes.[112]

The mean allelic peak ratios for the heterozygous individuals genotyped in this study differed substantially from unity. However, the detection of an incorporation product clearly signified the presence of the associated allele, so there was no urgent need to address this. It has been shown previously in model systems, however, that the peak ratios for 'heterozygotes' can be brought closer to unity by varying the initial concentrations of the aldehyde-modified nucleobases (see Chapter 2.4). Differences in allelic ratio between the model and clinical studies may have arisen as a result of differences in the reaction conditions (e.g. pH and PNA/DNA concentration). The allelic ratio showed greater variation for ΔF508 analysis than for G551D. This is to be expected given the additional probe needed for indel detection, and it is likely to be more difficult to improve the peak ratios by altering aldehyde concentrations in this case.

MALDI-TOF allows direct read-out of allele-specific products for multiplex genotyping. However, there is no reason why dynamic chemistry would not also be amenable to indirect fluorescence detection using labelled nucleobase aldehydes, so long as the hydrogen bonding motif is still available for selection through Watson-Crick base-pairing. Furthermore, it may be hypothesized that the method for allele-discrimination reported herein could be readily extended to direct RNA analysis, without the need for reverse transcription into cDNA libraries. RNA provides a source of single stranded nucleic acid for analysis by PNA probes, and PNA/RNA

hybridization has been reported widely.[111] Indeed, RNA-templated incorporation has already been demonstrated using a synthetic RNA template (see Chapter 2.7).

The assay described here relies upon polymerase chain reaction, with all of its inherent problems, for DNA amplification[126]. Since nucleobase incorporation will only occur in the presence of templating DNA, any false positives or negatives will arise from the enzymatic amplification stage. However, this dependence on PCR is shared by virtually all genotyping technologies currently available. A notable exception is the INVADER assay,[64] although it is important to note that this method requires the input of an often prohibitively large amount of genomic DNA.[127] The requisite of ssDNA necessitates an additional step, although one which is also a drawback for other assays. An example is the solid-phase capture-single base extension (SPC-SBE) method, which overcomes this through the use of biotinylated primers and expensive streptavidin-functionalized solid supports. Asymmetric PCR may be performed in a single step as opposed to the two reported here,[128] although multiplex asymmetric PCR is harder to achieve (as found in this study, necessitating the use of separate PCR steps to reliably generate double- and single-stranded DNA for analysis of each mutation). Direct analysis of dsDNA may be possible through the use of conditions which favour PNA/DNA over DNA/DNA hybridization (such as low salt concentrations),[129] or modifications to the PNA backbone which have been shown to result in helical pre-organization of the PNA to favour duplex invasion.[119]

The need for a PCR step in genotyping by dynamic chemistry is dependent on the sensitivity of the detection method and the concentration of the templating nucleic acid. It is reasonable to predict, for example, that a realistic lowering of the detection limit to the 10^{-15} mol (1 femtomole) level would allow direct profiling of the higher copy number miRNAs by a dynamic chemical approach.[130]

CHAPTER 4
Dynamic Chemistry as a Tool for the Discovery of Non-Natural Nucleobases

4.1 Introduction

The genetic 'alphabet' of A-T/U and G-C Watson-Crick base-pairs is universally conserved in nature. Increasingly, efforts are aimed at expanding this genetic alphabet through the use of non-natural nucleobases.[131-133] A long-term goal of such research is the development of non-natural base-pairs which can be faithfully replicated, transcribed and translated *in vivo*.[134] Semi-synthetic organisms could thus be generated possessing novel functionality for biotechnological applications. In the more immediate future, however, non-natural nucleobases hold great potential for *in vitro* applications, such as molecular diagnostics, and antisense gene therapy.

Early work in this field by Benner and co-workers focused on the rational design of new base-pairs through alternative hydrogen-bonding motifs.[135-137] A so-called 'artificially expanded genetic information system' (AEGIS) was introduced which has already found applications in molecular diagnostics. For example, the *iso*G-*iso*C base-pair (Figure 4.1a) was used successfully to improve sensitivity in a diagnostic assay for the human immunodeficiency virus (HIV), which led to an FDA-approved AEGIS-based assay for monitoring HIV and hepatitis viral loads.[138, 139] This non-natural base-pair has also been used for real-time PCR and genotyping (including CF genotyping).[140, 141] One problem associated with the 'first-generation' AEGIS base-pairs such as *iso*G-*iso*C is their propensity to tautomerize, which results in reduced binding selectivity (e.g. *iso*G can form an enolic tautomer with the correct hydrogen-bonding pattern for T/U recognition).[142] However, a more recent non-natural base-pair (P-Z, Figure 4.1b) has been reported that does not suffer from this limitation. This P-Z base-pair has been found to contribute more to duplex stability than G-C base-pairs and exhibits greater mismatch discrimination *versus* natural nucleobases.[143-145] P-Z can be replicated by polymerase enzymes, and primers incorporating these nucleobases have been shown to give improved performance *versus* natural nucleobases when used as tags during nested PCR.[146] Benner and co-

workers have also recently reported a 'self-avoiding molecular-recognition system' (SAMRS), employing non-natural bases that do not recognize one-another, but bind selectively to natural nucleobases.[147] This provides a means to avoid primer interactions during PCR, and was found to give improved results during multiplexed PCR.

Other approaches to the design of non-natural base-pairs have been described, including expanded hydrogen-bonded motifs (i.e. with four hydrogen-bond recognition sites, Figure 4.1c)[148] and size-expanded nucleobases (so-called 'expanded' (x)DNA and 'wide' (y)DNA, Figure 4.1d and 4.1e).[149] However, a more radical approach to expanding the genetic alphabet dispensed with hydrogen-bonding altogether and instead focused on the development of hydrophobic base-pair analogues. This builds on pioneering work by Kool and co-workers, who demonstrated that hydrogen-bonds are not essential for DNA replication or hybridization.[150-152] Nucleobase analogues of this type rely upon hydrophobic effects and shape complementarity for duplex stabilization and selective base recognition. Eliminating the need for templating hydrogen-bonds has greatly widened the scope for design of novel base-pair analogues, and has inspired many further developments in this direction, most notably by Romesberg, Schultz, Hirao and co-workers.[153-157] Examples of hydrophobic nucleobase analogues include a propynyl-substituted derivative (PICS, Figure 4.1f) which can self-pair and increase duplex stability even over a G-C pair,[158] and a methoxynaphthyl (NaM)-5-methylthioisocarbostyril (5SICS) base-pair which can be replicated in DNA (Figure 4.1g) and was developed following an initial screen of 3600 candidate molecules.[134]

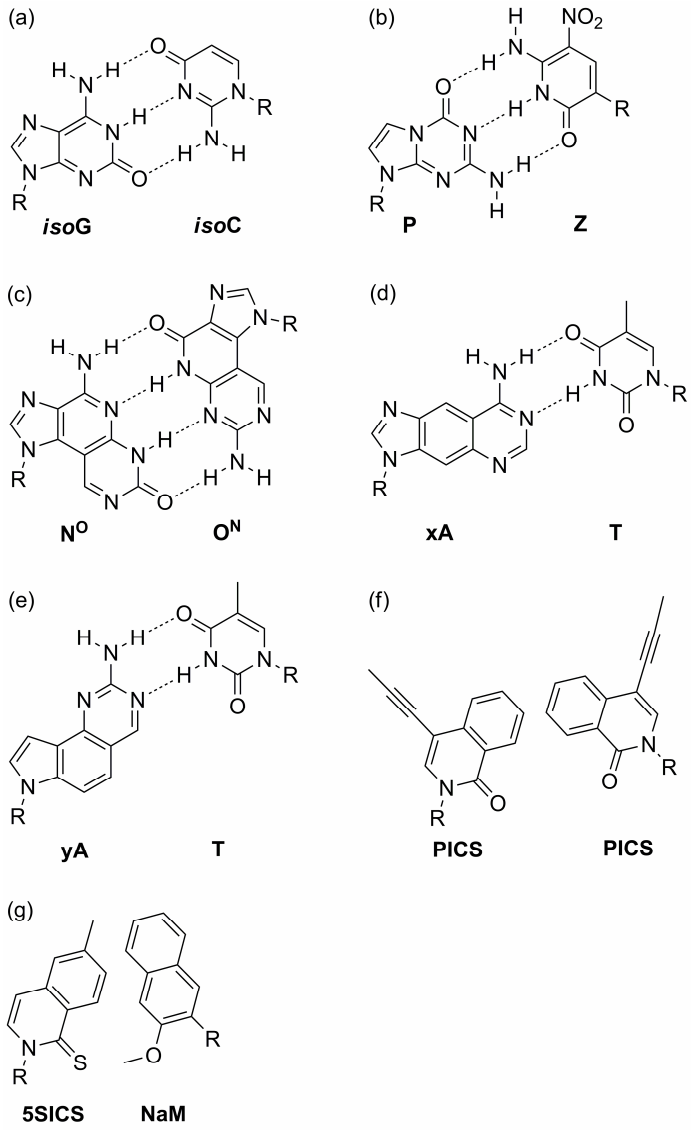


Figure 4.1 (a) AEGIS nucleobases introduced by Benner and co-workers, possessing a DDA-AAD pattern of hydrogen-bonding not found in natural nucleic acids (D = hydrogen-bond donor, A = hydrogen-bond acceptor). (b) A more recent AEGIS base-pair developed to remove the tautomerization observed in *iso*G-*iso*C. (c) A base-pair analogue displaying a DADA-ADAD pattern of 4 hydrogen-bonds. N^O-O^N refers to '(amin-oxo)-(oxo-amino)', a description of the hydrogen-bonding donor/acceptor at each end of the pattern. (d) An xDNA nucleobase (xA, shown paired with natural T) wherein the Watson-Crick hydrogen-bonding edge is distanced from the backbone by benzo fusion. (e) Benzo fusion with a different geometry generates yDNA. (f) A hydrophobic, propynylisocarbostyril (PICS) base-pair which lacks hydrogen-bonds for recognition. (g) A predominantly hydrophobic base-pair which can be enzymatically replicated. The *ortho* substituents are hydrogen-bond acceptors which aid enzymatic DNA extension following incorporation of one of these unnatural bases.

In the context of PNA, a number of nucleobase analogues have been reported that increase duplex melting temperatures with potential for improved recognition of target strands in molecular diagnostics and antisense technologies. Examples include 2,6-diaminopurine (D, Figure 4.2a) which has an extra hydrogen-bond donor (relative to adenine) with which to bind thymine,[159] 7-chloro-1,8-naphthyridin-2-(1*H*)-one ('bicyclic T' or bT, Figure 4.2b) which increases duplex stability (relative to thymine) through additional hydrophobic interactions,[160, 161] and 9-(2-guanidinoethoxy)phenoxazine (G-clamp, Figure 4.2c), which can form five hydrogen-bonds with guanine.[162]

Encouraged by the observation that templated base-filling reactions give selective incorporation of the natural nucleobases (Chapters 2 and 3), it was hypothesized that this chemistry could serve as a tool for the rapid screening and discovery of nucleobase analogues with improved binding to and selectivity for the natural bases. In this way, novel nucleobases could be screened without the need for their incorporation into full-length PNA (or DNA) oligomers. It was envisaged that the aldehydes required for this method of screening would be more easily synthesized than the Fmoc/Bhoc protected PNA monomers or DNA phosphoramidites necessary for oligomer synthesis. It might even be possible to identify novel base-pairs through the use of two PNA 'blank' oligomers if they were designed such that the 'blank' positions lay opposite one another in a PNA/PNA duplex. In relation to earlier work on genotyping (Chapters 2 and 3), it would also be desirable to find a replacement aldehyde for $\mathbf{T}_{CHO}$ and $\mathbf{A}_{CHO}$ to improve on the lower incorporation selectivity associated with these bases.

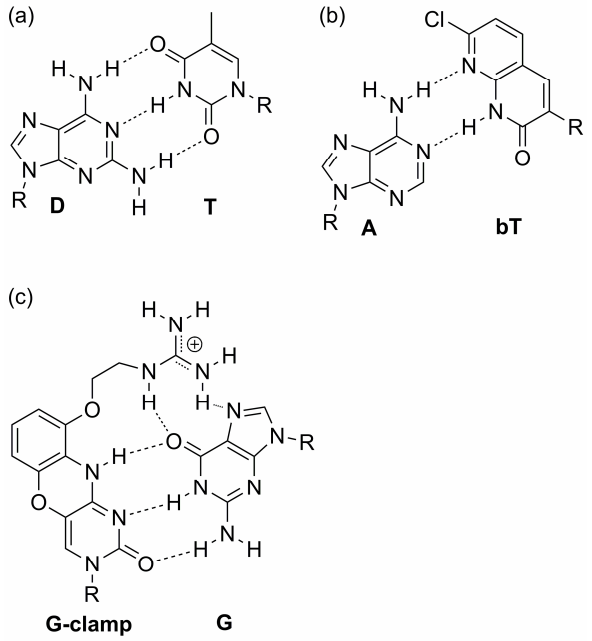


Figure 4.2 (a) 2,6-diaminopurine hybridizes more strongly with thymine as a result of an additional hydrogen-bond. (b) Increased duplex stability can also be affected through additional base-stacking interactions, as exemplified by a bicyclic thymine (bT) analogue. (c) The G-clamp base possesses guanidinium functionality which can form two additional hydrogen-bonds with G, whilst imparting a positive charge which provides additional stabilization through electrostatic interactions with the negatively charged sugar-phosphate backbone of DNA.

The protocol developed previously for templated incorporation could thus be applied to the discovery of new nucleobase analogues, again using mass spectrometry as a read-out tool. Two approaches were planned: firstly, a library of aldehydes available in-house would be screened for selective recognition of the natural nucleobases; secondly, a more targeted set of aldehyde-modified nucleobases would be synthesized and compared to the canonical nucleobases for incorporation selectivity. Any nucleobase found to strongly and/or selectively incorporate opposite a natural nucleobase could then be built into a full-length PNA oligomer for PNA/DNA duplex melting temperature (T_m) measurements, to determine whether or not this dynamic chemical approach gave a reliable estimate of base-pair stability and discrimination.

4.2 Library Screening for Nucleobase Analogues

A small library of eighteen aldehydes available in-house (Figure 4.3) was screened for template-selective incorporation using the protocol developed previously (Chapter 2.4). The aldehydes were chosen based upon in-house availability, the presence of aromaticity (to allow base-stacking), and ease of discrimination of the possible incorporation products in a mass spectrum (i.e. masses were chosen to avoid overlap). The aldehydes were screened for selective incorporation into PNA **P1** (Chapter 2.3, Table 2.1) using DNA templates **I-IV** (Chapter 2.3, Table 2.2).

The resulting mass spectra showed peaks associated with incorporation only of aldehydes **27** and **28** (Figure 4.4). However, the molecular ion associated with the incorporation of cinnamaldehyde **28** was always 2 Da greater in mass than the calculated value, although the isotopic distribution was poorly resolved. The reason for this was undetermined, but it is likely that the alkene functionality in **28** has also been reduced in the presence of excess cyanoborohydride. 1,4-addition of hydride to the conjugated iminium species generated by **28** would give an enamine which could be protonated under the acidic conditions of the experiment (at the carbon β to the aromatic ring) to afford a second iminium intermediate, reduction of which would generate a peak with the observed mass.[163] The mass was too low to result from incorporation of aldehyde **22** which has the next highest mass (6 Da greater than that of **28**), but the increase in 2 Da brings this peak close enough to overlap with that calculated for incorporation of **22**. Although the isotope resolution was not good enough to eliminate the possibility of a small amount of **22** incorporation, this seems unlikely in view of the lack of incorporation of similar aldehydes **26** and **29**.

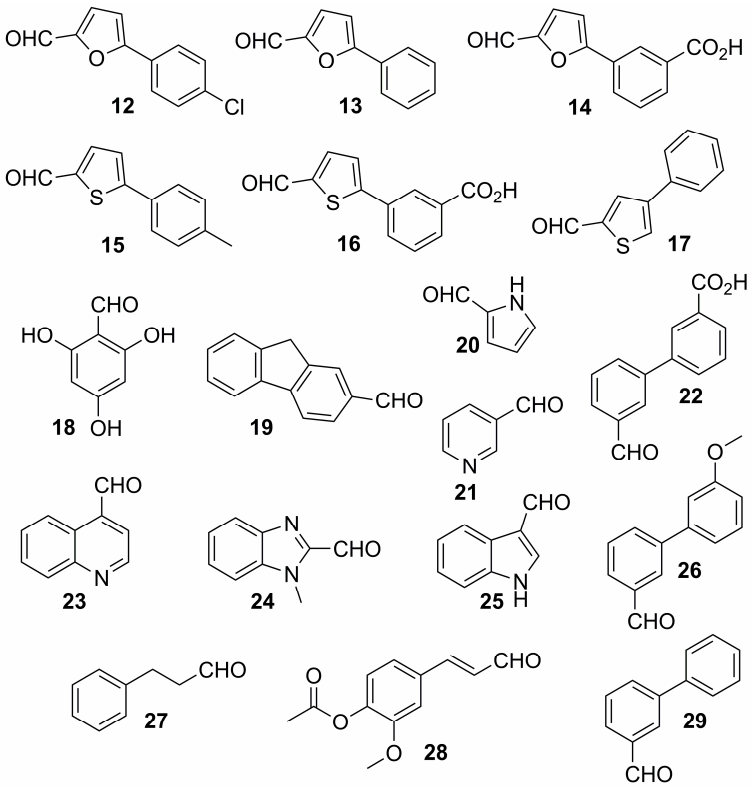


Figure 4.3 Aldehyde mixture screened for incorporation. Compounds **12-17**, **22**, **26**, and **29** had been previously synthesized as part of a separate medicinal chemistry project. All other aldehydes were commercially available.

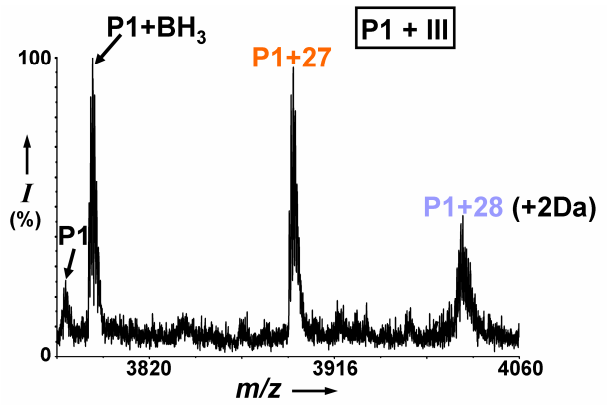
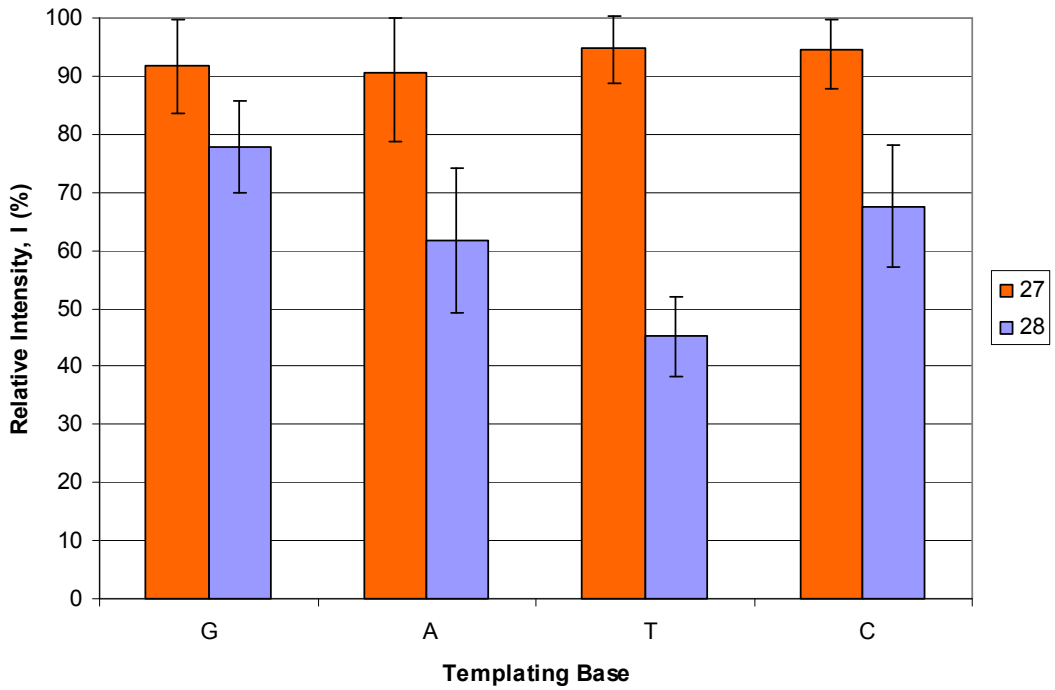


Figure 4.4 Representative mass spectrum obtained after dynamic incorporation of a library of aldehydes, showing peaks attributable to **27** and **28** incorporation (DNA **III** represents the T template). Although the peak attributed to **28** incorporation had a mass 2 Da higher than the calculated value, this is probably due to reduction of the alkene functionality proceeding *via* an enamine intermediate. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.48 mM in each of the eighteen aldehdyes; 5.0 μM in DNA template; 5.0 μM in **P1**; 100 mM in $NaBH_3CN$.



Graph 4.1 Mean peak intensities (of the most common isotope, relative to the most intense peak) resulting from the templated incorporation of aldehydes **27** and **28** from an equimolar library. Error bars indicate the standard deviation across ten mass spectra (five from each duplicate analysis).

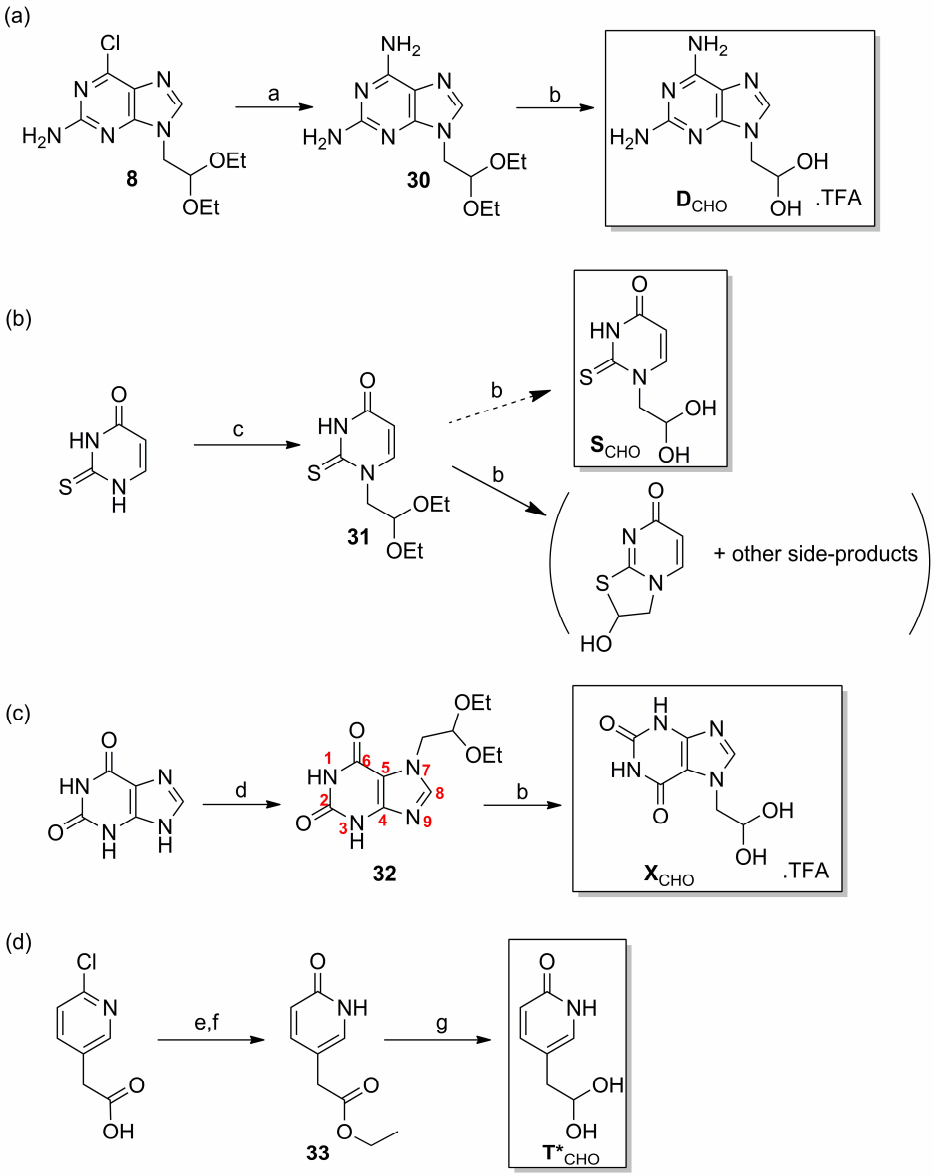
No additional incorporation products were distinguishable above the background noise in the MALDI spectra. This is remarkable, given that each of the eighteen aldehydes screened could potentially react with the secondary amine on the PNA probe in a reductive amination. Both **27** and **28** (a propionaldehyde and cinnamaldehyde derivative respectively) possess an aromatic ring connected to the aldehyde functionality by a two-carbon linkage. For each of the other compounds screened, the aldehyde functionality was attached directly to an aromatic ring, which may have been disfavourable for base-stacking and hence templated incorporation.

Aside from a slightly lower incorporation of **28** in the case of DNA **IV** (T template), equal selectivity (within standard deviation) was observed for **27** and **28** between the four templates (Graph 4.1), so these compounds are evidently unsuitable for sequence-specific recognition of DNA. However, bases which allow sequence-independent duplex formation can usefully serve as 'universal' bases, for example in PCR primers which span sections of genomic DNA of variable sequence (i.e. a

universal base could be integrated into the primer at a position opposite a site of ambiguous sequence, such as a SNP location).[125]

4.3 Targeted Synthesis and Screening of Nucleobase Analogues

A smaller set of four aldehydes was targeted for screening. Firstly, aldehyde-modified 2,6-diaminopurine ($\mathbf{D}_{CHO}$; Scheme 4.1a) was designed to allow an initial comparison with $\mathbf{A}_{CHO}$ (Chapters 2 and 3). This would provide a good indication of the suitability of templated dynamic incorporation as a means for the discovery of nucleobase analogues, given that D has already been reported to improve PNA/DNA duplex stability.[159] For further comparison, thiouracil aldehyde ($\mathbf{S}_{CHO}$; Scheme 4.1b) was targeted as an analogue of $\mathbf{T}_{CHO}$, since S-A is known to give melting temperatures similar to a T-A base pair.[164] Other aldehydes would be synthesized with a view to improving on the relatively poor selectivity observed for $\mathbf{T}_{CHO}$ incorporation. Thus, the *N*9-alkylated xanthine aldehyde was initially targeted for synthesis in light of the improved selectivity observed for purine *versus* pyrimidine incorporation (Chapter 2.4), although the *N*7-alkylated isomer ($\mathbf{X}_{CHO}$; Scheme 4.1c) was eventually isolated. Building on the observation that increased hydrogen bonding also gives improved selectivity (Chapter 2.4), pyridone aldehyde ($\mathbf{T^*}_{CHO}$; Scheme 4.1d) was prepared, as a recent publication calculated an increased hydrogen-bond strength for a T*-A base-pair relative to T-A.[165]



Scheme 4.1 Synthesis of a targeted set of aldehydes. (a) NH_3 in methanol, microwave, 65 → 90 °C, 150 min; (b) 1:1 v/v TFA:H_2O, microwave, 100 °C, 30 min; (c) *N,O*-bis(trimethylsilyl) acetamide, bromoacetaldehyde diethyl acetal, KI, DiPEA, RT → 100 °C (microwave); (d) bromoacetaldehyde diethyl acetal, Cs_2CO_3, DMF, microwave, 100 → 130 °C, 60 min, then RT, 40 h; (e) 10 M KOH aq, microwave, 205 °C, 25 min; (f) EtOH, cat. HCl, microwave, 85 °C, 30 min; (g) DIBAL-H, *n*-hexane/DCM, -78 °C → RT.

2,6-diaminopurine aldehyde ($\mathbf{D}_{CHO}$) was prepared *via* 2-amino-6-chloropurine acetal (**8**; Chapter 2.2) by nucleophilic aromatic substitution with methanolic ammonia in a microwave-assisted reaction to afford intermediate acetal **30**. Microwave-assisted hydrolysis of **30** in aqueous TFA then afforded $\mathbf{D}_{CHO}$ as the hydrate trifluoroacetate.

Although thiouracil could be successfully alkylated with bromoacetaldehyde diethyl acetal to afford **31**, attempted hydrolysis of **31** to the target aldehyde in aqueous acid was fruitless. Instead, several products were present (as judged by ^{1}H NMR), and ES-MS showed the presence of a compound with mass corresponding to the bicyclic derivative resulting from nucleophilic attack of sulfur on the liberated aldehyde. Thiouracil is able to tautomerize more readily than uracil owing to the greater polarizability of the larger sulfur atom as compared to the analogous oxygen in uracil. The target $\mathbf{S}_{CHO}$ was therefore dismissed as being inherently unstable to cyclization, and serves as an example to show that these 'simple' aldehyde-modified nucleobases may not always be synthetically accessible, owing to the reactivity of the aldehyde functionality with any nucleophilic groups present in the target molecule. An attempted incorporation on **P1** (Chapter 2.3, Table 2.1) templated by A (DNA **II**; Chapter 2.3, Table 2.2) using the crude hydrolysis product of **31** returned the starting PNA only, suggesting the absence of any reactive aldehydic functionality and the stability of any cyclic thioacetal to ring-opening.

Xanthine derivative $\mathbf{X}_{CHO}$ was synthesized *via* **32**, which was prepared by direct alkylation of xanthine with bromoacetaldehyde diethyl acetal. Although the *N*9 isomer was originally targeted for synthesis and analysis, only products resulting from alkylation at the *N*3 and *N*7 positions were obtained as a mixture in vanishingly low yield after column chromatography. These isomers were distinguished by 2D (HMBC) NMR spectroscopy. The *N*7 derivative was isolated and taken forward for testing in dynamic incorporation experiments. Thus, microwave-assisted hydrolysis of **32** in aqueous TFA afforded $\mathbf{X}_{CHO}$ as the hydrate trifluoroacetate.

In order to synthesize the pyridone derivative, commercially available 2-chloro-3-pyridylacetic acid was subjected to nucleophilic aromatic substitution in a high-pressure microwave-assisted reaction in concentrated aqueous potassium hydroxide. The resulting carboxylic acid was not isolated in this instance but converted directly to the corresponding ethyl ester **33** in a further microwave-assisted step, then reduced with DIBAL-H to afford a separable mixture of the alcohol and target aldehyde $\mathbf{T^*}_{CHO}$.

The three synthesized aldehydes were tested in a series of competition experiments with their analogous natural nucleobases. Thus, $\mathbf{D}_{CHO}$ and $\mathbf{A}_{CHO}$ were

mixed in equimolar ratio and the resulting solution used in a dynamic incorporation reaction templated by T (DNA **III**; Chapter 2.2, Table 2.2). As anticipated given the additional stabilization imparted by its extra hydrogen-bond, only D incorporation was observed in the resulting mass spectrum (Figure 4.5a; no A incorporation was discernible above the background noise).

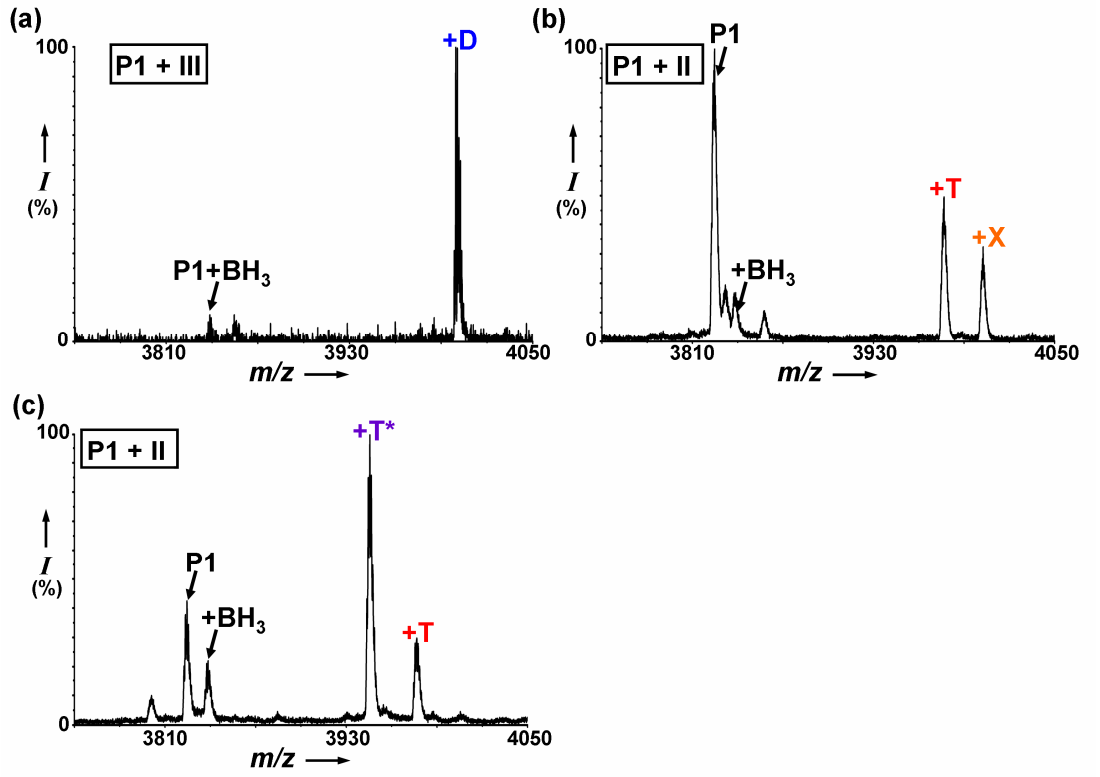


Figure 4.5 Mass spectra obtained by dynamic incorporation of the non-natural nucleobase aldehydes in competition with their natural analogues. (a) Incorporation of only D was observed in the presence of A, with T as the templating base; (b) X was incorporated to a lesser extent than T in a reaction templated by A; (d) T* was incorporated to a greater extent than T in a reaction templated by A. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.27 mM in each of the two competing aldehdyes; 5.0 μM in DNA template; 5.0 μM in **P1**; 100 mM in $NaBH_3CN$.

Templated incorporation of an equimolar mixture of $\mathbf{X}_{CHO}$ and $\mathbf{T}_{CHO}$ resulted in a mass spectrum showing a slightly larger peak height associated with T incorporation (Figure 4.5b), suggesting that duplex stability is greater for adenine paired with the natural base than with the *N*7-xanthine derivative. X:T peak ratios were 0.6:1, averaged over ten spectra (five for each of two duplicate experiments).

However, repeating the experiment with $\mathbf{T^*}_{CHO}$ in place of $\mathbf{X}_{CHO}$ resulted in a peak indicating greater incorporation of pyridone (T*) relative to thymine (Figure 4.5c). T*:T peak ratios were 3:1, again averaged over ten spectra (five for each of two duplicate experiments). Thus, T* was selected as a potential candidate nucleobase analogue that may impart greater duplex stability than T when paired with A.

Further dynamic chemical incorporation experiments using $\mathbf{T^*}_{CHO}$, $\mathbf{C}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$ and the four DNA templates **I-IV** (Figure 4.6) were performed to investigate the selectivity for T* against mis-incorporation.

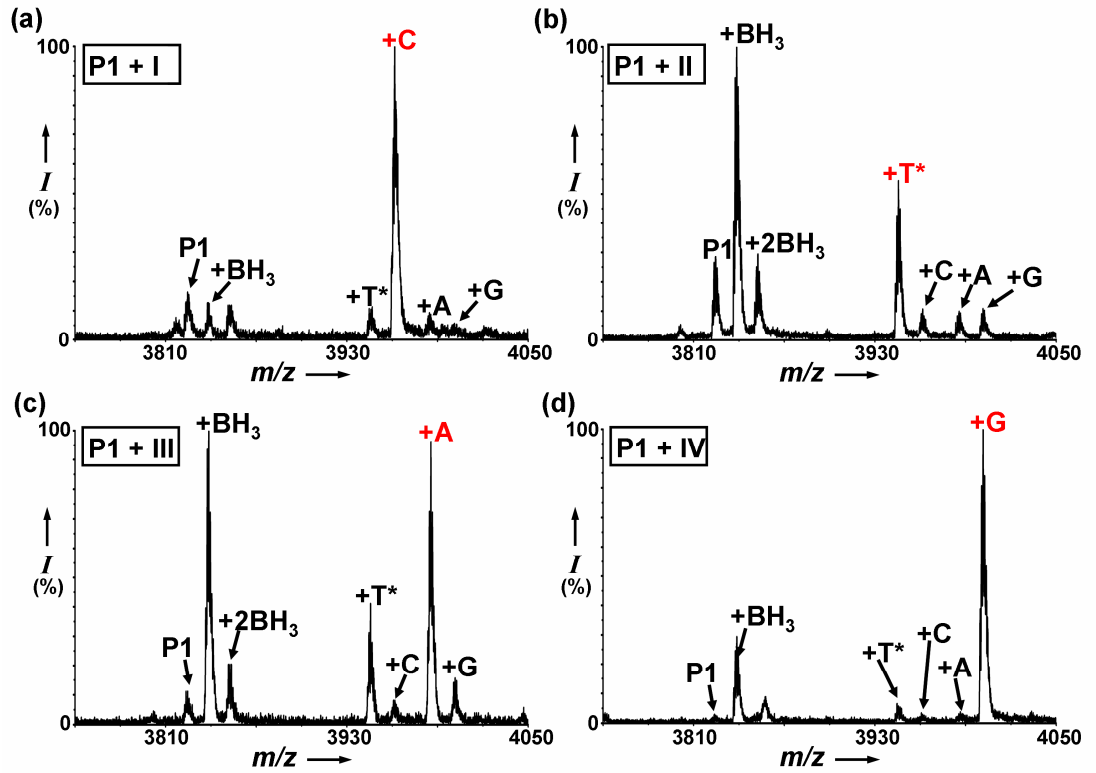
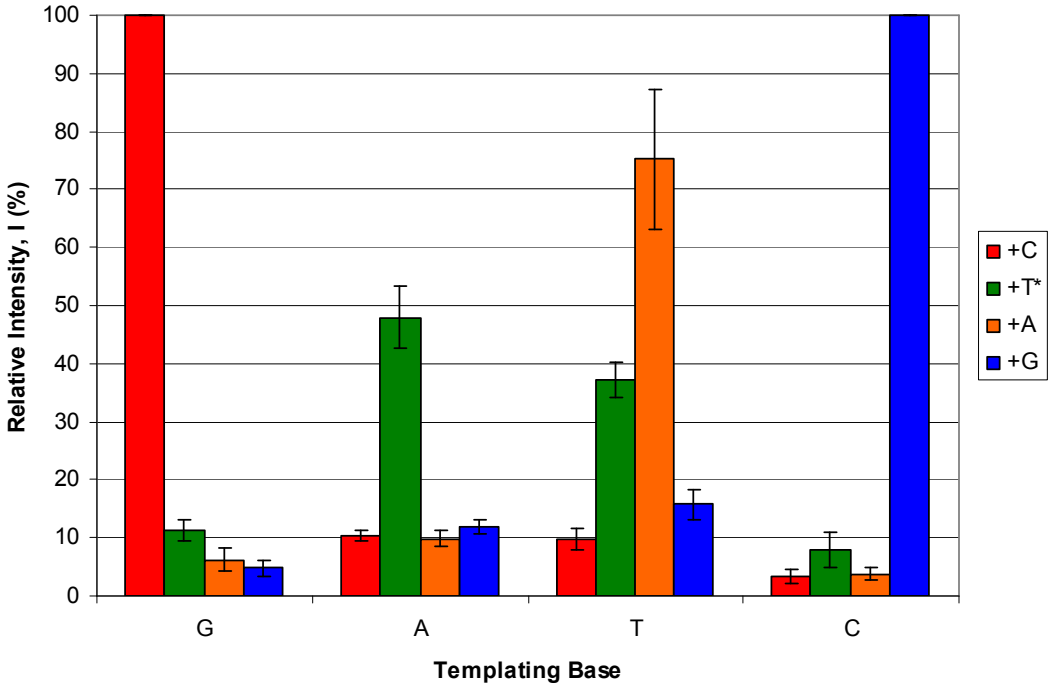


Figure 4.6 Representative mass spectra recorded after DNA-templated reductive aminations using an equimolar ratio of aldehydes $\mathbf{C}_{CHO}$, $\mathbf{T^*}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$. Peaks shown in red are due to products with the 'correct' base incorporated according to Watson-Crick base-pairing. (a) DNA template **I** directs incorporation of C; (b) **II** directs incorporation of T*; (c) **III** directs incorporation of A; (d) **IV** directs incorporation of G. Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate buffer; 0.14 mM in each of $\mathbf{C}_{CHO}$, $\mathbf{T^*}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$; 5 μM in DNA template and **P1**; 100 mM in $NaBH_3CN$.

Thus, in the presence of the A template, T* was incorporated more selectively than T in the presence of the other three natural nucleobase aldehydes (Graph 4.2 and Table 4.1), again suggesting that replacement of T-A with T*-A base-pairs would increase duplex melting temperatures. However, mis-templated incorporation of T* in the presence of the C and G templates was slightly higher than for T, and in the presence of the T template mis-templated incorporation of T* was much higher (the selectivity for A *versus* T was only 2:1). On the basis of these results, it was predicted that the duplex melting temperatures for T* would be slightly higher for single-base T*-C and T*-G mismatches, and substantially higher for a T*-T mismatch, than would be observed for natural T.



Graph 4.2 Mean peak intensities (of the most common isotope, relative to the most intense peak) resulting from the templated reaction of an equimolar mixture of $\mathbf{C}_{CHO}$, $\mathbf{T^*}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$ with **P1**. Error bars indicate the standard deviation across ten mass spectra (five from each duplicate analysis).

Table 4.1 MALDI signal ratios resulting from an equimolar mixture of $\mathbf{C}_{CHO}$, $\mathbf{T^*}_{CHO}$, $\mathbf{A}_{CHO}$ and $\mathbf{G}_{CHO}$, compared with the results for an equimolar mixture of the natural nucleobases.

DNA Oligomer	Templating Base	MALDI Signal Ratios[a]	
		C:T*:A:G	C:T:A:G[b]
I	G	**21**:2:1:1	**19**:1:1:1
II	A	1:**5**:1:1	1:**4**:1:1
III	T	1:4:**8**:2	1:1:**8**:1
IV	C	1:2:1:**30**	1:1:1:**39**

[a]Based upon the mean relative intensities of the most common isotope and reported to the nearest integer. The value for the nucleobase complementary to the position under interrogation on the DNA template is in bold. [b]See Chapter 2.4.

4.4 Comparison of Dynamic Incorporation Results and Duplex Melting Temperatures

To determine whether or not the results for dynamic chemical incorporation of the pyridone aldehyde $\mathbf{T^*}_{CHO}$ provided a reliable estimate of the duplex stability and selectivity conferred by this nucleobase mimic, PNA oligomers **P9-12** containing zero, one, two and three pyridone (T*) nucleobases respectively were synthesized, complementary to DNA oligomers **XII-XVI** which represented a fully matched and T, G and C single-base mismatched sequences respectively (Table 4.2). Incorporation of pyridone into oligomers **P9-12** was achieved using monomer **36**, which was synthesized as shown (Figure 4.7).

Table 4.2 PNA and DNA sequences used for T_m measurements.

Oligomer	Sequence[a]
P9	*N*-AGT GAT CTA C Lys-*C*
P10	*N*-AGT GA**T*** CTA C Lys-*C*
P11	*N*-AGT GA**T*** C **T***A C Lys-*C*
P12	*N*-AG**T*** GA**T*** C**T***A C Lys-*C*
XIII	*5'*-GTA G**A**T CAC T-*3*'
XIV	*5'*-GTA G**T**T CAC T-*3'*
XV	*5'*-GTA G**G**T CAC T-*3'*
XVI	*5'*-GTA G**C**T CAC T-*3'*

[a]All PNA oligomers possessed a *C*-terminal lysine residue.

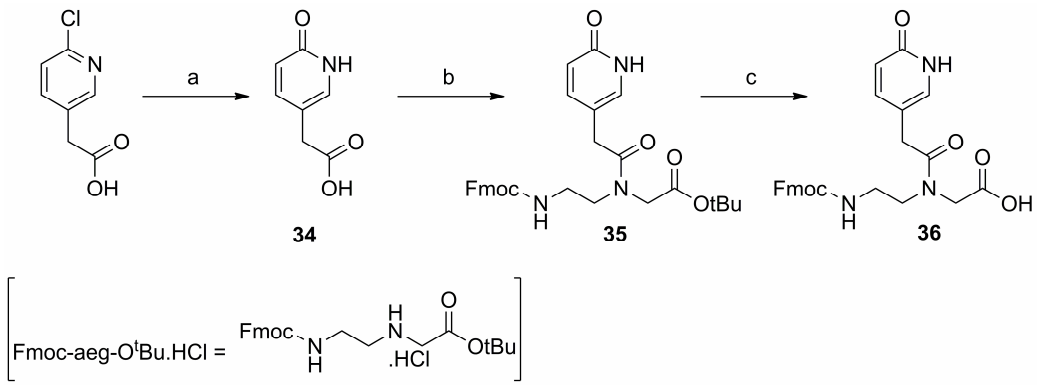


Figure 4.7 (a) 10 M KOH aq, microwave, 205 °C, 25 min; (b) Fmoc-aeg-O^{t}Bu.HCl, DCC, oxyma, DiPEA, DMF, microwave, 60 °C, 30 min; (c) 1:1 v/v TFA:DCM, 8 h.

The melting temperatures for PNA/DNA duplexes containing T* in the PNA strand (Table 4.3) did not agree with the predictions made on the basis of the dynamic incorporation experiments performed with **T***$_{CHO}$. Specifically, the replacement of T with T* at one position lowered the T_m by 13 °C. No substantial difference was observed between T_m values of duplexes containing one and two T*-A base-pairs, but a further decrease in T_m of 14 °C resulted from a third T*-A pair. However, comparison of the duplex T_m values for **P10** with DNA oligomers **XIV-XVI** showed poor mismatch discrimination for T* as expected, although with greater stability observed for the T*-G 'mismatched' base-pair than the predicted T*-T 'mismatch'. The stability recorded for the T*-G 'mismatch' was even higher than for the supposed T*-A match, and is indicative of the tautomerization of pyridone to allow the formation of a T*-G 'wobble' base-pair which is more stabilizing than T*-A.

Table 4.3 Duplex melting temperatures for PNA oligomers incorporating T*.

PNA Oligomer	DNA Oligomer	Base-Pair (PNA/DNA)	T_m (± 1)/ °C[a]	ΔT_m (°C)[b]
P9	**XIII**	T/A	49[c]	(0)
P10	**XIII**	T*/A	36	-13
P11	**XIII**	T*/A (× 2)	36	-13
P12	**XIII**	T*/A (× 3)	22	-27
P10	**XIV**	T*/T	36[d]	-13
P10	**XV**	T*/G	42	-7
P10	**XVI**	T*/C	33	-16

[a]Determined using CaryWin UV software, from the maximum of the first derivative of a plot of T *vs* A_{260nm}. Measurements were performed at 1 °C intervals. Uncertainty is based on the standard deviation across four measurements. [b]Change in T_m relative to fully matched canonical PNA/DNA duplex **P9/XIII**. [c]This is in precise agreement with a literature value.[159] [d]Uncertainty in this reading is ± 6 °C due to poorly defined melting curves.

4.5 Discussion and Conclusions

Dynamic chemical incorporation of aldehydes $\mathbf{D}_{CHO}$ and $\mathbf{A}_{CHO}$ in the presence of T template (DNA **III**) showed clear incorporation of the 2,6-diaminopurine (D) only. This is entirely consistent with the hypothesis that the additional hydrogen-bond of D provides extra stability to the iminium species generated from $\mathbf{D}_{CHO}$ and the secondary amine at the PNA blank position. This iminium species is therefore the more concentrated (relative to that formed from $\mathbf{A}_{CHO}$ and the PNA blank) at equilibrium, which is reflected in the product distribution after reduction and mass-spectrometric detection. This ties in well with reported studies on the effect of D on the T_m of PNA/DNA melting temperatures, wherein D was shown to increase duplex melting temperatures.[159]

Of the T analogues investigated, the aldehyde derived from xanthine, $\mathbf{X}_{CHO}$, was incorporated slightly less selectively than $\mathbf{T}_{CHO}$. This was predicted, as xanthine can be thought of as a size-expanded version of T, and as such will destabilize a duplex in which all of the other bases are natural (i.e. not size-expanded). In addition, alkylation at *N*7 may present a sub-optimal hydrogen-bonding motif for recognition by an adenine template. The pyridone aldehyde, $\mathbf{T^*}_{CHO}$, on the other hand, was

incorporated slightly more selectively than $\mathbf{T}_{CHO}$. This was also expected, given the additional hydrogen-bond strength calculated for this base analogue.

However, for T*, the dynamic chemical incorporation results were not predictive of the T_m measurements made for PNA/DNA duplexes in which the PNA oligomer contained T*. While this work was in progress, Benner and co-workers reported the use of 3-methyl-2-pyridone as a potential nucleobase mimic, but it was found to cause duplex destabilization and poor mismatch discrimination in DNA/DNA duplexes.[147] This has been attributed to the absence of a minor-groove hydrogen-bond donor in the pyrimidine ring (which is present in thymine), resulting in duplex stabilization, as this is necessary for hydrogen-bonding with water in the minor grove of the duplex (Figure 4.8).[166]

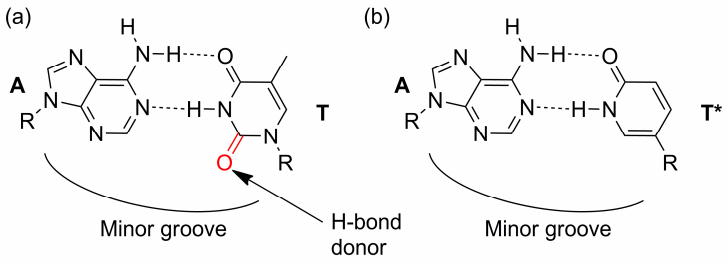


Figure 4.8 (a) Hydrogen-bond donors in the minor groove stabilize a B-form DNA double-helix through hydration. (b) Pyridone lacks a hydrogen-bond donor in this position.

Based upon the discrepancy between the greater dynamic chemical incorporation for T* *versus* T and the duplex destabilization caused by the presence of T*-A in place of T-A base-pairs, it can be concluded that the method of dynamic chemical incorporation reported herein does not always provide a true estimate of duplex stability, and therefore is not a reliable means of screening for nucleobase analogues with improved properties for diagnostic and antisense applications. The reason for this discrepancy has yet to be fully investigated and determined, but potential reasons may be offered. Firstly, the dynamic chemistry employed selects for the most stable iminium species at equilibrium. In some instances (e.g. for the templated incorporation of the natural nucleobases or D), this may correspond to the relative stabilities of PNA/DNA duplexes in which the nucleobase is connected to the PNA backbone by the canonical amide linkage, but this may not always be the case. Secondly, although it has been supposed that a single base change in the PNA

sequence does not substantially affect the ionizability of the PNA oligomer and hence mass spectrometric signal strengths, this cannot be ruled out because a calibration curve to determine peak heights at, for example, different relative concentrations of PNA oligomers **P9** and **P10** (to measure the effect of T*/T substitution on peak heights) was not performed.

In spite of the failure of dynamic chemistry to reliably predict T_m values by the method described, it has been demonstrated that novel aldehydes can be screened to identify those which are more selectively incorporated, as exemplified by the experiment using $\mathbf{D}_{CHO}$. Such aldehydes may serve to improve the method of genetic analysis developed previously, by reducing the possibility of mis-primed incorporation. $\mathbf{D}_{CHO}$ could thus replace $\mathbf{A}_{CHO}$ for SNP and indel genotyping by dynamic chemistry (see Chapter 3), although this may require derivatization with a mass tag to prevent overlap with peaks associated with G incorporation (since D has a mass only 1 Da less than that of G).

In a recent development, Leumann and co-workers reported a very similar approach to that described herein for the rapid screening of a parallel library of aromatic heterocyclic amines for the discovery of novel artificial nucleobases.[167] In this study, reversible hemiaminal formation was employed at an abasic site on a fluorophore-labelled dodecameric DNA oligomer. The nucleobase candidates were screened in parallel for sequence selective incorporation opposite a templating DNA strand which was labelled with a fluorescence quencher to facilitate direct T_m determination by fluorescence measurements. One drawback of this hemiaminal approach, however, was that both anomers may have been produced, and so carbocyclic derivatives had to be synthesized representing the α and β derivatives to establish which anomer was responsible for the observed T_m values.

CHAPTER 5
Aldehydes for DNA-Templated Extension of PNA Oligomers

5.1 Introduction

Several examples of non-enzymatic, nucleic acid templated extension of informational polymers have been reported. One strategy involves the irreversible chemical ligation of activated building blocks, notable examples of which have been described by Szostak and co-workers in studies focused on the development of an alternative genetic system. The reaction between an amine and an imidazole-activated phosphate group facilitated the template-directed synthesis of both N2'→P3' glycerol nucleic acid (npGNA) and N2'→P5' DNA.[168, 169] Template-directed ligation of DNA has also been achieved by Kool and co-workers, who employed an S_N2 reaction of a phosphorothioate with an alkyl iodide to ligate oligomers for DNA and RNA sequence analysis.[170] More recently, Brown and colleagues used the a Diels-Alder reaction to perform the simultaneous DNA-templated ligation of three oligonucleotides.[171] Examples of templated oligomer ligation chemistries applied to mutation detection include the native chemical ligation of PNA reported by Seitz and colleagues,[172] and the reaction between an amine and an azaoxybenzotriazolide activated phosphate reported by Richert and co-workers in the ligation of both DNA and RNA.[173, 174]

As an alternative to these irreversible chemical ligation strategies, reversible reactions have been employed in a number of dynamic chemical approaches to the template-directed extension of oligomers. Such reactions permit the reversal of any mis-incorporation and subsequent insertion of the 'correct' building block to improve fidelity in the daughter strand. This is more akin to biological systems, in which mis-incorporated nucleotides can be removed by the hydrolytic domains of polymerase enzymes. For example, reversible imine formation for DNA template-directed ligation has been reported by Lynn and colleagues, who joined modified DNA trimers and also polymerized amino-aldehyde DNA dimers by reductive amination.[175-177] In a strategy similar to that described herein, Liu and colleagues

have performed the DNA-templated polymerization of PNA aldehyde tetramers and pentamers by reductive amination.[178-180] Sequence-selective single-base extension has also been reported by Micklefield and colleagues, using DNA to template a reductive animation between a hybridized strand of PNA and di-aldehyde morpholino precursors to generate a morpholino-PNA chimera for analysis by MALDI-TOF mass spectrometry.[181]

To facilitate the method of genotyping and sequencing by terminal extension outlined in Chapter 1.2.3, a set of four *N*-terminal protected PNA aldehydes was required (Figure 5.1). The aim of the work presented in this Chapter was to synthesize such aldehydes and test them for DNA-templated dynamic extension of a PNA probe. Early work targeted a set of unlabelled monomers for direct read-out of the extended PNA products by MALDI-TOF MS analysis as described in previous Chapters.

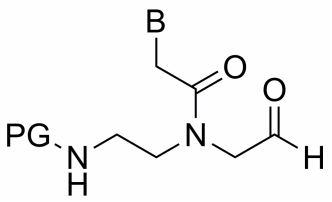
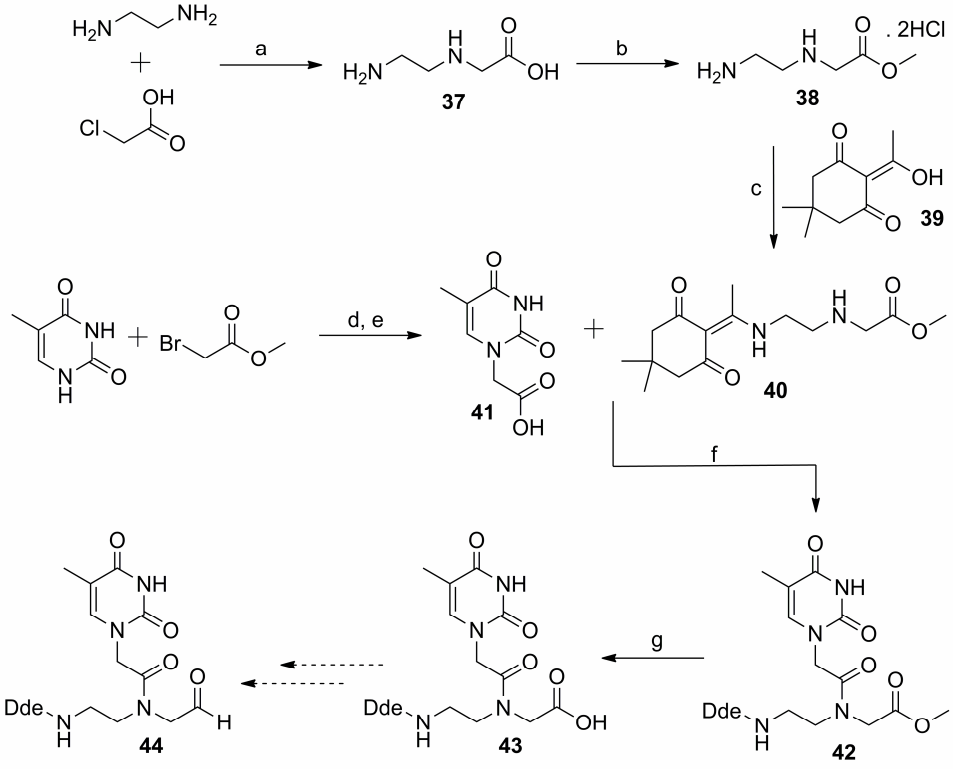


Figure 5.1 General structure of targeted PNA aldehydes for terminal extension. B = nucleobase (A, G, C or T); PG = protecting group.

5.2 Synthesis of a Thymine PNA Aldehyde

The initial phase of this work concerned the synthesis of a Dde-protected PNA-aldehyde thymine monomer (**44**, Scheme 5.1) from the analogous carboxylic acid (**43**). This route was chosen because **43** was readily prepared *via* a well-established synthetic route.[83] The Dde (1-(4,4-dimethyl-2,6-dioxacyclohexylidene)ethyl) protecting group is classically removed with hydrazine, and is orthogonal to the acid labile groups (Bhoc and Mmt) that are routinely used for the protection of the exocyclic amine functionality of adenine, guanine and cytosine.[182] A compound very similar to the target aldehyde **44** but bearing a Boc protecting group in place of Dde had been reported by Liu.[178] However, the preparation of this poorly characterized PNA-aldehyde thymine monomer involved a low-yielding synthesis culminating in

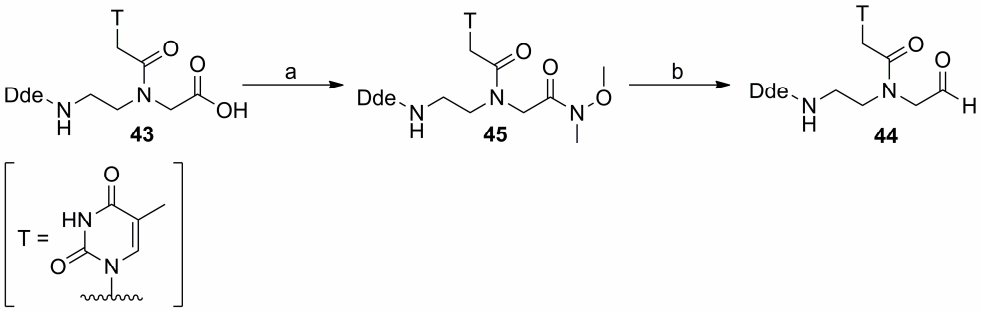
an OsO_4-catalyzed cleavage of an allylic intermediate, and an alternative route to **44** was therefore sought.



Scheme 5.1 Synthesis of carboxylic acid **43** as a precursor to thymine aldehyde **44**: (a) 10 °C → RT, 16 h; (b) $SOCl_2$, MeOH, 0 °C → reflux, 20 h; (c) DiPEA, DCM, 16 h; (d) K_2CO_3, DMF, 16 h; (e) NaOH aq, reflux, 3h; (f) DCC, HOBt.H_2O, DMF, 16 h; (g) Cs_2CO_3, 1:1 v/v MeOH:H_2O, 4.5 h.

The synthesis of peptide aldehydes has been widely investigated as they have many applications. For example, they are potent inhibitors of serine and cysteine proteases, as they react with these enzymes to yield, respectively, hemiacetals and hemithioacetals which resemble the transition state of amide hydrolysis and are bound tightly in the active site.[183] One of the most widely reported methods of synthesising *N*-protected peptide aldehydes involves the reduction of Weinreb amides, often with lithium aluminium hydride or DIBAL-H.[184, 185] Thus **43** was converted into Weinreb amide **45** (Scheme 5.2), then reduced to the target aldehyde. To prevent over-reduction of **44**, the milder reducing agent lithium tri-*tert*-

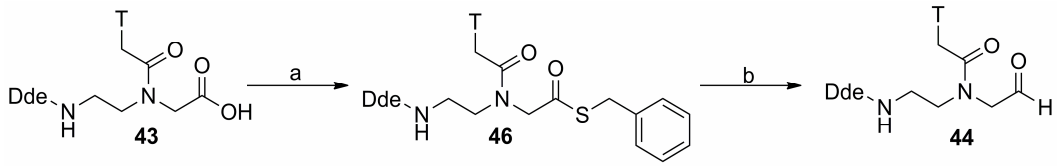
butoxyaluminium hydride ($LiAlH(O\text{-}t\text{-}Bu)_3$) was employed in place of the more commonly used $LiAlH_4$.[186]



Scheme 5.2 Synthesis of target aldehyde *via* reduction of a Weinreb amide: (a) MeONHMe.HCl, EDC.HCl, $HOBt.H_2O$, Et_3N, DMF, 20 h; (b) $LiAlH(O\text{-}t\text{-}Bu)_3$, THF, 95 min.

However, reduction of **45** and isolation of the resulting aldehyde is not facile. The use of three equivalents of $LiAlH(O\text{-}t\text{-}Bu)_3$ affords **44** with approximately 63 % conversion as judged by HPLC. The use of fewer equivalents of the hydride gives a lower conversion of the amide, whilst greater equivalents result in over-reduction to the corresponding alcohol and unidentified impurities. Attempts to purify the aldehyde by column chromatography on silica gel were unsuccessful, and resulted in decomposition of the target. A small, analytically pure sample of **44** was finally isolated for partial characterization after purification by preparative HPLC, although this compound degraded over time in solution (*vide infra*).

As an alternative route to aldehyde **44**, the *S*-benzyl thioester **46** was prepared and reduced (Scheme 5.3).



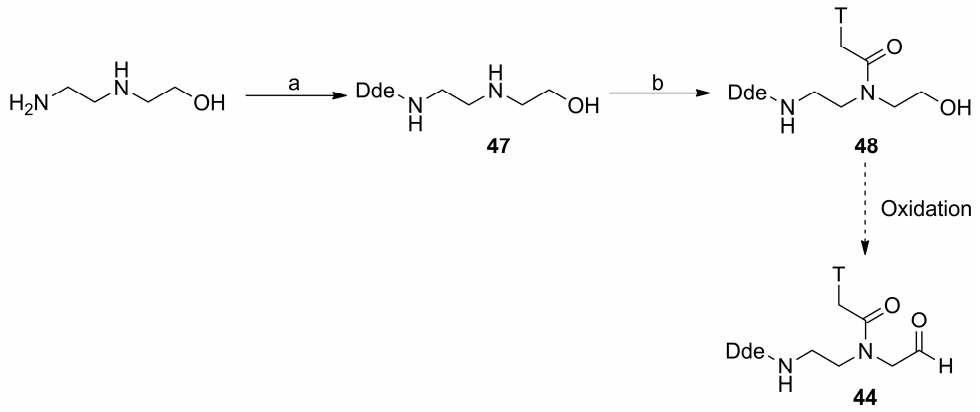
Scheme 5.3 Synthesis of target aldehyde *via* reduction of an *S*-benzyl thioester: (a) benzyl mercaptan, DMAP, DCC, DMF, 22 h; (b) triethylsilane, Pd/C, THF, 2 h.

HPLC of the reaction mixture for reduction of **45** suggested approximately 83 % conversion of the thioester to the aldehyde **44**. However, it was not possible to drive the reaction to completion through the addition of extra catalyst or triethylsilane.

Furthermore, work-up of the reaction mixture was less straightforward than for reduction of the Weinreb amide as the crude product was contaminated with catalyst and unreacted triethylsilane.

A 'cleaner' catalytic hydrogenation reaction of **46** was attempted with 10 % Pd/C using the HC-Tutor™. This is an educational version of the H-Cube™ developed by ThalesNano, which is a bench-top continuous-flow reactor capable of heterogeneous catalytic reductions.[187] The instrument generates hydrogen by the electrolysis of water, and mixes it with a solution of the starting material before it is pumped through a heated column containing a solid-supported catalyst. Unfortunately, the HC-Tutor™ generated only a trace of **44** even after repeated flows through the catalyst column. This may be because the catalyst is poisoned by the sulfur-containing starting material or by-products.

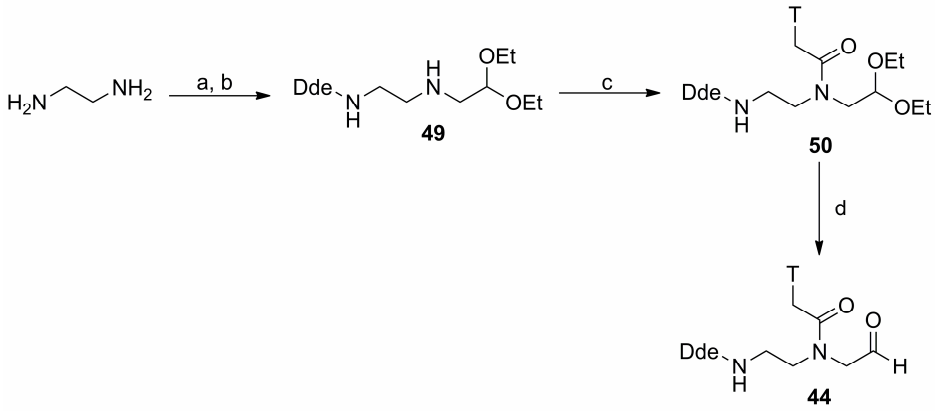
The difficulties (moderate yields and problematic purifications) experienced in the conversion of carboxylic acid **43** to **44** stimulated the investigation of faster and cleaner routes to the aldehyde. It was envisaged that methyl ester **42**, which is a precursor in the synthesis of the acid (Scheme 5.1) could be reduced to the corresponding alcohol. Test reactions with $LiAlH(O\text{-}t\text{-}Bu)_3$ showed this to be possible. The alcohol might then be oxidized to aldehyde **44**, thereby removing two steps from the overall synthesis. However, a faster route to the alcohol (**48**, Scheme 5.4) was devised, which would yield **44** in only three steps from commercially available 2-(2-aminoethylamino)ethanol.



Scheme 5.4 Synthesis of target aldehyde *via* oxidation of a primary alcohol: (a) Dde-OH (**39**), DCM, 16 h; (b) **41**, PyBOP, DiPEA, DCM, 0 °C → RT, 2 h.

Thus **48** was rapidly prepared by a coupling of acid **41** with the Dde protected intermediate **47**. The low yield of this reaction was attributable to the formation of the ester by coupling of the terminal hydroxyl group with **41**. Trial oxidations of **48** using Dess-Martin periodinane[188] and sulphur trioxide-pyridine complex[189] afforded crude aldehyde (judged by TLC and ^{1}H NMR of the crude material) although it was not isolated or purified. With some reaction optimization this synthesis of aldehyde **44** offers a much faster and more economical route than that proceeding *via* carboxylic acid **43**. However, focus was shifted from this synthetic route by the development of an alternative methodology (*vide infra*).

A more straightforward route to aldehyde **44** was envisaged in which the aldehyde functionality was incorporated through a diethyl acetal. Thus, acetal **50** prepared in a microwave assisted synthesis starting from 1,2-diaminoethane and bromoacetaldehyde diethyl acetal (Scheme 5.5).



Scheme 5.5 Synthesis of target aldehyde *via* acetal hydrolysis: (a) bromoacetaldehyde diethyl acetal, K_2CO_3, MeCN, microwave at 130 ^oC, 30 min; (b) DdeOH (**39**), DCM, 16 h; (c) **41**, DCC, HOBt.H_2O, DMF, microwave at 60 ^oC, 30 min; (d) 90 % TFA aq, -10 → 0 ^oC, 1 h.

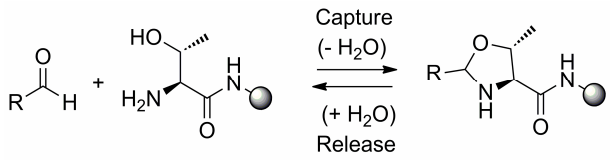
Several attempts were made to hydrolyze this acetal to the target aldehyde: refluxing in aqueous HCl; pTsOH in acetone/water;[190] iodine in acetone/DCM.[191] However, the best result was obtained by stirring with aqueous trifluoroacetic acid at between 0 and -10 ^oC.[192] This gave the target aldehyde in approximately 90 % yield as judged by HPLC. The synthesis and deprotection of acetal **50** appears to offer the fastest and most economical route to aldehyde **44**.

Mention should be made of the instability of the target aldehyde **44**. As noted above, the final target is difficult to purify, and tends to degrade rapidly in solution

(aqueous or organic) to a mixture of products. Although a thorough investigation of the degradation products was not undertaken, it may be speculated that one possible cause is an intramolecular cyclization *via* a favourable 6-*exo*-trig ring-closure. Intramolecular cyclization in this way has been reported previously, wherein cyclic enamides were synthesized *via* intramolcular attack of the amide nitrogen of *N*-carboxybenzyl (Cbz) protected amines on *N*α-acetals in aqueous TFA.[193-196] Furthermore, the relatively mild conditions (aqueous acid at between -10 and 0 ^{o}C) used to hydrolyze acetal **50** are in stark contrast to those needed in the hydrolysis of the nucleobase acetals described in previous chapters (e.g. microwave heating in aqueous acid at 100 ^{o}C, used in the hydrolysis of diethyl acetals **4** and **7** to yield aldehydes $\mathbf{C}_{CHO}$ and $\mathbf{A}_{CHO}$ respectively; see Chapter 2, Scheme 2.1), which is consistent with neighbouring-group participation in the hydrolysis of **50**.

5.3 Resin Capture of a Thymine PNA Aldehyde

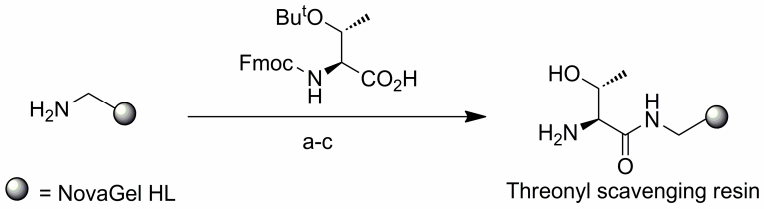
It was anticipated that aldehyde **44** could be resin-captured using a threonyl scavenging resin, as reported by Liu (Scheme 5.6).[178, 180] This would allow purification of **44**, and also enable Dde deprotection on the resin for alternative derivatization of the *N*-terminus of the aldehyde (e.g. by attachment of fluorophores, with or without cleavable linkers for sequencing by cyclic reversible termination). Cleavage from the resin would then afford the pure PNA aldehyde monomer for templated terminal extension of a PNA strand.



Scheme 5.6 Condensation of an impure aldehyde with a threonine modified resin yields a supported oxazolidine. After washing, cleavage affords the pure aldehyde.

Thus, a threonyl scavenging resin was prepared from a commercially available aminomethyl functionalized PEG-PS (i.e. polyethylene glycol/ polystyrene composite) resin, namely NovaGel HL, by coupling with Fmoc-Thr(*t*-Bu)-OH and removal of the Fmoc and *t*-Bu protecting groups (Scheme 5.7) according to a literature method.[188] A model study was then performed in which the scavenger was

used to purify 3-phenylpropionaldehyde from an equimolar mixture with 3-phenylpropan-1-ol by capture/release in the microwave (Figure 5.2).



Scheme 5.7 Preparation of an aldehyde-scavenging resin: (a) TBTU, DiPEA, DMF; (b) 20 % v/v piperidine/DMF; (c) 80 % v/v TFA/DCM.

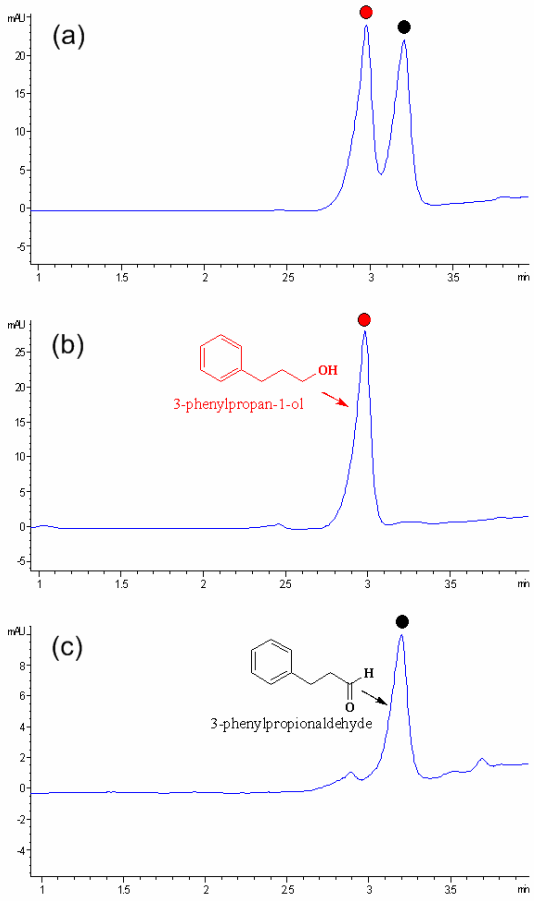


Figure 5.2 HPLC traces of: (a) an equimolar mixture of 3-phenylpropan-1-ol and 3-phenylpropionaldehyde in MeOH + DiPEA; (b) the supernatant after resin capture in the microwave at 60 ^{o}C for 60 min; (c) the cleavage cocktail obtained after heating in the microwave at 60 ^{o}C for 30 min with 60:40:1 MeCN:H_2O:TFA.

Initial studies towards capturing **44** using the threonyl scavenging resin were met with some success. Following capture, washing and release of the crude aldehyde obtained by reduction of Weinreb amide **45**, HPLC analysis of the cleavage cocktail

showed the presence of aldehyde **44** and a slight reduction in the amount of **45** present (Figure 5.3).

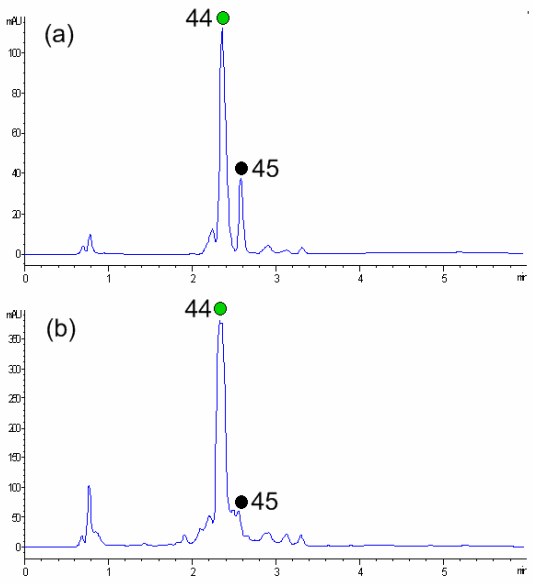


Figure 5.3 Capture/release of **44** using a threonyl scavenging resin. HPLC traces of: (a) crude product; (b) cleavage cocktail. Capture performed from 86:9:6:1 MeOH:DCM:DMF:AcOH at room temperature, 1 h. Release with 80:10:5:5 MeOH:AcOH:DCM:H_2O at room temperature, 80 min. Milder conditions than for the model system were employed during capture/release owing to the poorer stability of **44**.

5.4 Templated Terminal Extension of a PNA Oligomer

A DNA-templated extension was planned using PNA oligomer **P13**, which was synthesized with a free *N*-terminus (Table 5.1). DNA oligomer **XVII** was purchased, with A as the templating base for terminal extension.

Table 5.1 PNA and DNA sequences used for terminal extension.

Oligomer	Sequence[a]
P13	*N*-CAT TCT TCC TCT-*C*
XVII	*5'*-AGA GGA AGA ATG AAA CAT AGA C-*3'*

[a]**P13** has a free *N*-terminus. The section of DNA **XVII** to which **P13** is complementary is underlined; as the *N*-terminus of PNA aligns with the 3'-end of DNA, **XVII** presents A as the templating base for extension of **P13** by reductive amination.

Owing to the problems associated with the stability of aldehyde **44**, it was envisaged that the easiest way to obtain DNA-templated extension of a PNA oligomer would be to hydrolyze the PNA acetal immediately prior to the extension reaction. Thus acetal **50** was treated with aqueous TFA and the resulting aldehyde was dissolved in water and used directly in extension reactions of **P13** templated by DNA **XVII** using the protocol described previously (see Chapter 2.4). A control reaction was performed in the absence of DNA template. The resulting mass spectra (Figure 5.4) showed incorporation of the thymine aldehyde **44** in the presence of DNA only, demonstrating the templated extension of a PNA oligomer by reductive amination. A significant signal for the extended product was only observed when the freshly prepared aldehyde solution was added after the initial PNA/DNA hybridization at 80 ^{o}C. If the aldehyde was present from the start, only very weak signals were observed for extension, which provides further evidence for the instability of **44** in solution.

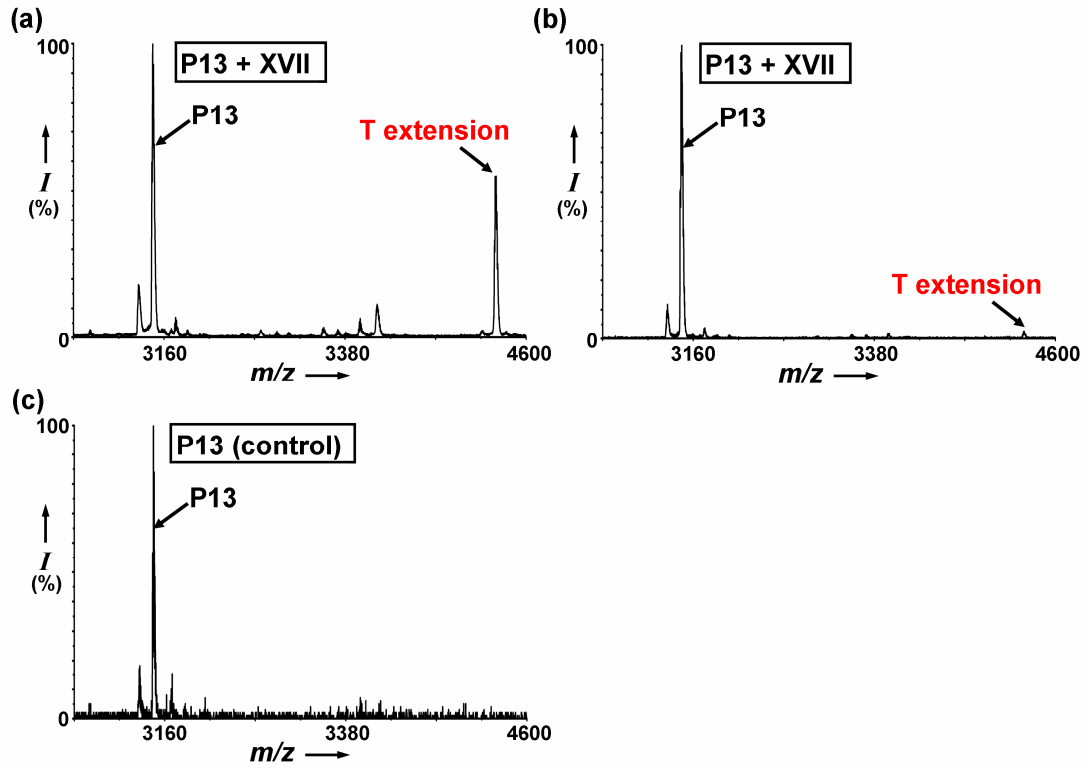


Figure 5.4 Extension of **P13** by reductive amination: (a) in the presence of DNA template **XVII**, with the aldehyde added at 40 ^{o}C immediately before reduction; (b) in the presence of DNA **XVII**, with the aldehyde added at the start prior to hybridization at 80 ^{o}C; (c) control reaction in the absence of DNA **XVII**, with the aldehyde added after the initial hybridization (as for (a)). Final concentrations in a 20 μL reaction volume: 4 mM pH 6.0 sodium phosphate

buffer; 3.2 mM in crude, freshly prepared **44**; 5 μM in DNA template (for (a) and (b) only) and **P13**; 100 mM in $NaBH_3CN$.

5.5 Discussion and Conclusions

A PNA aldehyde thymine monomer **44** was successfully synthesized by a number of routes, but the fastest involved the microwave assisted synthesis and mild deprotection of a diethyl acetal derivative, **50**. The aldehyde was used in a DNA-templated extension of a PNA strand by reductive amination, although sequence-selective templated extension has yet to be demonstrated by this approach as the cytosine, guanine and adenine analogues remain to be synthesized. However, the chemistry employed in the synthesis of the thymine PNA aldehyde monomer is readily applicable to the preparation of these analogues, as the exocyclic amines of these bases are routinely protected with the acid-labile Mmt or Bhoc groups, which could be removed in the same step as the deprotection of the acetal to the target aldehyde. Difficulties encountered in the isolation of **44** stimulated investigation of purifying the aldehyde by capture with a threonine modified resin, and a trial reaction met with some success. This could allow the Dde deprotection of resin-bound aldehyde for attachment of alternative protecting groups or fluorophores.

Problems during the synthesis of aldehyde **44** related to its instability in solution. Although the mechanism of degradation of **44** remains to be investigated, it is noteworthy that similar issues have been encountered by others when preparing monomers for the terminal extension of informational polymers. For example, Szostak and co-workers resorted to the use of dimers during the templated extension of N2'→P3' GNA, as the monomers were prone to cyclization.[168] It would be interesting to learn whether PNA aldehyde dimers would be less refractory and more useful for sequencing studies. Such dimers could be prepared after resin capture of those aldehydes reported herein and extension by solid-phase synthesis, as described by Liu and colleagues during the synthesis of PNA aldehyde tetramers.[178] It would also be interesting to test aromatic intercalating molecules (or 'molecular midwives') of the type described by Hud and co-workers in these extension reactions, as such molecules have been shown to hinder the cyclization of oligonucleotides and promote strand extension reactions.[197-199]

CHAPTER 6
Experimental

6.1 General Information

The **commercially available reagents** were used without further purification. All **DNA and RNA oligomers** were purchased in desalted form from Microsynth AG, Switzerland. **Dry solvents** were obtained using a Pure Solv Solvent Purification System (Innovate Technology, USA) or from commercial sources. **^{1}H NMR** spectra were recorded on Bruker AVA600, DMX500, AVA500, DPX360, or ARX250 spectrometers and **^{13}C NMR** on Bruker AVA600, DMX500, AVA500 or DPX360 spectrometers in the solvents indicated at 298 K (300 K for spectra recorded on the Bruker ARX250). Chemical shifts are reported on the δ scale in parts per million and are referenced *via* residual non-deuterated solvent resonances.[200] ^{13}C chemical shifts in D_2O or 10 % D_2O:H_2O were referenced *via* internal 1,4-dioxane (67.19 ppm) unless otherwise stated. **Low resolution electrospray (ES) mass spectra** were recorded on an Agilent Technologies LC/MSD 1100 Quadrupole Mass Spectrometer (QMS) with an electrospray ion source. **Electron impact (EI) ionization and high resolution electrospray mass spectra** were recorded by the MS Section of the University of Edinburgh on a Finnigan MAT 900 XLP high resolution, double-focussing mass spectrometer. **Microwave reactions** were performed using a Biotage Initiator instrument. **MALDI-TOF mass spectra** were recorded on an Applied Biosystems Voyager-DE™ STR instrument in positive ionization reflector mode (delay 175 ns, 20 kV accelerating voltage, variable laser intensity, typically 200 shots or fewer). Sinapic acid matrix consisted of 10 mg/mL sinapic acid in 50 % v/v acetonitrile in water with 0.1 % v/v TFA. **Melting points** were determined using a Gallenkamp melting point apparatus. **IR spectra** were recorded neat using a Bruker Tensor 27 FT-IR spectrometer. **TLC** was performed on aluminium-backed silica gel 60 plates using the eluents described. **Aldehyde and acetal products were visualized on TLC plates** using a 2,4-dinitrophenylhydrazine dip, which was prepared by dissolving 12 g of 2,4-dinitrophenylhydrazine in a mixture of conc. H_2SO_4 (60 mL), H_2O (80 mL) and 95 % EtOH (200 mL). **Flash column**

chromatography was carried out on silica gel 60 (40-63 μm particle size). Unless stated otherwise, compounds for purification were loaded directly onto the column as a solution in a small volume of eluting solvent. Alternatively (when stated), compounds were loaded onto the column by dissolving in a greater volume of solvent and 'pre-adsorbing' (with evaporation of the solvent) onto a small amount of silica which was then layered onto the silica column. **Automated flash column chromatography** was performed on a Biotage Isolera One instrument using pre-packed columns of the size specified. Prior to use, **Q Sepharose® Fast Flow** (GE Healthcare; 1 mL of a suspension in ethanol) was centrifuged and the supernatant removed. The resin was subsequently washed centrifugally with H_2O (1 mL) and 10 mM phosphate buffer, pH 7 (2 × 1 mL), before resuspending in the same buffer (0.5 mL). Immediately before use, the pre-equilibrated Q Sepharose® was agitated to resuspend the resin beads.

HPLC analyses were performed on an Agilent 1100 analytical system with a Supelco Discovery® C18, 5 μm, 50 × 4.6 mm column unless otherwise stated. Detection was by UV absorbance at 254 nm. The following eluents were used: (A) H_2O + 0.1 % TFA; (B) MeCN + 0.04 % TFA; (C) MeOH + 0.1 % FA; (D) H_2O + 0.1 % FA; (E) MeCN + 0.1 % FA; (F) MeOH; (G) H_2O; (H) MeCN + 0.1 % TFA. HPLC grade eluents were employed, at a flow rate of 1 mL/min with samples prepared to a concentration of approximately 1 mg/mL and filtered prior to injection. The following HPLC methods were used:

Method 1 (A and B): 10 % to 90 % B over 3 min, then 90 % B over 1 min.

Method 2 (C and D): 5 % to 95 % C over 6 min, then 95 % C over 3 min.

Method 3 (C and D): 5 % to 95 % C over 3 min, then 95 % C over 1 min.

Method 4 (F and G): 5 % to 95 % F over 3 min, then 95 % F over 1 min.

Method 5 (D and E): 5 % to 95 % E over 3 min, then 95 % E over 1 min.

Method 6 (A and H): 5 % to 95 % H over 10 min, then 95 % H over 4 min.

Method 7 (A and H; Phenomenex Jupiter® Proteo 4 μm, 90 Å, 250 × 4.6 mm column): 0 % to 18 % H over 13 min.

Preparative HPLC was performed on an Agilent 1100 preparative system. HPLC grade eluents were employed, at a flow rate of 3 mL/min with samples prepared to a concentration of approximately 10-20 mg/mL and filtered prior to

injection of a volume containing up to 10 mg. The following eluents were used: (A) H_2O + 0.1 % TFA; (B) MeCN + 0.1 % TFA. The following methods were used:

Method 1 (A and B, Phenomenex® Prodigy ODS (3), 5 μm, 100Å, 250 × 10 mm column): 30 % B over 1 min, then 30 to 90 % B over 20 min, then 90 % B over 5 min.

Method 2 (A and B, Hichrom C18, 5 μm, 100Å, 250 × 21.2 mm column): 5 % B over 1 min, then 5 to 50 % B over 20 min, then 50 % B over 14 min.

Method 3 (A and B, Hichrom C18, 5 μm, 100Å, 250 × 21.2 mm column): 0 % B over 1 min, then 0 to 65 % B over 30 min, then 65 to 100 % B over 4 min.

6.2 General Solid-Phase Synthesis (SPS) Procedures and Information

The PNA oligomers reported herein were prepared by standard solid-phase synthesis techniques on polymer supports by repeated rounds of coupling of activated (amino-protected) PNA monomers followed by deprotection of the terminal amino group, with washing steps after each stage.[201] Unless stated otherwise, all reactions were performed at room temperature in Supelco Solid-Phase Extraction (SPE) tubes fitted with frits and Teflon taps (Sigma-Aldrich, UK) which allowed washing steps to be carried out using a vacuum manifold. Resin agitation during reactions was by gentle rotation of the capped SPE tubes on a blood tube rotator (Stuart Scientific, UK).

6.2.1 Calculation of Theoretical Loading

The theoretical loading of a resin after a reaction can be calculated using the following equation:

$$N_L = S_L / [1 + (S_L \times \Delta M \times 10^{-3})]$$

Where: N_L is the 'new' theoretical loading (mmol/g);
S_L is the original, 'starting' loading (mmol/g);
ΔM is the change in formula weight (g/mol).

6.2.2 Qualitative Ninhydrin Test

A qualitative ninhydrin test was used to monitor amine deprotection and coupling steps. To a few resin beads was added three drops of reagent A and one drop of reagent B (see below), and the resulting mixture was heated at 100 ^{o}C for 10 min. A blue colour indicated the presence of free amine, whilst a yellow colour showed no free amine.

Reagent A:

Solution 1: Phenol (40 g) was dissolved in ethanol (10 mL) with gentle warming and stirred over Amberlite mixed-bed MB-3 resin (4 g) for 45 min. The resulting solution was obtained by filtration.

Solution 2: Potassium cyanide (65 mg) was dissolved in water (100 mL) and an aliquot (2 mL) was diluted with pyridine (38 mL, freshly distilled from ninhydrin) and stirred over Amberlite mixed-bed MB-3 resin (4 g). The solution was collected by filtration and added to 'solution 1' to give 'reagent A'.

Reagent B:

Ninhydrin (2.5 g) was dissolved in ethanol (50 mL) to give 'reagent B'.

6.2.3 Quantitative Fmoc Test[201]

A quantitative Fmoc test was used to determine the new loading of (aminomethyl)polystyrene resin after functionalization with the Fmoc-Rink amide linker. In duplicate, a small sample (approximately 3 mg) of dried resin was weighed accurately and transferred to a 10 mm quartz UV cuvette. 20 % piperidine in DMF (3.00 mL) was measured accurately and added to the cuvette which was capped and shaken gently for 2 h. The resin beads were allowed to settle before the absorbance at 290 nm (A_{290}) was measured and the new loading was calculated using the following equation (based upon an extinction coefficient, ε_{290} of 5253 $M^{-1}cm^{-1}$ for the piperidyl-fulvene adduct), averaging over the duplicate measurements:

$$\text{Loading (mmol/g)} = A_{290} / (\text{mg of resin used} \times 1.75)$$

6.2.4 Cleavage of Final PNA Oligomers from Solid-Support

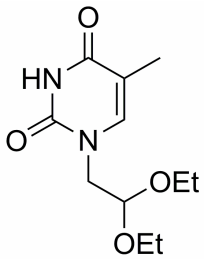
Final cleavages of full length PNAs with concomitant Bhoc and Boc deprotection were carried out using a TFA:TIS:DCM (90:5:5 v/v) cocktail for 2.5 h. The resulting

cleavage mixtures were evaporated to a small volume (< 0.1 mL) under a stream of nitrogen then the PNA oligomers were obtained by precipitation with Et_2O. The precipitates were collected by centrifugation and removal of the supernatant before being dried *in vacuo*. Aqueous solutions were prepared and PNA concentrations determined by measuring the absorbance at 260 nm (A_{260}) on an Agilent 8453 spectrophotometer using ε_{260} values of 6.6, 8.6, 13.7, 11.7 and 2.5 $mL\mu mol^{-1}cm^{-1}$ for C, T, A, G and the triphenylphosphonium tag respectively.[202, 203] PNA oligomers were characterized by MALDI-TOF mass spectrometry and HPLC.

6.3 Chapter 2 Experimental

6.3.1 Synthesis of Aldehydes and 'Blank' PNA Monomer

1-(2,2-Diethoxyethyl)-thymine (1)

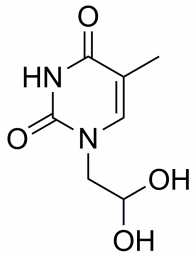


Route (a):[104-106] To a stirred suspension of thymine (1.60 g, 13 mmol) and dry K_2CO_3 (1.75 g, 13 mmol, 1 eq) in dry DMF (40 mL) under N_2 (g) was added bromoacetaldehyde diethyl acetal (1.97 mL, 13 mmol, 1 eq) and the resulting suspension was heated to 130 ^oC for 20 h. The reaction mixture was then cooled to room temperature and stirred for an additional 10 h before the solvent was removed *in vacuo* to afford a sticky brown solid. The crude product was dissolved in H_2O (65 mL) and extracted with EtOAc (3 × 100 mL). The combined organics were then washed with brine (65 mL), dried ($MgSO_4$), filtered and concentrated *in vacuo* to afford a brown oil. Purification by column chromatography (3.5 × 15 cm silica, eluting with 5 % MeOH:DCM) afforded **1** as an oil which solidified to an off-white solid on standing (864 mg, 3.6 mmol, 27 %).

Route (b): A mixture of thymine (800 mg, 6.3 mmol), bromoacetaldehyde diethyl acetal (984 μL, 6.3 mmol, 1 eq) and Cs_2CO_3 (4.1 g, 12.6 mmol, 2 eq) in DMF (20 mL) was heated at 100 ^oC (microwave) for 30 min. This reaction was repeated in a second vial, and both batches were combined and concentrated *in vacuo* to give a

brown solid. This was resuspended in H_2O (65 mL) and extracted with EtOAc (3 × 100 mL). The combined organics were then washed with brine (65 mL), dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford a yellow solid. This crude product was dissolved in 50 % MeOH:DCM and pre-adsorbed by evaporation onto silica, then purified by automated column chromatography (50 g column size, 3 CV DCM, 24 CV 0 → 5 % MeOH:DCM, 3 CV 5 % MeOH:DCM) to afford **1** as a yellow oil which solidified to an off-white solid on standing (333 mg, 1.4 mmol, 11 %). $\mathbf{R_f}$ = 0.33 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3212 (w), 2975 (w), 1668 (s), 1351 (m), 1034 (s); **mp** 101-102 °C, lit.[106] 106-109 °C; **^{1}H NMR** (600 MHz, $CDCl_3$) δ_H 8.04 (br s, 1H, NH), 7.09 (q, 1H, ^4J = 1.2 Hz, C=CH), 4.63 (t, 1H, ^3J = 5.4 Hz, C<u>H</u>$(OEt)_2$), 3.76 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.6 Hz, $^3J_{AX}$ = 7.2 Hz, OCH_2), 3.76 (d, 2H, ^3J = 5.4 Hz, NCH_2), 3.54 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.6 Hz, $^3J_{AX}$ = 7.2 Hz, OCH_2), 1.91 (d, 3H, ^4J = 1.2 Hz, C=CCH_3), 1.20 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7.2 Hz, 2 × CH_3); **^{13}C NMR** (150.9 MHz, $CDCl_3$) δ_C 164.6 (CO), 151.3 (CO), 142.0 (CH), 109.8 (C), 100.3 (CH), 64.2 (CH_2), 50.8 (CH_2), 15.2 (CH_3), 12.1 (CH_3); ***m/z*** **(ES$^+$)** 243 $(M+H)^+$, 265 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{11}H_{19}O_4N_2$ $(M+H)^+$: calcd 243.13486, found 243.13503; **HPLC** t_R = 3.56 min (method 2).

2-(2,4-Dihydroxy-5-methylpyrimidin-1-yl)-ethanal hydrate ($\mathbf{T_{CHO}}$)

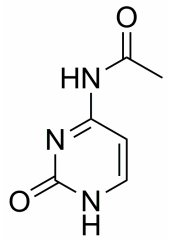


<u>Route (c):</u>[105, 106] A stirred suspension of **1** (157 mg, 0.65 mmol) in 1 M HCl aq (30 mL) was heated to reflux for 70 min, then cooled to room temperature and concentrated *in vacuo* to afford an oil which solidified to yield crude $\mathbf{T_{CHO}}$ as a colourless solid (164 mg).

<u>Route (d):</u> A suspension of **1** (111 mg, 0.46 mmol) in 1 M HCl aq (20 mL) was heated at 105 °C (microwave) for 30 min, then concentrated *in vacuo* to a yellow oil which solidified on standing. This was triturated with DCM to yield $\mathbf{T_{CHO}}$ as an off-

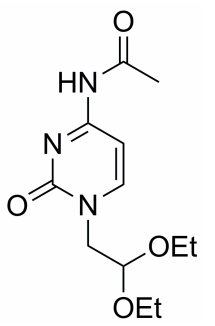
white solid which was collected by suction filtration (46 mg, 0.25 mmol, 54 %). $\mathbf{R_f}$ = 0.32 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3305 (m), 3164 (m), 3048 (m), 1672 (s), 1020 (s); **mp** decomp. > 164 °C, lit.[106] 200-210 °C; **^{1}H NMR** (250 MHz, D_2O) δ_H 7.44 (s, 1H, C=CH), 5.24 (t, 1H, ^3J = 5.2 Hz, C<u>H</u>$(OH)_2$), 3.80 (d, 2H, ^3J = 5.2 Hz, CH_2), 1.85 (s, 3H, C=CCH_3); **^{13}C NMR** (150.9 MHz, D_2O referenced to internal MeOH at δ_C 49.5) δ_C 167.7 (CO), 153.1 (CO), 144.5 (CH), 111.1 (C), 88.0 (CH), 53.7 (CH_2), 11.9 (CH_3); ***m/z* (ES$^+$)** 169 $(M+H)^+$, 187 $(M+H_3O)^+$; **HRMS (ES$^+$)** for $C_7H_9O_3N_2$ $(M+H)^+$: calcd 169.06187, found 169.06209; **HPLC** t_R = 1.55 min (method 3).

4-*N*-Acetylcytosine (2)[204]



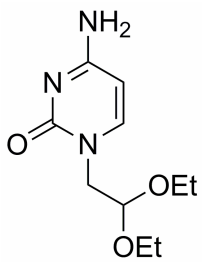
A stirred suspension of cytosine (1 g, 9 mmol) in acetic anhydride (10 mL) and glacial acetic acid (2 mL) was refluxed overnight (17 h) under N_2 (g). The reaction mixture was then cooled to room temperature and the insoluble material was collected by filtration, washed with cold EtOH (10 mL), then Et_2O (10 mL), and dried to afford **2** as an off-white solid (1.2 g, 7.8 mmol, 87 %). $\mathbf{R_f}$ = 0.33 (10 % MeOH:DCM); **^{1}H NMR** (500 MHz, d_6-DMSO) δ_H 11.50 (s, 1H, H-1), 10.74 (s, 1H, N<u>H</u>Ac), 7.80 (d, 1H, ^3J = 7 Hz, H-6), 7.09 (d, 1H, ^3J = 7 Hz, H-5), 2.08 (s, 1H, CH_3); **^{13}C NMR** (125.7 MHz, CD_3OD) δ_C 170.8 (CO), 163.1 (CO), 156.1 (C), 147.1 (CH), 94.4 (CH), 24.2 (CH_3); ***m/z* (ES$^+$)** 154 $(M+H)^+$, 176 $(M+Na)^+$.

4-*N*-Acetyl-1-(2,2-diethoxyethyl)-cytosine (3)



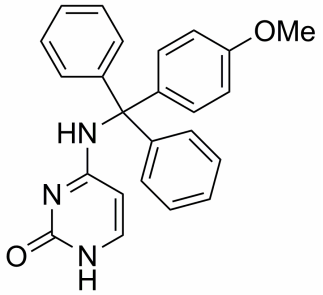
A mixture of **2** (250 mg, 1.6 mmol), dry K_2CO_3 (226 mg, 1.6 mmol, 1 eq) and bromoacetaldehyde diethyl acetal (253 µL, 1.6 mmol, 1 eq) in dry DMF (4.5 mL) was heated at 130 ^oC in the microwave for 30 min. The resulting dark brown/black reaction mixture was concentrated to a dark brown oil *in vacuo* then resuspended in H_2O (10 mL) and extracted with EtOAc (3 × 15 mL). The combined organics were washed with brine (10 mL), dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford a yellow solid. Purification by column chromatography (1.5 × 16 cm silica, eluting with 5 % MeOH:DCM) afforded **3** as an off-white solid (103 mg, 0.38 mmol, 24 %). $\mathbf{R_f}$ = 0.46 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2974 (w), 2931 (w), 1712 (m), 1655 (s), 1051 (s); **mp** 131-132 ^oC; **^{1}H NMR** (500 MHz, d_6-DMSO) δ_H 10.83 (s, 1H, NH), 7.96 (d, 1H, ^3J = 7 Hz, H-6), 7.14 (d, 1H, ^3J = 7 Hz, H-5), 4.71 (t, 1H, ^3J = 5.5 Hz, C<u>H</u>$(OEt)_2$), 3.84 (d, 2H, ^3J = 5.5 Hz, CH_2), 3.65 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 3.44 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 2.09 (s, 3H, NH(CO)C<u>H</u>$_3$), 1.06 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7 Hz, 2 × OCH_2C<u>H</u>$_3$); **^{13}C NMR** (125.7 MHz, CD_3OD) δ_C 172.9 (CO), 164.4 (CO), 158.7 (C), 152.4 (CH), 100.8 (CH), 97.5 (CH), 64.8 (CH_2), 54.1 (CH_2), 24.6 (CH_3), 15.7 (CH_3); ***m/z* (ES$^+$)** 270 $(M+H)^+$, 292 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{12}H_{20}O_4N_3$ $(M+H)^+$: calcd 270.14483, found 270.14456; **HPLC** t_R = 3.16 min (method 3).

1-(2,2-Diethoxyethyl)-cytosine (4)[105]



A solution of **3** (303 mg, 1.1 mmol) in methanolic ammonia (2M NH_3, 10 mL) was stirred at room temperature for 92 h, then the reaction mixture was concentrated *in vacuo* to a white solid and purified by column chromatography (2 × 16 cm silica, eluting with 5 % MeOH:DCM, with the crude sample dissolved in MeOH and pre-adsorbed on silica by evaporation) to afford **4** as a white solid (148 mg, 0.65 mmol, 58 %). $\mathbf{R_f}$ = 0.20 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3349 (m), 2975 (m), 1660 (s), 1616 (s), 1058 (s); **mp** 232-234°C, lit.[105] 232-234 °C; **^{1}H NMR** (500 MHz, CD_3OD) δ_H 7.49 (d, 1H, ^3J = 7.5 Hz, H-6), 5.82 (d, 1H, ^3J = 7.5 Hz, H-5), 4.73 (t, 1H, ^3J = 5 Hz, C<u>H</u>$(OEt)_2$), 3.81 (d, 2H, ^3J = 5 Hz, CH_2), 3.74 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 3.53 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 1.16 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7 Hz, 2 × $OCH_2C\underline{H}_3$); **^{13}C NMR** (125.7 MHz, CD_3OD) δ_C 168.2 (CO), 159.2 (C), 148.8 (CH), 101.5 (CH), 95.3 (CH), 65.0 (CH_2), 53.6 (CH_2), 15.7 (CH_3); ***m/z* (ES$^+$)** 228 $(M+H)^+$; **HRMS (EI)** for $C_{10}H_{17}O_3N_3$ M^+: calcd 227.12644, found 227.12631; **HPLC** t_R = 2.02 min (method 1).

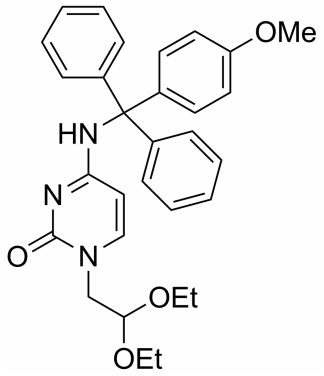
4-*N*-(4-Methoxytrityl)-cytosine (5)[85]



To a stirred suspension of cytosine (5 g, 45 mmol) and 4-methoxytrityl chloride (20.8 g, 67.5 mmol, 1.5 eq) in pyridine (225 mL) under N_2 (g) was added *N*-ethylmorpholine (5.7 mL, 45 mmol, 1 eq). The reaction mixture was heated to 40 °C for 30 min then stirred at room temperature for 16 h before the solvent was removed

in vacuo. The residue was resuspended in EtOAc (800 mL) and the insoluble material was collected by suction filtration, washed with EtOAc then Et_2O and dried in a vacuum oven at 40 ^oC overnight to afford crude **5** as a white solid (26.7 g). This was used directly without further purification. $\mathbf{R_f}$ = 0.63 (10 % MeOH:DCM); **^{1}H NMR** (500 MHz, d_6-DMSO) δ_H 10.25 (br s, 1H, NH), 8.25 (br s, 1H, NH), 7.37-6.78 (m, 16H, Ar-H), 3.72 (s, 1H, CH_3); ***m/z*** **(ES^+)** 406 $(M+Na)^+$, 422 $(M+K)^+$.

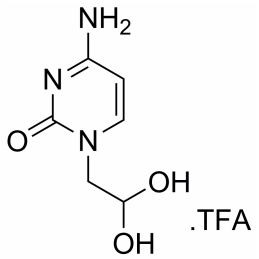
4-*N*-(4-Methoxytrityl)-1-(2,2-diethoxyethyl)-cytosine (6)



A suspension of crude **5** (2 g, 5.2 mmol), Cs_2CO_3 (3.4 g) and bromoacetaldehyde diethyl acetal (810 μL, 5.2 mmol, 1 eq) in DMF (15 mL) was heated at 100 ^oC in the microwave for 30 min. The solvent was then removed *in vacuo* and the residue was resuspended in H_2O (50 mL) and collected by filtration to afford an off-white solid which was washed further with H_2O (10 mL). This crude product was then dissolved in DCM (150 mL) and washed with 1M $KHSO_4$ aq (50 mL), 1M $NaHCO_3$ aq (2 × 50 mL), brine (50 mL) then dried (Na_2SO_4), filtered and concentrated *in vacuo* to yield 1.1 g of a white solid. Purification by column chromatography (4.5 × 16 cm silica, eluting with 5 % MeOH:DCM, with the crude product applied to the column as a suspension in 5 % MeOH:DCM) afforded **6** as a white solid (431 mg, 0.86 mmol, 17 %). $\mathbf{R_f}$ = 0.33 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2974 (w), 1653 (m), 1624 (s), 1489 (m), 1058 (s); **mp** 196-197 ^oC; **^{1}H NMR** (250 MHz, $CDCl_3$) δ_H 7.40-7.10 (m, 12H, Ar-H), 6.99 (d, 1H, ^3J = 7.5 Hz, H-6), 6.92-6.73 (m, 2H, Ar-H), 4.99 (d, 1H, ^3J = 7.5 Hz, H-5), 4.71 (t, 1H, ^3J = 5.5 Hz, C<u>H</u>$(OEt)_2$), 3.93-3.57 (m, 7H, O<u>CH_3</u> and N<u>CH_2</u> and O<u>CH_2</u>), 3.46 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.2 Hz, $^3J_{AX}$ = 7 Hz, O<u>CH_2</u>), 1.11 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7 Hz, 2 × OCH_2<u>CH_3</u>); **^{13}C NMR** (62.9 MHz, $CDCl_3$) δ_C 165.7 (CO), 158.6 (C), 156.0 (C), 146.7 (CH),

144.1 (C), 135.9 (C), 129.9 (CH), 128.5 (CH), 128.2 (CH), 127.4 (CH), 113.4 (CH), 100.4 (CH), 93.9 (CH), 70.3 (C), 64.6 (CH_2), 55.1 (CH_3), 52.7 (CH_2), 15.2 (CH_3); ***m/z* (ES$^+$)** 500 $(M+H)^+$, 522 $(M+Na)^+$, 538 $(M+K)^+$; **HRMS (ES$^+$)** for $C_{30}H_{34}O_4N_3$ $(M+H)^+$: calcd 500.25438, found 500.25468; **HPLC** t_R = 3.54 min (method 1).

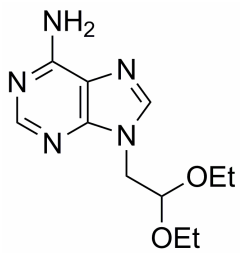
2-(Cytosin-1-yl)-ethanal hydrate trifluoroacetate (C_{CHO})



From **6**: A solution of **6** (263 mg, 0.53 mmol) in 1:1 v/v TFA:H_2O (10 mL) was heated at 100 ^oC in the microwave for 30 min then concentrated *in vacuo* to a viscous orange oil. Purification by column chromatography (2 × 15 cm silica, eluting with 25 % MeOH:DCM) afforded a straw-coloured oil which was redissolved in H_2O and lyophilized to give $\mathbf{C}_{CHO}$ as a white solid (137 mg, mmol, 91 %). NMR and LCMS suggest that some of the target is present as the methanolic hemiacetal.

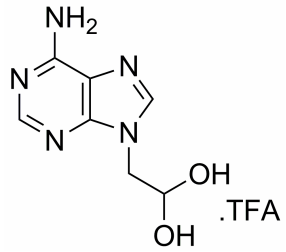
From **4**: A solution of **4** (10 mg, 44 μmol) in 1:1 v/v TFA:H_2O (0.5 mL) was heated at 100 ^oC in the microwave for 30 min then concentrated in vacuo to a viscous yellow oil which was lyophilized to give $\mathbf{C}_{CHO}$ as an off-white hygroscopic solid (13 mg, quantitative). $\mathbf{R}_f$ = 0.50 (25 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3205 (m), 16389 (s), 1610 (s), 1491(s), 1198 (s); **mp** decomp. > 110 ^oC; **^{1}H NMR** (360 MHz, D_2O) δ_H 7.83 (d, ^3J = 7.7 Hz, 1H, H-6), 6.17 (d, ^3J = 7.7 Hz, 1H, H-5), 5.30 (t, ^3J = 5.1 Hz, 1H, C<u>H</u>$(OH)_2$), 3.93 (d, ^3J = 5.1 Hz, 2H, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O:H_2O) δ_C 163.3 (q, ^2J = 36.4 Hz, <u>C</u>$(O)CF_3$), 160.4 (CO), 150.9 (CH), 149.9 (C), 116.8 (q, ^1J = 290.4 Hz, CF_3), 95.0 (CH), 87.7 (CH), 54.6 (CH_2); ***m/z* (ES$^+$)** 154 $(M+H)^+$, 172 $(M+H_3O)^+$, 176 $(M+Na)^+$; **HRMS (ES$^+$)** for aldehyde, $C_6H_8O_2N_3$ $(M+H)^+$: calcd 154.06110, found 154.06148; **HPLC** t_R = 0.67 min (method 5).

9-(2,2-Diethoxyethyl)-adenine (7)[104-106]



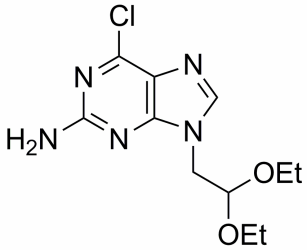
A mixture of adenine (500 mg, 3.7 mmol), K_2CO_3 (512 mg, 3.7 mmol, 1 eq) and bromoacetaldehyde diethyl acetal (574 μL, 3.7 mmol, 1 eq) in DMF (9 mL) was heated at 130 °C in the microwave for 30 min. This reaction was repeated twice and all three batches were combined and concentrated to dryness *in vacuo*. The resulting brown solid was resuspended in H_2O (20 mL), extracted with EtOAc (3 × 30 mL) and the combined organics were washed with brine (20 mL), dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford an off-white solid. Purification by column chromatography (4.5 × 24 cm silica, eluting with 5 % MeOH:DCM, sample dissolved in MeOH and pre-adsorbed on silica by evaporation) afforded **7** as an off-white solid (730 mg, 2.9 mmol, 26 %). $\mathbf{R_f}$ = 0.51 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3278 (w), 3112 (m), 1673 (s), 1602 (m), 1056 (s); **mp** 206-208 °C, lit.[85, 105] 212 °C; **^{1}H NMR** (500 MHz, d_6-DMSO) δ_H 8.14 (s, 1H, H-8), 8.05 (s, 1H, H-2), 7.20 (s, 2H, NH_2), 4.84 (t, 1H, ^3J = 5.5 Hz, C<u>H</u>$(OEt)_2$), 4.21 (d, 2H, ^3J = 5.5 Hz, NCH_2), 3.64 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 3.42 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7 Hz, OCH_2), 1.02 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7 Hz, 2 × $OCH_2C\underline{H}_3$); **^{13}C NMR** (62.9 MHz, $CDCl_3$) δ_C 155.8 (C), 152.4 (CH), 149.6 (C), 141.2 (CH), 118.3 (C), 99.4 (CH), 62.2 (CH_2), 45.2 (CH_2), 15.0 (CH_3); ***m/z* (ES$^+$)** 252 $(M+H)^+$, 274 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{11}H_{18}O_2N_5$ $(M+H)^+$: calcd 252.1455, found 252.1453; **HPLC** t_R = 3.37 min (method 4).

2-(Adenin-9-yl)-ethanal hydrate trifluoroacetate (A_{CHO})



9-(2,2-diethoxyethyl)-adenine, **7** (300 mg, 1.2 mmol) in 1:1 v/v TFA:H_2O (10 mL) was heated at 100 ^oC in the microwave for 30 min then concentrated *in vacuo* to give a viscous yellow oil which was triturated with Et_2O (2 × 20 mL) and dried *in vacuo* to give $\mathbf{A}_{CHO}$ as an off-white solid (376 mg, quantitative). $\mathbf{R}_f$ = 0.07 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3216 (w), 2970 (w), 1698 (s), 1622 (m), 1045 (s); **mp** decomp. > 120 ^oC; **^{1}H NMR** (360 MHz, D_2O) δ_H 8.42 (s, 1H, H-8), 8.35 (s, 1H, H-2), 5.44 (t, ^3J = 4.7 Hz, 1H, C<u>H</u>$(OH)_2$), 4.41 (d, ^3J = 4.7 Hz, 2H, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O:H_2O) δ_C 163.5 (q, ^3J = 35.2 Hz, <u>C</u>(O)CF_3), 150.7 (C), 149.4 (C), 145.9 (CH), 145.3 (CH), 118.5 (C), 116.9 (q, ^2J = 291.6 Hz, CF_3), 88.2 (CH), 49.9 (CH_2); ***m/z* (ES$^+$)** 178 $(M+H)^+$; **HRMS (ES$^+$)** for aldehyde, $C_7H_8O_1N_5$ $(M+H)^+$: calcd 178.07234, found 178.07234; **HPLC** t_R = 0.71 min (method 3); Anal. calc. for $C_9H_{10}N_5O_4F_3$: C 34.96, H 3.26, N 22.65; found: C 35.19, H 3.10, N 22.13.

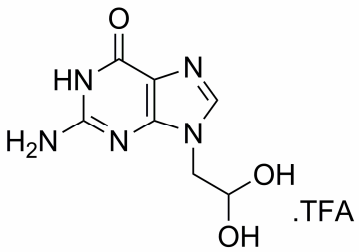
2-Amino-6-chloro-9-(2,2-diethoxyethyl)-purine (8)[205]



A suspension of 2-amino-6-chloropurine (500 mg, 2.9 mmol), Cs_2CO_3 (1.92 g, 5.9 mmol, 2 eq) and bromoacetaldehyde diethyl acetal (460 μL, 3.0 mmol, 1 eq) was heated at 100 ^oC in the microwave for 30 min. The reaction mixture was then concentrated *in vacuo* to a brown solid which was resuspended in H_2O (20 mL) and washed with EtOAc (3 × 30 mL). The combined organics were washed with brine (20 mL), dried (Na_2SO_4), filtered and concentrated *in vacuo* to give a yellow oil. Purification by column chromatography (3.5 × 16 cm silica, eluting with 5 %

MeOH:DCM, crude product dissolved in EtOAc/DCM/MeOH and pre-adsorbed on silica by evaporation) afforded **8** as a white solid (201 mg, 0.7 mmol, 24 %). $\mathbf{R_f}$ = 0.37 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3300 (w), 3183 (m), 1613 (s), 1558 (s), 1051 (s); **mp** 132-133 ^oC; **^{1}H NMR** (250 MHz, CD_3OD) δ_H 8.02 (s, 1H, H-8), 4.82 (t, 1H, ^3J = 5.1 Hz, C<u>H</u>$(OEt)_2$), 4.22 (d, 2H, ^3J = 5.1 Hz, NCH_2), 3.72 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.2 Hz, $^3J_{AX}$ = 7.1 Hz, OCH_2), 3.52 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.2 Hz, $^3J_{AX}$ = 7.1 Hz, OCH_2), 1.12 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7.1 Hz, 2 × $OCH_2C\underline{H}_3$); **^{13}C NMR** (62.9 MHz, CD_3OD) δ_C 161.7 (C), 155.5 (C), 151.4 (C), 145.4 (CH), 124.5 (C), 101.2 (CH), 64.6 (CH_2), 46.9 (CH_2), 15.6 (CH_3); ***m/z* (ES$^+$)** 286 $(M+H)^+$, 308 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{11}H_{17}O_2N_5Cl$ $(M+H)^+$: calcd 286.10653, found 286.10679; **HPLC** t_R = 2.58 min (method 1).

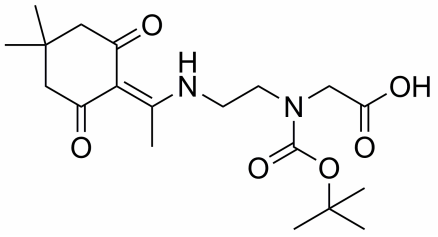
2-(Guanin-9-yl)-ethanal hydrate trifluoroacetate (G_{CHO})



2-Amino-6-chloro-9-(2,2-diethoxyethyl)-purine, **8** (10 mg, 35 mmol) in 1:2 v/v TFA:H_2O (0.5 mL) was heated at 100 ^oC (microwave) for 30 min then concentrated *in vacuo* to give a viscous yellow oil which was lyophilized to give $\mathbf{G_{CHO}}$ as an off-white solid (11 mg, quantitative). The ^{1}H and ^{13}C NMR spectra of this compound were very complicated when acquired immediately after dissolution in D_2O/H_2O, possibly due to the presence of multiple hydrogen bonded or imine species which were long-lived on the NMR timescale. However, upon standing overnight at room temperature, the NMR spectra became much simpler and consistent with the structure of the target hydrate trifluoroacetate. $\mathbf{R_f}$ = 0.07 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3120 (m), 1673 (s), 1598 (s), 1191 (s), 1048 (s); **mp** decomp. > 290 ^oC; **^{1}H NMR** (360 MHz, D_2O) δ_H 8.76 (s, 1H, H-8), 5.43 (t, ^3J = 4.7 Hz, 1H, C<u>H</u>$(OH)_2$), 4.31 (d, ^3J = 4.7 Hz, 2H, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O:H_2O) δ_C 163.2 (q, ^2J = 36.4 Hz, <u>C</u>$(O)CF_3$), 156.2 (C), 155.7 (C), 150.8 (C), 138.8 (CH),

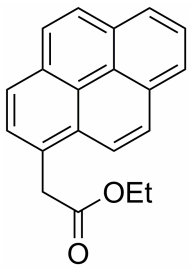
116.7 (q, ^1J = 290.8 Hz, CF_3), 108.0 (C), 87.4 (CH), 50.34 (CH_2); ***m/z* (ES$^+$)** 194 $(M+H)^+$, 212 $(M+H_3O)^+$, 216 $(M+Na)^+$; **HRMS (ES$^+$)** for aldehyde, $C_7H_8O_2N_5$ $(M+H)^+$: calcd 194.06725, found 194.06746; **HPLC** t_R = 0.71 min (method 5).

_N_-(2-[1-(4,4-Dimethyl-2,6-dioxocyclohexylidene)-ethylamino]-ethyl)-_N_-(*tert*-butoxycarbonyl)-glycine (10)



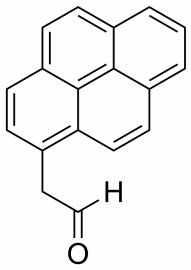
To a solution of methyl *N*-(2-[1-(4,4-Dimethyl-2,6-dioxocyclohexylidene)-ethylamino]-ethyl)-glycinate (**9**; for the synthesis of **9**, see compound **40**, Chapter 6.6.1)[83] (300 mg, 1.0 mmol) in THF (10 mL) was added di-*tert*-butyl dicarbonate (250 µL, 1.1 mmol) and triethylamine (150 µL, 1.1 mmol) and the reaction mixture was stirred for 5 hours at room temperature. After removal of the solvent in vacuo, the crude product was dissolved in DCM (200 mL) and washed with 1 M $NaHCO_3$ aq (2 × 50 mL), 1 M $KHSO_4$ aq (2 × 50 mL) and brine (50 mL). The organic phase was dried over Na_2SO_4, filtered and concentrated in vacuo to give a yellow solid. Without any further purification the crude was dissolved in MeOH (100 mL) and a 2 M solution of Cs_2CO_3 in water (100 mL) was added. After stirring at room temperature for 1.5 h, the reaction was acidified to pH 3 with sat. $KHSO_4$ aq and the precipitate was collected by filtration, washed with water (20 mL) and dried *in vacuo* to afford **10** as a white solid (279 mg, 0.7 mmol, 73 %). $\mathbf{R_f}$ = 0.09 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2935 (w), 1731 (m), 1683 (s), 1587 (s), 1139 (s); **mp** 117-118 ^oC; **^{1}H NMR** (360 MHz, $CDCl_3$) two rotamers: δ_H 4.03 and 3.95 (s, 2H, CH_2COO), 3.66 (m, 2H, CH_2), 3.53 (m, 4H, CH_2), 2.57 (s, 3H, CCH_3), 2.35 (s, 4H, Dde-C$\underline{H_2}$), 1.41 and 1.44 (s, 9H, Boc-C$\underline{H_3}$), 1.01 (s, 6H, Dde-C$\underline{H_3}$) ppm; **^{13}C NMR** (90.6 MHz, $CDCl_3$) two rotamers: δ_C 198.2 (CO), 174.8 (CO), 172.6 and 172.4 (C), 155.3 and 155.2 (CO), 108.1 (C), 81.5 and 81.1 (C), 52.5 (CH_2), 50.8 (CH_2), 48.4 and 48.2 (CH_2), 42.1 and 42.0 (CH_2), 30.1 (CH_3), 28.3 and 28.1 (C), 28.2 (CH_3), 18.0 (CH_3); ***m/z* (ES$^-$)** 381 $(M-H)^-$, 763 $(2M-H)^-$; **HRMS (EI)** for $C_{19}H_{30}O_6N_2$ $(M)^+$: calcd 382.20984, found 382.20990; **HPLC** t_R = 3.88 min (method 3).

Ethyl 1-pyreneacetate (11)[206]



1-Pyreneacetic acid (500 mg, 1.9 mmol) was suspended in a mixture of EtOH (2 mL) toluene (10 mL) and conc. H_2SO_4 and heated to reflux (oil bath at 120 ^oC) under N_2 (g) with a Dean-Stark water trap for 4 h. The black solution was then cooled to room temperature and washed with 5 % $NaHCO_3$ aq (2 × 10 mL) and brine (10 mL), dried (Na_2SO_4), filtered (washing the residue with toluene) and concentrated to a brown oil *in vacuo*. Purification by column chromatography (2.2 × 5.7 cm silica, eluting with toluene) afforded a yellow oil which was recrystallized from *n*-hexane to yield **11** as a pale yellow solid (410 mg, 1.4 mmol, 74 %). $\mathbf{R_f}$ = 0.36 (toluene); **IR** ν_{max}/ cm^{-1} (neat) 2982 (w), 1725 (s), 1602 (m), 1368 (m), 1029 (s); **mp** 63-65 ^oC, lit.[206] 66.5-67.5 ^oC; **^{1}H NMR** (250 MHz, $CDCl_3$) δ_H 8.29-7.93 (m, 9H, Ar<u>H</u>), 4.35 (s, 2H, Ar<u>C</u>H_2), 4.16 (q, 2H, ^3J = 7.5 Hz, OCH_2), 1.22 (t, 6H, ^3J = 7.5 Hz, CH_3); **^{13}C NMR** (125.7 MHz, $CDCl_3$) δ_C 171.6 (CO), 131.3 (C), 130.8 (C), 130.8 (C), 129.5 (C), 128.4 (CH), 128.3 (C), 127.9 (CH), 127.4 (CH), 127.3 (CH), 126.0 (CH), 125.2 (CH), 125.1 (CH), 125.0 (C), 124.9 (CH), 124.8 (C), 123.3 (CH), 61.0 (CH_2), 39.6 (CH_2), 14.2 (CH_3); ***m/z* (ES$^+$)** 289 $(M+H)^+$, 311 $(M+Na)^+$.

1-Pyreneacetaldehyde (Y_{CHO})[206]



A stirred solution of ethyl 1-pyreneacetate, **11** (200 mg, 0.69 mmol) in anhydrous toluene (6.9 mL) under N_2 (g) was cooled to -78 ^oC before 1 M DIBAL-H in *n*-hexane (0.69 mL) was added dropwise and stirring was continued at -78 ^oC for 2 h. The reaction mixture was then quenched carefully (i.e. keeping the temperature

below -74 °C) with a solution of 37 % w/w HCl (0.17 mL) in THF (1.55 mL) and stirring was continued at -78 °C for 5 min. The reaction mixture was warmed to room temperature, decanted from white insoluble precipitate and washed with H_2O (2 × 7 mL) and brine (7 mL), dried ($MgSO_4$), decanted and concentrated *in vacuo* to afford a yellow oil which was recrystallized from *n*-hexane to afford crude $\mathbf{Y}_{CHO}$ (154 mg, of approximately 85 % purity as judged by HPLC) as a yellow solid. An analytical sample (16 mg) of $\mathbf{Y}_{CHO}$ was obtained by preparative HPLC (method 1) of a portion (23 mg) of this crude material. **IR** ν_{max}/ cm^{-1} (neat) 2828 (w), 2726 (w), 1709 (m), 1379 (m), 1058 (m); **mp** 109-110 °C, lit.[206] 111.5-112.5 °C; **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 9.90 (t, 1H, ^3J = 2.5 Hz, CHO), 8.23-7.96 (m, 9H, Ar<u>H</u>), 4.40 (d, 2H, ^3J = 2.5 Hz, Ar<u>C</u>H_2); **^{13}C NMR** (125.7 MHz, $CDCl_3$) δ_C 199.3 (CO), 131.3 (C), 131.1 (C), 130.8 (C), 129.8 (C), 128.6 (CH), 128.4 (CH), 127.6 (CH), 127.4 (CH), 12.2 (CH), 125.6 (C), 125.5 (CH), 125.4 (CH), 125.2 (C), 125.1 (CH), 124.7 (C), 122.8 (CH), 48.8 (CH_2); ***m/z* (ES$^+$)** 267 $(M+Na)^+$, 299 $(M+MeOH+Na)^+$; **HPLC** t_R = 4.47 min (method 3).

6.3.2 Synthesis of PNA Oligomers and T_m Measurements

PNA oligomers **P1-5** (Table 2.1) were synthesized* on a solid support using a Rink amide linker attached to PEGA resin (Polymer Labs, UK)[207] according to a literature method.[83] Fmoc/Bhoc protected PNA monomers (Link Technologies, UK) were employed alongside **10**, which was used as the monomer to insert 'blank' positions. Couplings were carried out using PyBOP/NEM (for 3 h) as described elsewhere.[83, 208] Fmoc deprotections were carried out using 20% piperidine in DMF for (2 × 6) min while Dde deprotection was achieved using a solution of $NH_2OH.HCl$/imidazole (1.25g and 0.9 g respectively) in NMP (5 mL) and DMF (1 mL) for (2 × 1.5 h).[208] Phosphonium tags (PNA **2-5**) were added using (4-carboxybutyl)triphenyl-phosphonium bromide (Sigma-Aldrich, UK) while *N*-Fmoc protected PEG units, PEG1 (PNA **2** and **4**) and PEG2 (PNA **3**) were commercially available (Sigma-Aldrich, UK and Polypeptide Labs, France respectively).

* **P1-5** and **P6** (Chapter 6.4) were synthesised by Dr Juan Jose Diaz-Mochon. Characterization was carried out by the author.

P1: ***m/z* (MALDI-TOF MS)**; for $C_{154}H_{201}N_{74}O_{46}$ $(M+H)^+$: calcd 3824.57, found 3824.36; **HPLC** t_R = 5.40 min (method 6).

P2: ***m/z* (MALDI-TOF MS)**; for $C_{182}H_{230}N_{86}O_{44}P^+$ $(M)^+$: calcd 4356.82, found 4356.64; **HPLC** t_R = 7.57 min (method 6).

P3: ***m/z* (MALDI-TOF MS)**; for $C_{171}H_{216}N_{84}O_{40}P^+$ $(M)^+$: calcd 4118.72, found 4118.77; **HPLC** t_R = 7.28 min (method 6).

P4: ***m/z* (MALDI-TOF MS)**; for $C_{169}H_{219}N_{81}O_{39}P^+$ $(M)^+$: calcd 4039.74, found 4039.42; **HPLC** t_R = 7.42 min (method 6).

P5: ***m/z* (MALDI-TOF MS)**; for $C_{165}H_{206}N_{72}O_{43}P^+$ $(M)^+$: calcd 3916.59, found 3916.22; **HPLC** t_R = 6.99 min (method 6).

The T_m values (Table 2.3) for **P1-5** and complementary DNA oligomers **I-VI** (Table 2.2) were determined using a Varian Cary 300 Bio UV/Vis spectrometer. An initial heating cooling cycle was performed over the range 20-80 ^oC to permit hybridization of a 1:1 mixture of the PNA and DNA strands (2 μM) in 4.5 mM pH 6 phosphate buffer (500 μL final volume). The change in A_{260nm} was then recorded over this range at 1 ^oC intervals, heating at a rate of 1 ^oC/min, and the T_m was determined from the maximum of the first derivative of the resulting curve.

6.3.3 Base-Filling Reactions

General

Aqueous solutions of each aldehyde were prepared and concentrations confirmed by ^1H NMR in 10 % D_2O:H_2O. For $\mathbf{T}_{CHO}$, a solution of thymine of known concentration was used as an internal standard, and the concentration confirmed by the relative integral peak areas for the CH_3 protons. This solution of $\mathbf{T}_{CHO}$ of known concentration was used as an internal standard to confirm the concentrations of the other aldehydes by comparing the relative integral peak areas for the CH_2 protons.

For MALDI-TOF analysis, $+BH_3$, +C, +T, +A, +G and +Y incorporation result in mass increases of +14, +137, +152, +161, +177 and +228 Da respectively.

(i) Reactions at equimolar aldehyde concentrations (Chapter 2.4: Figure 2.3, Graph 2.1, and Table 2.4)

A PNA blank (2.5 μL, 40 μM aq), DNA template (1 μL, 100 μM aq), aldehydes $\mathbf{A}_{CHO}$, $\mathbf{G}_{CHO}$, $\mathbf{C}_{CHO}$, and $\mathbf{T}_{CHO}$ (1.6 μL of each, 1.7 mM aq), and pH 6 phosphate buffer (8.1 μL, 10 mM aq) were combined in a 1.5 mL Eppendorf tube (Eppendorf AG) and placed in an Eppendorf Thermomixer comfort (Eppendorf AG) at 80 ^{o}C and 1200 rpm for 5 min. The reaction mixture was then cooled to 40 ^{o}C (3 ^{o}C/min) before $NaBH_3CN$ (2 μL, 1 M aq) was added and shaking continued for 1 h. Pre-treated Q Sepharose® Fast Flow (5 μL; see above) was then added before the reaction mixture was agitated at room temperature for 20 min. The reaction tube was centrifuged and the supernatant removed, then the resin was washed centrifugally with 3% MeCN in water (3 × 200 mL). Sinapic acid matrix (10 μL) was added to the resin, and this mixture was spotted (1 μL in duplicate) onto a stainless steel MALDI plate (Applied Biosystems). Reactions were performed in duplicate, and five MALDI spectra acquired for each reaction. Spectra are presented unprocessed. Relative peak intensities were determined for the most common isotopes of the PNA-incorporation products. Product signal ratios were determined by averaging over the ten spectra.

(ii) Test of reversibility (Chapter 2.4: Figure 2.4)

To test the reversibility of iminium formation, a reaction was performed as above (i) but without $\mathbf{G}_{CHO}$ (water was added to maintain the final reaction volume and component concentrations). The reaction was then repeated, but this time $\mathbf{A}_{CHO}$, $\mathbf{C}_{CHO}$ and $\mathbf{T}_{CHO}$ were allowed to react for 1 h at 40 ^{o}C, before $\mathbf{G}_{CHO}$ was added followed by 2 μL of 1 M $NaBH_3CN$, and the reaction left for 1 h at 40 ^{o}C. Treatment with Q Sepharose® and MALDI-TOF analysis were performed as above.

(iii) Reactions at non-equimolar aldehyde concentrations (Chapter 2.4: Figure 2.5, Graph 2.2, and Table 2.5)

Conditions as reported above (i) with the following amounts of nucleobase aldehydes: 1.6 μL of $\mathbf{C}_{CHO}$ (2.2 mM), 1.6 μL of $\mathbf{T}_{CHO}$ (5.6 mM), 1.6 μL of $\mathbf{A}_{CHO}$ (3.3 mM) and 1.6 μL of $\mathbf{G}_{CHO}$ (1.7 mM). A single reaction was run under each set of

conditions, and three MALDI spectra acquired for each. Product signal ratios were determined by averaging over the three spectra.

(iv) Effect of pH (Chapter 2.4: Figure 2.6)

For investigation of the pH dependence of the incorporation reaction, conditions were as above (iii), using PNA **P1** (Table 2.1) and DNA **II** (Table 2.2) but employing the relevant buffer (8.1 μL of a 10 mM aqueous buffer solution). pH 8.5 was buffered with *N*-tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid (TAPS), pH 5.0 with sodium acetate, and intermediate pH values with sodium phosphate.

(v) Analysis of mixtures of DNA templates (Chapter 2.4: Figure 2.7 and 2.8, and Table 2.6)

For initial investigations (Figure 2.7), conditions as above (iii) but using 0.5 μL of each of two DNA templates (100 μM). Improved results (Figure 2.8) were obtained using different amounts of the four aldehydes, specifically 0.6 μL of $\mathbf{C}_{CHO}$ (2.2 mM), 2.6 μL of $\mathbf{T}_{CHO}$ (5.6 mM), 2.6 μL of $\mathbf{A}_{CHO}$ (3.3 mM) and 0.6 μL of $\mathbf{G}_{CHO}$ (1.7 mM). A control reaction was performed with the same aldehyde ratio but with water (1 μL) in place of DNA.

(vi) Templated reaction at multiple contiguous blanks (Chapter 2.5)

Conditions as above (iii) using PNA oligomers **P2-4** (Table 2.1), DNA **V** (Table 2.2) and the following amounts of the four aldehydes: 0.6 μL of $\mathbf{C}_{CHO}$ (2.2 mM), 2.6 μL of $\mathbf{T}_{CHO}$ (5.6 mM), 2.6 μL of $\mathbf{A}_{CHO}$ (3.3 mM) and 0.6 μL of $\mathbf{G}_{CHO}$ (1.7 mM).

(vii) Abasic site analysis (Chapter 2.6)

For abasic site-templated base-filling in the absence of $\mathbf{Y}_{CHO}$, conditions were as above (iii) using PNA **P5** (Table 2.1) and DNA **VI** (Table 2.2). For abasic site analysis with $\mathbf{Y}_{CHO}$, conditions were as for (i), but with the addition of 1.6 μL of $\mathbf{Y}_{CHO}$ (1.7 mM) and less buffer (6.5 μL) to maintain a constant reaction volume.

(viii) Analysis of RNA (Chapter 2.7)

Conditions as above (i) with PNA **P1**, but with RNA **RI** in place of DNA.

6.4 Chapter 3 Experimental

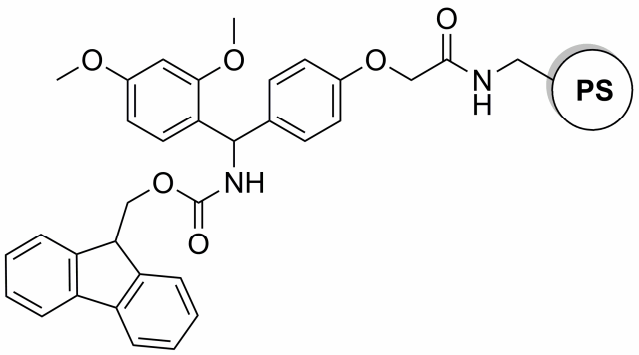
6.4.1 Synthesis of PNA Oligomers

Probes **P5-6** were synthesized as described above (Chapter 6.2.2).

P5: (See Chapter 6.3.2 for characterization data).

P6: ***m/z*** **(MALDI-TOF MS)**; for $C_{165}H_{208}N_{66}O_{46}P^+$ $(M)^+$: calcd 3882.58, found 3882.60; **HPLC** t_R = 7.09 min (method 6).

Fmoc-Rink amide PS Resin



Low-loading aminomethylpolystyrene resin (1.0 g, ~ 0.5 mmol/g) was swollen in DCM for 1 h. Meanwhile, Fmoc-Rink linker (809 mg, 1.5 mmol) was dissolved in DMF (15 mL) and activated with DIC (189 mg, 1.5 mmol) and HOBt.H_2O (203 mg, 1.5 mmol) for 15 min. The resin was filtered and the solution of activated Rink linker was added. After agitating for 3 h, the resin was filtered, washed with DMF (3 × 3 mL) then DCM (3 × 3 mL), and a small sample of resin was analyzed by qualitative ninhydrin test to confirm the absence of free amines. The resin was washed with MeOH (3 × 3 mL) then Et_2O (3 × 3 mL) and dried *in vacuo*. **Loading** = 0.20 mmol/g (theoretical loading = 0.40 mmol/g, yield = 50 %). Prior to use, Fmoc-Rink PS resin was pre-swollen in DCM and deprotected using 20 % piperidine in DMF (2 × 6 min) and washed with DMF (3 ×) and DCM (3 ×) to afford Rink amide PS.

Probes **P7-8** were synthesized on Rink amide PS (37.5 mg each, 7.5 μmol) as follows, using monomers as *per* Chapter 6.3.2 (yields are based upon UV determination of the concentration of a solution of the total PNA dissolved in water).

General procedure for PNA monomer couplings

PNA monomer (3 eq) was dissolved in DMF (to 0.1 M concentration) and activated with DIC (3 eq) and oxyma (3 eq) for 5 min, then added to the pre-swollen resin in an SPE tube. The resulting mixture was then transferred to a glass microwave vial (5 mL size, Biotage AB, Sweden) using a plastic Pasteur pipette. A small magnetic stirrer was added to the vial, which was then sealed and transferred to a heating block (aluminium, containing holes of 17 mm diameter and 30 mm depth to hold microwave vials, prepared by the mechanical workshop of the School of Chemistry, University of Edinburgh) on a hot plate at 60 ^{o}C. The reaction mixture was stirred very slowly at 60 ^{o}C for 30 min, then transferred back to the SPE tube (again with a plastic Pasteur pipette) for washing with DMF (3 ×) and DCM (3 ×).

Fmoc and Dde deprotection steps

As *per* Chapter 6.3.2, which were always followed by DMF (3 ×) and DCM (3 ×) washings before the next coupling or capping step.

P7: ***m/z* (MALDI-TOF MS)**; for $C_{167}H_{204}N_{82}O_{36}P^+$ $(M)^+$: calcd 3966.64, found 3966.63; **HPLC** t_R = 5.33 min (method 6); **Yield** = 50 % (95 % per monomer).

P8: ***m/z* (MALDI-TOF MS)**; for $C_{178}H_{217}N_{89}O_{38}P^+$ $(M)^+$: calcd 4241.76, found 4241.82; **HPLC** t_R = 5.21 min (method 6); **Yield** = 46 % (95 % per monomer).

6.4.2 Model Studies using Synthetic DNA

(i) G551D and W1282X (Chapter 3.3: Figure 3.2 and 3.4)

Conditions for the homo- and heterozygous SNP (G551D and W1282X) analyses were as *per* Chapter 6.3.3(i), using the optimized ratios of the four nucleobase aldehydes described in Chapter 6.3.3(v). Homozygous models employed 1 μL of the relevant DNA template (100 μM), whilst heterozygous studies used 0.5 μL of both templates (100 μM). The experiment that modelled duplex analysis of both SNPs (Figure 3.4a) employed 0.5 μL of each DNA template (100 μM), 2.5 μL of PNA **5** (40 μM) , 2.5 of μL of PNA **6** (40 μM) and 5.6 μL pH 6 10mM phosphate buffer

(final reaction volume of 20 μL after addition of aldehydes and $NaBH_3CN$ solution). A substoichiometric amount of PNA **6** was used to normalize the product ratios (Figure 3.4b) using 2.5 μL of PNA **5** (40 μM) , 1.5 of μL of PNA **6** (40 μM) and 6.6 μL pH 6 10mM phosphate buffer (to maintain a final reaction volume of 20 μL).

(ii) ΔF508 (Chapter 3.3: Figure 3.3)

Model analyses for the ΔF508 mutation were performed as above (i) but this time 2.5 μL of each of two probes (**P7** and **P8**; 40 μM) was used along with 5.6 μL of pH 6 10mM phosphate buffer (to maintain a final reaction volume of 20 μL).

(iii) Investigation of the effect of [PNA] >> [DNA] (Chapter 3.3: Figure 3.5)

Conditions were as above (i) but 6 μL of **P5** (167 μM) and 4.6 μL of pH 6 10 mM phosphate buffer (final reaction volume 20 μL) were used along with 1 μL of 10 μM DNA **VII** (Figure 3.5a) or 1 μM DNA **VII** (Figure 3.5b) to give PNA:DNA ratios of 100:1 and 1000:1 respectively.

(iv) Detection limit for PNA following dynamic incorporation

To determine the detection limit for the charge-tagged PNA, varying amounts of DNA template **VI** (10 pmol, 1 pmol and 100 fmol) were used in three templated incorporation reactions (conditions as *per* (i), using 1 μL of DNA solutions of concentration 10 μM, 1 μM and 0.1 μM respectively) with a parallel decrease in the amount of PNA **6** blank to maintain a 1:1 stoichiometry (i.e. by using 2.5 μL of PNA solutions of concentration 4 μM, 0.4 μM and 0.04 μM respectively). In each case the desired incorporation product was observable in unprocessed spectra, although for 100 fmol the product was only weakly detectable at 'sweet spots' on the MALDI plate. However, given that only $1/10^{th}$ of the reaction product (i.e. 1 μL of a 10 μl dilution with sinapic acid matrix) was spotted onto the plate, then it can be concluded that a reliable detection limit of 100 fmol of DNA is possible with the charge tagged PNA (although detection at the 10 fmol level was possible in some cases). Investigation of base-filling on acyl-capped **P1** using DNA **III** (Chapter 2.3) in the

same way showed a detection limit of 1 pmol, which supports the use of the triphenylphosphonium charge tag to improve detection limits.

6.4.3 Human Genomic DNA samples

For 'in-house' and clinical samples, DNA was isolated from buccal (cheek) swabs (Isohelix for in-house samples, medical wire (MWE) Dryswab for clinical samples) using a commercial (Isohleix) DNA isolation kit and kept at -20 ^{o}C in the storage buffer provided until analysis. The concentrations of DNA in the samples were determined by UV absorbance using a GE Healthcare NanoVue.

To obtain clinical samples (Chapter 3.5 and 3.6), subjects were identified from the Western General Hospital Adult Cystic Fibrosis Unit database. This database consists of 159 adults with CF; 147 have fully identified genotypes, of whom 17 possess the G551D mutation and 63 are ΔF508 homozygotes. Inclusion criteria included ability to provide written informed consent. Ethical approval was granted by Lothian Regional Ethics Committee. 13 Patients were approached to participate in the study on attendance at CF outpatient clinic with an invitation letter and patient information sheet. 12 (6 male) agreed to provide a sample. 7 of these were homozygous or heterozygous for G551D and 5 were ΔF508 homozygotes. The genotyping reported herein was performed by individuals who were blind to the previously determined patient genotypes. All participants provided written consent before a buccal swab was provided by each. All swabs were taken by the same CF clinical research fellow (Dr Philip Andrew Reid) and stored at -20 ^{o}C until isolation.

Commercial DNA samples (Coriell Cell Repositories, USA) were supplied as solutions of known concentration in 10 mM Tris, 1 mM EDTA buffer at pH 8.0. The repository numbers for samples 'Cor1' and Cor2' were NA08338 and NA12785 respectively. More information on these samples (including genotypic and phenotypic data) is available at http://ccr.coriell.org/.

6.4.4 PCR Amplification

General

PCR primers were selected from a previous publication[55] and purchased from Microsynth. For the G551D SNP, primer sequences were:

Forward: 5'-CTTGGAGAAGGTGGAATCAC-3';

Reverse: 5'-AAATGCTTGCTAGACCAATA-3'.

For the ΔF508 indel, primer sequences were:

Forward 5'-AGTTTTCCTGGATTATGCCT-3';

Reverse 5'-TTGGGTAGTGTGAAGGGTTC-3'.

PCR amplifications were performed on a Techne TC-312 Thermocycler and all water was nuclease free. Cycling conditions were an initial denaturation at 95 ^{o}C for 3 min, 40 cycles of 95 ^{o}C for 30 s, annealing at 47 ^{o}C for 45 s, and extension at 72 ^{o}C for 45 s, a final extension at 78 ^{o}C for 5 min, and a final hold at 4 ^{o}C.

(i) Two-stage asymmetric PCR (Chapter 3.4 ('in-house' samples), 3.5 and 3.6 (clinical samples))

For the first symmetric PCR step (final volume, 50 μL), 2 μL of isolated genomic DNA solution was amplified (for DNA concentrations see Table 3.4 and 3.5; these were not normalized) and reagent concentrations were: 1X PCR mastermix (Promega), 0.4 μM forward and reverse primers. The genomic DNA was replaced with water for negative controls.

For the subsequent asymmetric PCR (final volume, 50 μL), 3 μL of the first crude PCR mixture was used. Reagent concentrations were: 1X PCR mastermix (Promega), 1 μM forward primer. Negative controls were performed with 3 μL of the negative controls from the first PCR stage. For individual mutation analysis, crude PCR products were purified individually into 28 μL of 10 mM pH 7.4 PBS using a QIAGEN® QIAquick® PCR purification kit. For duplex analysis, the ssDNA PCR products for both G551D and ΔF508 were combined and purified together into 28 μL of 10 mM pH 7.4 PBS as above.

(ii) One-stage asymmetric PCR (Chapter 3.4 (commercial samples))

For amplification (singleplex and duplex, Figure 3.8 and 3.9 respectively), 2 μL of isolated genomic DNA solution was amplified (for DNA concentrations see Table 3.4; these were not normalized) and reagent concentrations were: 1X PCR mastermix (Promega), 1.0 μM forward and 0.2 μM reverse primers. The genomic DNA was

replaced with water for negative controls. Crude PCR products were purified into 28 μL of 10 mM pH 7.4 PBS using a QIAGEN® QIAquick® PCR purification kit.

6.4.5 Allele Discrimination by Dynamic Chemistry

Depending upon the analysis (i.e. G551D, ΔF508 or duplex), the following were added to the purified, amplified DNA (28 μL):

G551D analysis - 1 μL of PNA **P5** (40 μM), 1.8 μL $\mathbf{C}_{CHO}$ (2.2 mM), 7.8 μL $\mathbf{T}_{CHO}$ (5.6 mM), and 15.4 μL water.

ΔF508 analysis - 1 μL of PNA probes **P7** and **P8** (40 μM), 1.8 μL $\mathbf{G}_{CHO}$ (1.7 mM), 7.8 μL $\mathbf{A}_{CHO}$ (3.3 mM), and 14.4 μL water.

Duplex G551D and ΔF508 analysis - 1 μL of PNA probes **P5**, **P7** and **P8** (40 μM), 1.8 μL $\mathbf{G}_{CHO}$ (1.7 mM), 1.8 μL $\mathbf{C}_{CHO}$ (2.2 mM), 7.8 μL $\mathbf{T}_{CHO}$ (5.6 mM), 7.8 μL $\mathbf{A}_{CHO}$ (3.3 mM), and 3.8 μL water.

The resulting mixtures were heated (in a Techne TC-312 Thermocycler) at 95 ^{o}C for 5 min, then cooled to 40 ^{o}C at ~ 3 ^{o}C/min. 6 μL of freshly prepared 1 M aqueous sodium cyanoborohydride was added (final volume 60 μL) and the mixture left at 40 ^{o}C for 1 h, then 10 μL of pre-treated Q sepharose® Fast Flow was added and the mixture left to shake at 20 ^{o}C for 20 min. The resin was then centrifuged, the supernatant removed, and the resin washed centrifugally (for 30 sec at 13,000 rpm) with 3 % aqueous acetonitrile (3 × 200 μL). Finally, 10 μL of sinapic acid matrix was added to the resin, and 1 μL of the resulting mixture spotted in duplicate directly onto a stainless steel MALDI-TOF plate for MS analysis. The unreacted PNA probe of lowest mass was used as an internal calibrant. Genotypes were determined by visual inspection of the unprocessed MALDI spectra. For the clinical samples, genotypes were also established by input of the MS peak table data into a Microsoft Excel file (briefly, raw data for peaks above a certain intensity threshold obtained from the MALDI spectra were analyzed using Excel formulae, which converted numerical peak m/z values within a set range into an output genotype; see Appendix 3 for representative examples of the resulting Excel spreadsheets for each clinical sample). Peaks of relative intensity < 5 % of the most intense peak were disregarded.

6.4.6 Agarose Gel Electrophoresis

Agarose gels were prepared using 1.8-2.0 % w/v Type XI agarose (Sigma-Aldrich) and 0.5 μg/mL ethidium bromide (Sigma-Aldrich, UK) in 1 X TBE buffer (Fischer Scientific, UK). Gels were run for 45-60 min at 90 V on a Bio-Rad Power PAC 300 instrument and imaged using UV light (Uvitec transilluminator). Size markers were low molecular weight DNA ladders (0.5 μg/lane, New England Biolabs, USA). PCR products (20 μL from 50 μL reactions) and ladders were loaded using blue/orange loading dye (Promega, UK). For the gel-shift assay of single-stranded DNA (lane 2, Figure 3.12), PCR product (20 μL) was pre-hybridized with PNA **P5** (0.4 μL of a 40 μM solution) by heating to 95 ^{o}C for 2 min and cooling to 4 ^{o}C in a PCR thermocycler. As a control for the gel-shift assay (lane 1, Figure 3.12), **P5** (0.4 μL of a 40 μM solution) was loaded directly with blue/orange loading dye.

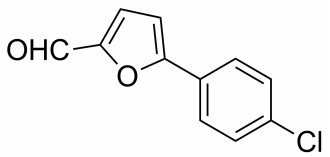
6.5 Chapter 4 Experimental

6.5.1 Library Screening for Nucleobase Analogues (Chapter 4.3)

(i) Library compounds (Figure 4.3)

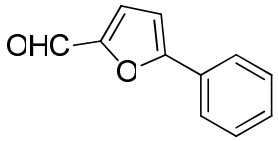
Compounds **18-21**, **23-25** and **27-28** were commercially available (from Sigma-Aldrich, with the exception of **24** from Acros Organics). Compounds **12-17**, **22**, **26** and **29** were synthesized as part of a separate project by Dr Jeff Walton and Dr Sunay Chankeshwara, who supplied small (< 4 mg) quantities of each together with ^{1}H NMR and ES-MS characterization data.

5-(4-Chlorophenyl)furan-2-carbaldehyde (12)



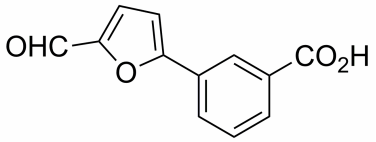
Beige solid. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 9.66 (s, 1H, CHO), 7.75 (d, 2H, ^{3}J = 8.6 Hz, Ar-H), 7.42 (d, 2H, ^{3}J = 8.6 Hz, Ar-H), 7.32 (d, 1H, ^{3}J = 3.7 Hz, Ar-H), 6.83 (d, 1H, ^{3}J = 3.7 Hz, Ar-H); ***m/z* (ES^{+})** 207 $(M+H)^{+}$.

5-Phenylfuran-2-carbaldehyde (13)



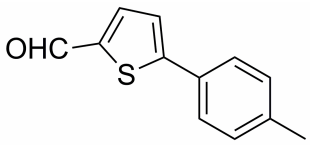
Colourless oil. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 9.66 (s, 1H, CHO), 7.82 (dd, 2H, ^3J = 8.4, ^4J = 2.0 Hz, Ar-H), 7.47-7.38 (3H, m, Ar-H), 7.32 (d, 1H, ^3J = 3.7 Hz, Ar-H), 6.85 (d, 1H, ^3J = 3.7 Hz, Ar-H); ***m/z*** **(ES$^+$)** 173 $(M+H)^+$.

3-(5-Formylfuran-2-yl)benzoic acid (14)



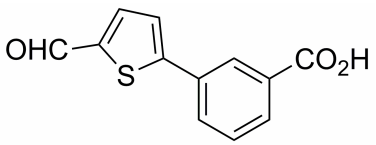
Brown solid. **^{1}H NMR** (360 MHz, d_6-DMSO) δ_H 9.64 (s, 1H, CHO), 8.38 (s, 1H, Ar-H), 8.13 (d, 1H, ^3J = 7.5 Hz, Ar-H), 8.00 (d, 1H, ^3J = 7.4 Hz, Ar-H), 7.68-7.64 (m, 2H, Ar-H), 7.64 (1H, d, ^3J = 7.7 Hz, Ar-H), 7.43 (d, 1H, J = 3.8 Hz, Ar-H); ***m/z*** **(ES$^+$)** 215 $(M-H)^-$.

5-*p*-Tolylthiophene-2-carbaldehyde (15)



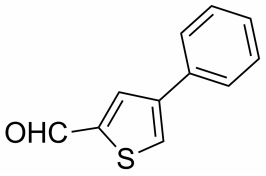
Off-white solid. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 9.80 (s, 1H, CHO), 7.65 (d, 1H, ^3J = 4.0 Hz, Ar-H), 7.49 (d, 2H, ^3J = 8.3 Hz, Ar-H), 7.29 (d, 1H, ^3J = 4.0 Hz, Ar-H), 7.16 (d, 2H, ^3J = 8.4 Hz, Ar-H), 2.32 (s, 3H, CH_3); ***m/z*** **(ES$^+$)** 203 $(M+H)^+$.

3-(5-Formylthiophen-2-yl)benzoic acid (16)



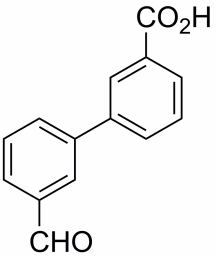
Yellow solid. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 13.34-13.11 (br s, 1H, OH), 9.93 (s, 1H, CHO), 8.26 (t, 1H, ^4J =1.7 Hz, Ar-H), 8.08-8.05 (m, 2H, Ar-H), 8.00-7.97 (m, 1H, Ar-H), 7.85 (1H, d, ^3J = 4.0 Hz, Ar-H), 7.63 (1H, t, ^3J = 7.8 Hz, Ar-H); ***m/z*** **(ES$^-$)** 231 $(M-H)^-$.

4-Phenylthiophene-2-carbaldehyde (17)



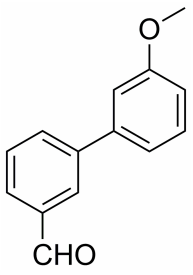
Pale yellow solid. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 9.97 (d, 1H, ^4J =1.4 Hz, CHO), 8.04 (s, 1H, ^4J =1.4 Hz, Ar-H), 7.86 (s, 1H, Ar-H), 7.81-7.56 (m, 2H, Ar-H), 7.47-7.34 (m, 3H, Ar-H); ***m/z* (ES$^+$)** 189 $(M+H)^+$.

3'-Formylbiphenyl-3-carboxylic acid (22)



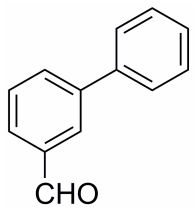
Beige solid. **^{1}H NMR** (360 MHz, d_6-DMSO) δ_H 13.29-13.01 (br s, 1H, OH), 10.12 (s, 1H, CHO), 8.26-8.25 (m, 2H, Ar-H), 8.07 (d, 1H, ^3J = 7.9 Hz, Ar-H), 8.03-7.98 (m, 2H, Ar-H), 7.94 (d, 1H, ^3J = 7.6 Hz, Ar-H), 7.72 (t, 1H, ^3J = 7.6 Hz, Ar-H), 7.65 (t, 1H, ^3J = 7.9 Hz, Ar-H); ***m/z* (ES$^-$)** 225 $(M-H)^-$.

4'-Methoxybiphenyl-3-carbaldehyde (26)



White solid. **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 10.08 (s, 1H, CHO), 8.06 (s, 1H, Ar-H), 7.83-7.80 (m, 2H, Ar-H), 7.61-7.56 (m, 3H, Ar-H), 7.01 (d, 2H, ^3J = 8.7 Hz, Ar-H), 3.87 (s, 3H, CH_3); ***m/z* (ES$^+$)** 213 $(M+H)^+$.

Biphenyl-3-carbaldehyde (29)



Colourless oil. **^{1}H NMR** (250 MHz, $CDCl_3$) δ_H 10.10 (s, 1H, CHO), 8.11 (t, 1H, ^4J = 1.6 Hz, Ar-H), 7.87 (dd, 2H, ^4J = 1.6 Hz, ^3J = 7.6 Hz, Ar-H), 7.66-7.62 (m, 3H, Ar-H), 7.52-7.40 (m, 3H, Ar-H); ***m/z*** **(ES$^+$)** 183 $(M+H)^+$.

(ii) Screening by dynamic chemistry (Figure 4.4 and Graph 4.1)

The library of eighteen aldehydes (Chapter 4, Figure 4.3) was dissolved in 5:5:1 v/v H_2O:MeOH:MeCN to a concentration of 1.5 mM in each aldehyde. The organic solvents were necessary to allow complete dissolution. Dynamic incorporation of this library was then investigated in four separate experiments using PNA **P1** (Chapter 2, Table 2.1) and one of the four DNA templates **I-IV** (Chapter 2, Table 2.2; templating bases are G, A, T and C respectively). The standard protocol described above was followed (Chapter 6.3.3(i)) but the aldehyde library solution (6.4 μL) was used in place of the natural nucleobase aldehydes. For MALDI-TOF analysis, mass increases expected for aldehydes **12-29** are given below (Table 6.1).

Table 6.1 Mass increases calculated for incorporation of library members.

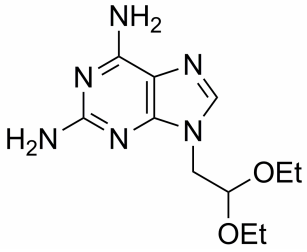
Aldehyde	Expected Mass Increase on Incorporation (Da)
12	190
13	156
14	200
15	186
16	216
17	172
18	138
19	178
20	79
21	91
22	210

23	141
24	144
25	129
26	196
27	118
28	204
29	166

6.5.2 Targeted Synthesis and Screening of Nucleobase Analogues (Chapter 4.3)

(i) Aldehyde synthesis (Scheme 4.1)

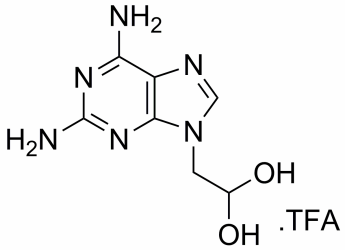
2,6-Diamino-9-(2,2-diethoxyethyl)-purine (30)



A solution of **8** (101 mg, 0.35 mmol) in 7 M NH_3 in MeOH (2 mL) was heated at 65 ^{o}C in the microwave for 30 min. TLC (10 % MeOH:DCM) showed the presence of mainly unreacted starting material, so the reaction mixture was saturated with NH_3 by bubbling NH_3 (g) through the solution for 15 min before heating at 90 ^{o}C in the microwave for 1 h (pressure = 4-5 bar). The reaction mixture was concentrated to dryness *in vacuo* and purified by automated column chromatography (10 g column size, 3 CV 5 % MeOH:DCM, 24 CV 5 → 10 % MeOH:DCM, 9 CV 10 % MeOH:DCM, sample dissolved in MeOH and pre-adsorbed on silica by evaporation) to afford **30** as a white solid (18 mg, 0.07 mmol, 19 %). $\mathbf{R_f}$ = 0.51 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3446 (w), 3306 (m), 3119 (m), 1656 (s), 1055 (s); **mp** 169-170 ^{o}C; **^{1}H NMR** (250 MHz, CD_3OD) δ_H 7.73 (s, 1H, H-8), 4.79 (t, 1H, ^{3}J = 5.4 Hz, C<u>H</u>$(OEt)_2$), 4.14 (d, 2H, ^{3}J = 5.4 Hz, NCH_2), 3.73 (dq, A of an ABX_3 spin system, 2H, $^{2}J_{AB}$ = 9.4 Hz, $^{3}J_{AX}$ = 7.0 Hz, OCH_2), 3.50 (dq, B of an ABX_3 spin system, 2H, $^{2}J_{AB}$ = 9.2 Hz, $^{3}J_{AX}$ = 7.0 Hz, OCH_2), 1.12 (t, X_3 of an ABX_3 spin system, 6H, $^{3}J_{AX}$ = 7.0 Hz, 2 × OCH_2C<u>H</u>$_3$); **^{13}C NMR** (125.7 MHz, CD_3OD) δ_C

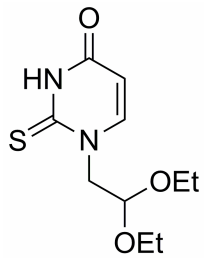
161.9 (C), 157.5 (C), 153.0 (C), 140.7 (CH), 113.8 (C), 101.5 (CH), 64.6 (CH_2), 46.8 (CH_2), 15.6 (CH_3); ***m/z*** **(ES$^+$)** 267 $(M+H)^+$, 289 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{11}H_{19}O_2N_6$ $(M+H)^+$: calcd 267.15640, found 267.15607; **HPLC** t_R = 2.59 min (method 3).

2-(2,6-Diaminopurin-9-yl)-ethanal hydrate trifluoroacetate (D_{CHO})



A suspension of **30** (5.5 mg, 21 μmol) in 1:1 v/v TFA:H_2O (250 μL) was heated at 100 ^oC in the microwave for 30 min then concentrated *in vacuo* to remove most of the TFA before lyophilizing to give $\mathbf{D}_{CHO}$ as an off-white solid (7.2 mg, quantitative). $\mathbf{R}_f$ = 0.19 (10 % MeOH:DCM); **^{1}H NMR** (600 MHz, 10 % D_2O in H_2O) δ_H 8.02 (s, 1H, H-8), 5.41 (t, 1H, ^3J = 4.8 Hz, C<u>H</u>$(OH)_2$), 4.22 (d, 2H, ^3J = 4.8 Hz, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O:H_2O) δ_C 163.5 (q, ^2J = 35.4 Hz, <u>C</u>$(O)CF_3$), 162.6 (C), 153.3 (C), 151.0 (C), 143.5 (CH), 116.9 (q, ^1J = 291.5 Hz, CF_3), 111.6 (C), 88.2 (CH), 49.4 (CH_2); ***m/z*** **(ES$^+$)** 211 $(M+H)^+$; **HRMS (ES$^+$)** $C_7H_{11}O_2N_6$ $(M+H)^+$: calcd 211.09380, found 211.09362; **HPLC** t_R = 0.78 min (method 3).

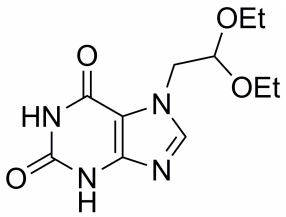
1-(2,2-Diethoxyethyl)-2-thiouracil (31)



To a stirred suspension of 2-thiouracil (1.0 g, 7.8 mmol) in dry MeCN (20 mL) under N_2 (g) was added *N,O*-bis(trimethylsilyl)acetamide (5.3 mL, 21.8 mmol, 2.8 eq). The resulting pale yellow solution was stirred for 5 min before DiPEA (1.4 mL, 8.0 mmol, 1 eq) then bromoacetaldehyde diethyl acetal (2.1 mL, 8.0 mmol, 1 eq) was added and the reaction mixture was stirred at room temperature for 16 h. TLC (10 %

MeOH:DCM) indicated the presence of mainly starting material, so a 20 mL aliquot of the reaction mixture was removed and heated at 100 ^{o}C for 30 min, before KI (0.9 g, 5.4 mmol) was added and the resulting mixture heated in the microwave at 100 ^{o}C for a further 30 min. The mixture was then quenched with MeOH (50 mL), neutralized with $NaHCO_3$ (3.0 g), and filtered (washing with MeOH) and concentrated *in vacuo* to give a yellow oil which was purified by automated column chromatography (50 g column size, 32 CV 10 % MeOH:DCM) to afford **31** as a white solid (149 mg, 0.6 mmol, 8 %). $\mathbf{R_f}$ = 0.65 (10 % MeOH:DCM);); **IR** ν_{max}/ cm^{-1} (neat) 2977 (m), 2876 (m), 2686 (m), 1684 (s), 1278 (s); **mp** 104-106 ^{o}C; **^{1}H NMR** (360 MHz, $CDCl_3$) δ_H 11.07 (br s, 1H, NH), 7.77 (d, 1 H, ^{3}J = 6.8 Hz, CH), 6.19 (d, 1 H, ^{3}J = 6.8 Hz, CH), 4.74 (t, 1H, ^{3}J = 5.0 Hz, C<u>H</u>$(OEt)_2$), 3.78 (dq, A of an ABX_3 spin system, 2H, $^{2}J_{AB}$ = 9.2 Hz, $^{3}J_{AX}$ = 7.1 Hz, OCH_2), 3.50 (dq, B of an ABX_3 spin system, 2H, $^{2}J_{AB}$ = 9.2 Hz, $^{3}J_{AX}$ = 7.1 Hz, OCH_2), 3.31 (d, 2H, ^{3}J = 5.0 Hz, NCH_2), 1.12 (t, X_3 of an ABX_3 spin system, 6H, $^{3}J_{AX}$ = 7.1 Hz, 2 × $OCH_2C\underline{H}_3$); **^{13}C NMR** (150.9 MHz, $CDCl_3$) δ_C 164.1 (CO), 161.6 (CS), 154.6 (CH), 111.3 (CH), 101.2 (CH), 62.8 (CH_2), 34.1 (CH_2), 15.2 (CH_3); ***m/z* (ES^{-})** 243 $(M-H)^{-}$; **HRMS (EI)** for $C_{10}H_{16}O_3N_2S$ $(M)^{-}$: calcd 244.08761, found 244.08778; **HPLC** t_R = 3.43 min (method 3).

7-(2,2-Diethoxyethyl)-xanthine (32)



A mixture of xanthine (1.0 g, 6.6 mmol), Cs_2CO_3 (4.3 g, 13.1 mmol, 2 eq) and bromoacetaldehyde diethyl acetal (1.02 mL, 6.6 mmol, 1 eq) in DMF (16 mL) was heated at 100 ^{o}C in the microwave for 30 min. TLC (10 % MeOH:DCM) showed mainly starting material, so the reaction mixture was heated at 130 ^{o}C in the microwave for a further 30 min then left to stand at room temperature for 40 h. The reaction mixture was then filtered (washing with DMF) and concentrated to a white solid *in vacuo*. Attempted recrystallization from hot MeOH was unsuccessful, so the suspension in hot MeOH was filtered and pre-adsorbed onto SiO_2 by evaporation and

purified by column chromatography (2 × 16 cm silica, eluting with 5 % MeOH:DCM) to afford 19 mg of a white solid. This 19 mg was sonicated in CD_3OD (1 mL), transferred to an Eppendorf tube, centrifuged and the supernatant removed for analysis by ^{1}H NMR which showed the presence of two isomers; isomer A and isomer B in a ratio of ~ 3:1. 2D HMBC NMR showed isomers A and B to be the *N*7- and *N*3-alkylated products respectively (Figure 6.1). The residual solid was sonicated in MeOH (1 mL), then centrifuged and the supernatant removed as before. Upon drying the residue *in vacuo*, a white solid was obtained which was shown to be the pure *N*7-alkylated isomer **32** (5 mg, 19 μmol, 0.3 %). $\mathbf{R_f}$ = 0.40 (5 % MeOH:DCM); **^{1}H NMR** (500 MHz, CD_3OD) δ_H 7.83 (s, 1 H, H-8), 4.83 (t, 1H, ^3J = 5.0 Hz, C<u>H</u>$(OEt)_2$), 4.36 (d, 2H, ^3J = 5.0 Hz, NCH_2), 3.73 (dq, A of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7.0 Hz, OCH_2), 3.50 (dq, B of an ABX_3 spin system, 2H, $^2J_{AB}$ = 9.5 Hz, $^3J_{AX}$ = 7.0 Hz, OCH_2), 1.13 (t, X_3 of an ABX_3 spin system, 6H, $^3J_{AX}$ = 7.0 Hz, 2 × $OCH_2C\underline{H}_3$); **^{13}C NMR** (125.7 MHz, CD_3OD) δ_C 157.6 (CO), 153.6 (CO), 150.8 (C), 145.0 (CH), 108.4 (C), 102.0 (CH), 64.8 (CH_2), 49.9 (CH_2), 15.7 (CH_3); ***m/z* (ES$^+$)** 269 $(M+H)^+$; **HRMS (EI)** for $C_{11}H_{16}O_4N_4$ $(M)^+$: calcd 268.11650, found 268.11661; **HPLC** t_R = 3.53 min (method 3).

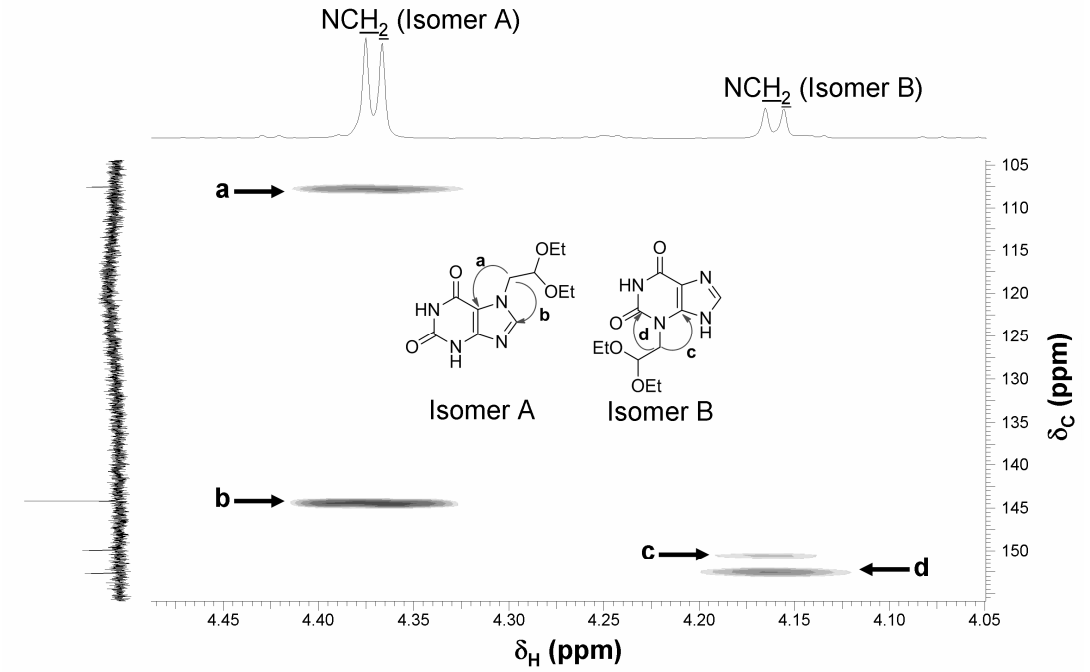
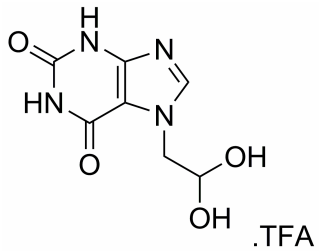


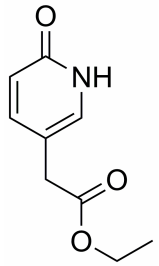
Figure 6.1 ^{1}H-^{13}C HMBC of a mixture of isomers A (major) and B (minor) obtained after alkylation of xanthine. Based upon the observed long-range (3 bond) couplings it can be concluded that A is the *N*7-alkylated product, and B is the *N*3-alkylated product.[209]

2-(Xanthin-9-yl)-ethanal hydrate trifluoroacetate (X_{CHO})



A solution of **32** (5 mg, 19 μmol) in 1:1 v/v TFA:H_2O (500 μL) was heated at 100 ^oC in the microwave for 30 min then concentrated *in vacuo* to remove most of the TFA before lyophilizing to give $\mathbf{X}_{CHO}$ as a white solid (5 mg, 16 μmol, 87 %). $\mathbf{R}_f$ = 0.08 (10 % MeOH:DCM); **^{1}H NMR** (600 MHz, 10 % D_2O in H_2O) δ_H 10.78 (s, 1H, NH), 7.93 (s, 1H, H-8), 5.37 (t, ^3J = 5.1 Hz, 1H, C<u>H</u>$(OH)_2$), 4.35 (d, ^3J = 5.1 Hz, 2H, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O:H_2O) δ_C 163.5 (q, ^2J = 35.4 Hz, <u>C</u>(O)CF_3), 157.6 (CO), 153.6 (CO), 149.9 (C), 144.9 (CH), 116.9 (q, ^1J = 291.9 Hz, CF_3), 108.1 (C), 88.7 (CH), 52.3 (CH_2); ***m/z* (ES$^+$)** 213 $(M+H)^+$; **HRMS (ES$^+$)** $C_7H_9O_4N_4$ $(M+H)^+$: calcd 213.06183, found 213.06196; **HPLC** t_R = 1.41 min (method 3).

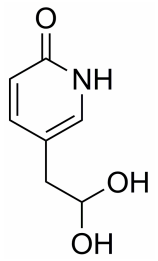
Ethyl 6-oxo-1,6-dihydro-3-pyridylacetate (33)



6-chloro-3-pyridylacetic acid (1.0 g, 5.8 mmol) was suspended in 10 M KOH aq (10 mL) and heated in the microwave at 205 ^oC for 25 min (17 bar pressure). **Warning**; these conditions cause obvious corrosion of the microwave vial, which should be disposed of after the reaction. The resulting black solution was acidified to pH ~ 1 with 2 M HCl aq and concentrated to a brown solid *in vacuo*. This was resuspended in EtOH (20 mL) with 4 drops of 37 % w/w HCl, and then heated in the microwave at 85 ^oC for 30 min. The reaction mixture was concentrated to a brown solid *in vacuo*, then suspended in saturated NH_4Cl aq (20 mL) and water (20 mL) and extracted with EtOAc (6 × 50 ml). The combined organics were dried ($MgSO_4$), filtered and concentrated *in vacuo* to yield **33** as a white solid (0.76 g. 4.2 mmol, 72

%). **R_f** = 0.59 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2804 (m), 1726 (s), 1657 (s), 1615 (s), 1177 (s); **mp** 133-135 °C; **^{1}H NMR** (500 MHz, $CDCl_3$) δ_H 12.63 (br s, 1H, NH), 7.44 (dd, 1H, ^4J = 2.0, ^3J = 9.5 Hz, H-4), 7.25 (d, 1 H, ^4J = 2.0 Hz, H-5), 6.57 (d, 1H, ^3J = 9.5 Hz, H-2), 4.16 (q, 2 H, ^3J = 7.2 Hz, OCH_2), 3.37 (s, 2H, CH_2), 1.26 (t, 3 H, ^3J = 7.2 Hz, CH_3); **^{13}C NMR** (125.7 MHz, $CDCl_3$) δ_C 170.8 (CO), 164.4 (CO), 143.4 (CH), 133.5 (CH), 120.4 (CH), 113.0 (C), 61.2 (CH_2) 36.9 (CH_2), 14.2 (CH_3); ***m/z* (ES$^+$)** 182 $(M+H)^+$, 204 $(M+Na)^+$; **HRMS (EI$^+$)** for $C_9H_{11}O_3N$ $(M)^+$: calcd 181.07324, found 181.07334; **HPLC** t_R = 2.75 min (method 3).

6-Oxo-1,6-dihydro-3-pyridylacetaldehyde hydrate (T*$_{CHO}$)



A stirred solution of **33** (25 mg, 0.14 mmol) in anhydrous DCM (1.4 mL) was cooled to -78 °C under N_2 (g) and 1 M DIBAL-H in *n*-hexane (0.21 mL, 0.21 mmol) was added dropwise. The reaction mixture was stirred at -78 °C for 2 h, then quenched with a solution of 37 % w/w HCl (40 μL) in THF (0.31 mL) before DCM (3 mL) was added and the mixture washed with water. TLC showed aldehyde in the aqueous phase only. Attempted extractions of the aqueous phase (with EtOAC and $CHCl_3$) were unsuccessful, so the water was removed *in vacuo* to afford a white solid (110 mg) which was purified by preparative HPLC (method 2) to yield **T*$_{CHO}$** as a yellow solid (8 mg, 0.05 mmol based on mass of hydrated product, 36 %). NMR showed peaks attributed to both the hydrated (major) and aldehydic (minor) products. **R_f** = 0.19 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3228 (m), 3131 (m), 1655 (s), 1593 (s), 1055 (s); **mp** decomp. > 250 °C; **^{1}H NMR** (500 MHz, 10 % D_2O in H_2O) hydrate: δ_H 7.67 (dd, 1H, ^4J = 2.0, ^3J = 9.5 Hz, H-4), 7.42 (d, 1 H, ^4J = 2.0 Hz, H-5), 6.61 (d, 1H, ^3J = 9.5 Hz, H-2), 5.17 (t, 1 H, ^3J = 5.5 Hz, OCH) and 2.74 (d, 2H, ^3J = 5.5 Hz, CH_2); aldehyde: δ_H 9.70 (s, 1H, CHO), 7.58 (dd, 1H, ^4J = 2.0, ^3J = 9.5 Hz, H-4), 7.42 (d, 1 H, ^4J = 2.0 Hz, H-5), 6.64 (d, 1H, ^3J = 9.5 Hz, H-2), 3.77 (s, 2H, CH_2); **^{13}C NMR** (125.7 MHz, 10 % D_2O in H_2O) hydrate: δ_C 164.6 (CO), 146.7 (CH),

134.9 (CH), 119.0 (CH), 118.6 (C), 91.2 (CH), 39.7 (CH_2); aldehyde: δ_C 204.6 (CO), 164.6 (CO), 146.5 (CH), 135.5 (CH), 119.4 (CH), 114.3 (C), 45.6 (CH_2); ***m/z* (ES$^+$)** 138 $(M+H)^+$, 156 $(M+H_3O)^+$; **HRMS (ES$^+$)** for $C_7H_8O_2N$ $(M+H)^+$: calcd 138.05496, found 138.05496; **HPLC** t_R = 1.86 min (method 3).

(ii) Dynamic incorporation of non-natural aldehydes in the presence of their natural counterparts (Figure 4.5 and 4.6, Graph 4.2 and Table 4.1)

Aqueous solutions of aldehydes $\mathbf{D}_{CHO}$, $\mathbf{X}_{CHO}$ and $\mathbf{T^*}_{CHO}$ were prepared and concentrations confirmed by ^{1}H NMR in 10 % D_2O:H_2O using an internal standard ($\mathbf{A}_{CHO}$ was used as a standard for $\mathbf{D}_{CHO}$ and $\mathbf{X}_{CHO}$, and 6-chloro-3-pyridylacetic acid for $\mathbf{T^*}_{CHO}$). Dynamic incorporation of each non-natural nucleobase aldehyde was investigated in separate experiments using PNA **P1** (Chapter 2, Table 2.1) and the 'correct' DNA template (**III** for $\mathbf{D}_{CHO}$, **II** for $\mathbf{X}_{CHO}$ and $\mathbf{T^*}_{CHO}$; Chapter 2, Table 2.2). The standard protocol described above was followed (Chapter 6.3.3(i)) but the natural nucleobase aldehydes were replaced with an equimolar mixture of the non-natural nucleobase under investigation (3.2 μL, 1.7 mM) and its natural counterpart (3.2 μL, 1.7 mM). For MALDI-TOF analysis, +D, +X, and +T* incorporation result in mass increases of +176, +178, and +121 Da respectively.

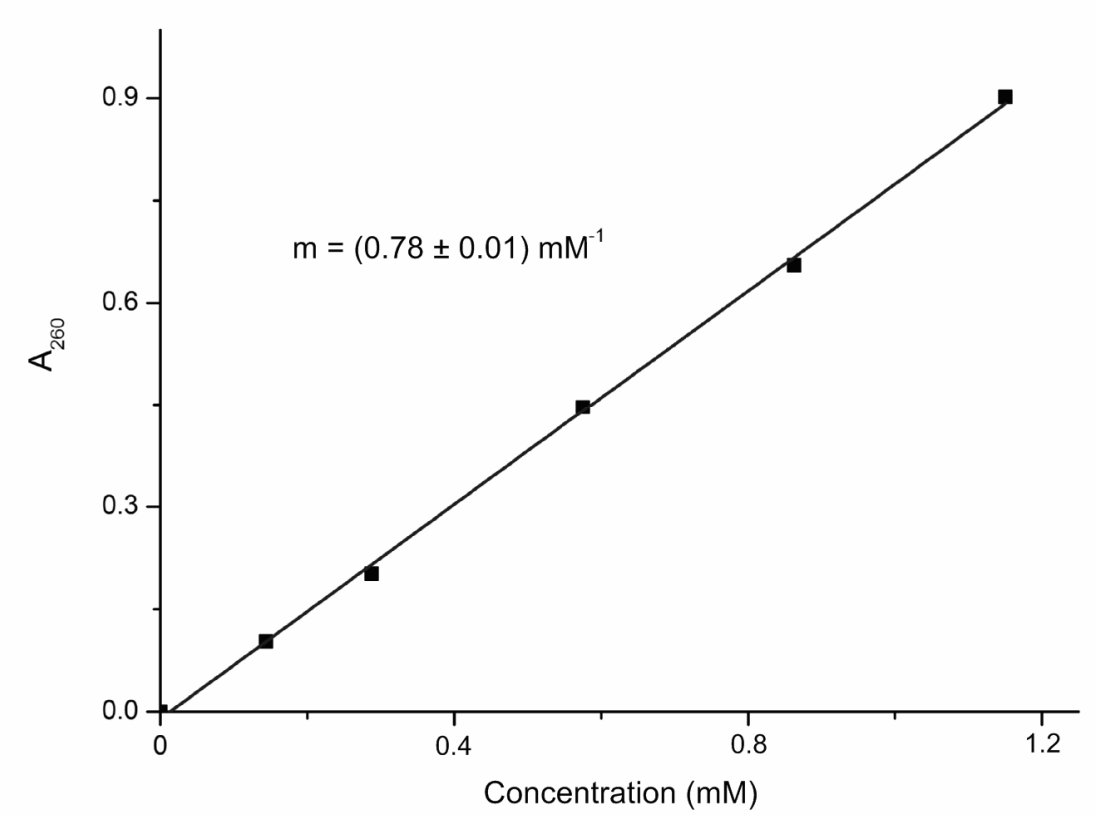
For the investigation of the selectivity for $\mathbf{T^*}_{CHO}$ against mis-incorporation, four separate experiments using PNA **P1** (Chapter 2, Table 2.1) and one of the four DNA templates **I-IV** (Chapter 2, Table 2.2; templating bases are G, A, T and C respectively) were performed according to the standard protocol described above (Chapter 6.3.3(i)), but with $\mathbf{T^*}_{CHO}$ in place of $\mathbf{T}_{CHO}$.

6.5.3 Comparison of Dynamic Incorporation Results with T_m Values for T* (Chapter 4.4)

(i) PNA synthesis (Table 4.2 and Figure 4.7)

PNA oligomers **P9-12** were synthesized on Rink amide PS (25 mg each, 7.4 μmol; Chapter 6.4.1, method as for **P7-8**). The non-natural nucleobase T* was added using monomer **36**, which was synthesized as detailed below. Yields are based upon UV determination of the concentration of a solution of the total PNA dissolved in water.

For the pyridone monomer, the extinction coefficient at 260 nm (ε_{260}) was estimated to be 0.8 mLμmol^{-1}cm^{-1} (see Graph 6.1).



Graph 6.1 Effect of concentration on the A_{260} of ethyl 6-oxo-1,6-dihydro-3-pyridylacetate **33**, used to estimate the extinction coefficient of T*. Measurements were made at 28 ^{o}C. The extinction coefficient at 260 nm, ε_{260} = (Absorbance at 260 nm, A_{260} /(concentration, c x path length, l))= (gradient, m)/l = 0.78 mM^{-1}cm^{-1} = 0.78 mLμmol^{-1}cm^{-1} (since path length, l = 1 cm).

Crude PNA oligomers **P10-12** were purified by preparative HPLC (method 1). **P9** was used directly without purification.

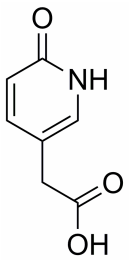
P9: ***m/z*** **(MALDI-TOF MS)**; for $C_{114}H_{149}N_{60}O_{31}$ $(M+H)^+$: calcd 2855.19, found 2855.26; **HPLC** t_R = 12.90 min (method 7); **Yield** = 16 % (85 % per monomer).

P10: ***m/z*** **(MALDI-TOF MS)**; for $C_{114}H_{148}N_{59}O_{30}$ $(M+H)^+$: calcd 2824.19, found 2824.64; **HPLC** t_R = 12.70 min (method 7); **Yield** = 4 % (75 % per monomer).

P11: ***m/z*** **(MALDI-TOF MS)**; for $C_{114}H_{147}N_{58}O_{29}$ $(M+H)^+$: calcd 2793.18, found 2793.79; **HPLC** t_R = 12.50 min (method 7); **Yield** = 6 % (77 % per monomer).

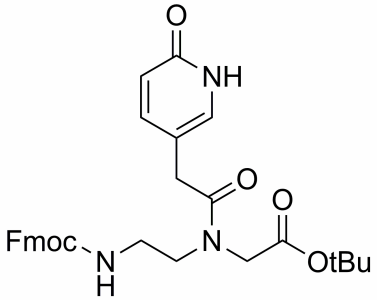
P12: ***m/z*** **(MALDI-TOF MS)**; for $C_{114}H_{146}N_{57}O_{28}$ $(M+H)^+$: calcd 2762.18, found 2762.17; **HPLC** t_R = 12.13 min (method 7); **Yield** = 7 % (79 % per monomer).

6-Oxo-1,6-dihydro-3-pyridylaceti acid (34)[210]



6-chloro-3-pyridylacetic acid (4.0 g, 23 mmol) was suspended in 10 M KOH aq (10 mL) and heated in the microwave at 205 °C for 25 min (18 bar pressure). **Warning**; these conditions cause obvious corrosion of the microwave vial, which should be disposed of after the reaction. The resulting black solution was acidified to pH ~ 1-2 with 2 M HCl aq, and the resulting precipitate was collected by suction filtration, washed with H_2O, and dried *in vacuo* overnight to afford **34** as a brown solid (3.6 g, 23 mmol, quantitative). $\mathbf{R_f}$ = 0.04 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 3066 (w), 2896 (w), 1703 (m), 1605 (s), 1080 (s); **mp** 249 °C decomp.; **^{1}H NMR** (500 MHz, d_6-DMSO) δ_H 11.90 (br s, 2H, NH/CO_2H), 7.33 (dd, 1H, ^4J = 2.5, ^3J = 9.0 Hz, H-4), 7.23 (d, 1 H, ^4J = 2.5 Hz, H-5), 6.28 (d, 1H, ^3J = 9.0 Hz, H-2), 3.33 (s, 2 H, CH_2); **^{13}C NMR** (125.7 MHz, d_6-DMSO) δ_C 172.7 (CO), 161.7 (CO), 143.2 (CH), 134.2 (CH), 119.3 (CH), 111.8 (C), 35.6 (CH_2); ***m/z*** **(ES$^+$)** 152 $(M-H)^-$; **HRMS (EI$^+$)** for $C_7H_7O_2N$ M^+: calcd 153.04314, found 153.04321; **HPLC** t_R = 1.89 min (method 3).

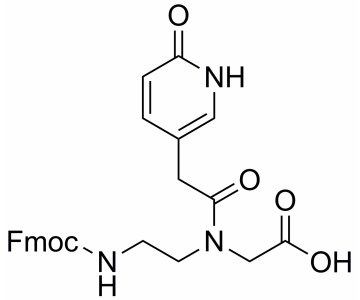
***tert*-Butyl *N*-[2-(*N*-9-Fluorenylmethoxycarbonyl)aminoethyl]-*N*-[(6-oxo-1,6-dihydro-3-pyridyl)acetyl]glycinate (35)**



A suspension of **34** (53 mg, 0.35 mmol), oxyma (49 mg, 0.35 mmol), *N,N'*-dicyclohexylcarbodiimide (71 mg, 0.34 mmol) and diisopropylethylamine (0.05 mL) in DMF (2.3 mL) was heated in the microwave at 60 °C for 2 min, then *tert*-butyl *N*-

[2-(*N*-9-fluorenylmethoxycarbonyl)-aminoethyl]-glycinate hydrochloride (100 mg, 0.23 mmol) was added and the reaction mixture was heated in the microwave at 60 ^{o}C for a further 30 min. The resulting suspension was filtered (washed with DMF) and concentrated to an oily orange solid *in vacuo*. The crude product was purified by flash chromatography (2 × 15 cm SiO_2, eluting with 5 % MeOH:DCM) to afford **35** (91 mg, 0.17 mmol, 74 %) as a white solid. $\mathbf{R_f}$ = 0.11 (5 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2975 (w), 2930 (w), 1715 (m), 1659 (s), 1150 (s); **mp** 101-102 ^{o}C; **^{1}H NMR** (500 MHz, $CDCl_3$) two rotamers: δ_H 12.68 (br s, 1H, NH), 7.75 and 7.74 (d, 2H, ^{3}J = 7.4 Hz, Ar-H), 7.59-7.57 (m, 2 H, Ar-H), 7.38 (t, 2 H, ^{3}J = 7.4 Hz, Ar-H), 7.34 (d, 1 H, ^{3}J = 9.4 Hz, Ar-H), 7.28 (t, 2H, ^{3}J = 7.4 Hz, Ar-H), 7.14 and 7.13 (s, 1H, Ar-H), 6.51 and 6.48 (d, 1H, ^{3}J = 9.4 Hz, Ar-H), 6.03 and 5.69 (br s, 1H, NH), 4.40 and 4.33 (d, 2H, ^{3}J = 7.0 Hz, CH_2O), 4.20 and 4.18 (t, 1H, ^{3}J = 7.0 Hz, $CHCH_2O$), 3.99 and 3.90 (s, 2H, CH_2CO), 3.55-3.48 (m, 2H, NCH_2), 3.39 and 3.29 (s, 2H, CH_2CO), 3.37-3.33 (m, 2H, NCH_2), 1.48 and 1.47 (s, 9 H, CH_3); **^{13}C NMR** (125.7 MHz, $CDCl_3$) two rotamers: δ_C 171.5 and 170.8 (CO), 169.2 and 168.5 (CO), 164.4 and 164.3 (CO), 156.7 and 156.6 (CO), 143.9 and 143.6 (C), 143.8 and 143.7 (CH), 141.2 and 141.2 (C), 133.6 and 133.5 (CH), 127.7 and 127.6 (CH), 127.1 and 127.0 (CH), 125.1 and 125.0 (CH), 120.1 and 120.0 (CH), 119.9 (CH), 113.6 and 113.4 (C), 83.3 and 82.4 (C), 66.9 and 66.7 (CH_2), 51.5 and 49.8 (CH_2), 49.4 and 48.4 (CH_2), 47.2 and 47.1 (CH), 39.3 and 39.2 (CH_2), 35.6 and 35.2 (CH_2), 28.0 (CH_3); ***m/z* (ES$^+$)** 532 $(M+H)^+$, 554 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{30}H_{34}O_6N_3$ $(M+H)^+$: calcd 532.24531, found 532.24410; **HPLC** t_R = 4.32 min (method 3).

_N_-[2-(_N_-9-Fluorenylmethoxycarbonyl)aminoethyl]-_N_-[(6-oxo-1,6-dihydro-3-pyridyl)acetyl]glycine (36)



35 (578 mg, 1.1 mmol) was dissolved in 50 % v/v TFA/DCM (30 mL) and stirred at room temperature for 8 h. The mixture was then concentrated to an oily solid *in vacuo* and purified by column chromatography (4 × 16 cm SiO_2, eluting with 10 % MeOH:DCM containing 0.4 % FA) to afford an off-white solid which was resuspended in water (50 mL) and lyophilized to give **36** as an off-white solid (526 mg, 1.1 mmol, quantitative). $\mathbf{R_f}$ = 0.30 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2945 (w), 1699 (m), 1654 (s), 1607 (s); **mp** 129-130 ^oC; **^{1}H NMR** (500 MHz, CD_3OD) two rotamers: δ_H 7.80 (d, 2H, ^3J = 7.4 Hz, Ar-H), 7.72-7.54 (m, 2 H, Ar-H), 7.46 (d, 1 H, ^3J = 9.5 Hz, Ar-H), 7.40-7.33 (m, 2H, Ar-H), 7.33- 7.23 (m, 3H, Ar-H), 6.47 and 6.46 (d, 1H, ^3J = 9.5 Hz, Ar-H), 4.40 and 4.35 (d, 2H, ^3J = 7.0 Hz, CH_2O), 4.22 and 4.07 (s, 2H, CH_2CO), 4.21- 4.18 (m, 1H, CHCH$_2$O), 3.59 and 3.46 (s, 2H, NCH_2), 3.54-3.48 (m, 2H, NCH_2), 3.35-3.32 (m, 2H, NCH_2); **^{13}C NMR** (125.7 MHz, $CDCl_3$) two rotamers: δ_C 173.9 and 173.4 (CO), 173.0 (CO), 164.9 (CO), 158.9 (CO), 146.2 and 146.0 (C), 145.2 and 145.1 (CH), 142.6 (C), 135.3 and 135.1 (CH), 128.8 and 128.2 (CH), 126.2 and 126.1 (CH), 125.0 (CH), 121.0 (CH), 120.3 and 120.1 (CH), 116.5 and 116.4 (C), 67.8 and 67.7 (CH_2), 51.5 and 49.9 (CH_2), 49.3 and 48.9 (CH_2), 48.4 (CH), 40.0 and 39.6 (CH_2), 36.1 and 35.8 (CH_2); ***m/z* (ES$^+$)** 476 $(M+H)^+$, 951 $(2M+H)^+$; **HRMS (ES$^+$)** for $C_{26}H_{26}O_6N_3$ $(M+H)^+$: calcd 476.18161, found 476.18183; **HPLC** t_R = 4.04 min (method 3).

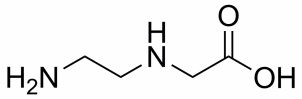
(ii) T_m measurements (Table 4.2)

T_m values were determined as before (Chapter 6.3.2) using five heating-cooling cycles over the range 15-90 ^oC. The reported T_m values were averaged across the latter four heating cycles (the first heating cooling cycle was used to permit full hybridization of the initial PNA/DNA mixture).

6.6 Chapter 5 Experimental

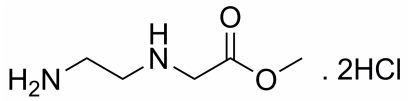
6.6.1 Synthesis of a Thymine PNA Aldehyde (Chapter 5.2)

N-(2-aminoethyl)glycine (37)[83]



1,2-diaminoethane (142 mL, 2.1 mol, 10 eq) was cooled to 10 °C and 2-chloroacetic acid (20.0 g, 0.21 mol) was added portionwise with stirring. After the acid had dissolved, the reaction mixture was warmed to room temperature and stirred for 16 h, then concentrated in vacuo (at 50 °C – higher temperatures lead to decomposition of the target) to a colourless oil. DMSO (500 mL) was added and the mixture was swirled, sonicated and left to stand at room temperature for 1 h. The resulting white precipitate was collected by suction filtration, washed with DMSO (3 × 100 mL) then Et_2O (2 × 50 mL) and dried *in vacuo* at 40 °C to afford the target as a white solid (14.4 g). A second crop of target (3.7 g) was collected by filtration of the DMSO washings which had been left to stand at room temperature overnight. This second crop was washed with DMSO (2 × 50 mL) then Et_2O (2 × 50 mL), dried *in vacuo* at 40 °C and combined with the first crop (total yield 18.1 g, 0.15 mol, 72 %). **^{1}H NMR** (600 MHz, D_2O) δ_H 3.30 (s, 2H, $C\underline{H_2}CO_2H$), 3.06 (t, 2H, ^3J = 6.3 Hz, $NH_2C\underline{H_2}$), 2.93 (t, 2H, ^3J = 6.3 Hz, $C\underline{H_2}NH$); **^{13}C NMR** (125.7 MHz, D_2O) δ_C 178.5 (CO), 51.8 (CH_2), 46.8 (CH_2), 38.8 (CH_2); ***m/z* (ES$^+$)** 119 $(M+H)^+$, 141 $(M+Na)^+$; **HPLC** t_R = 0.99 min (method 2).

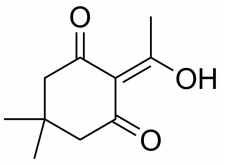
Methyl *N*-(2-aminoethyl)glycinate.2HCl (38)[83]



To a stirred suspension of **37** (13.0 g, 0.11 mol) in MeOH (156 mL) at -5 °C under N_2 (g) was added thionyl chloride (80.5 mL, 1.1 mol, 10 eq) slowly over 5 h (**caution**; this is a violent reaction liberating HCl (g)). The reaction mixture was then refluxed for 15 h before it was cooled to -5 °C. The resulting precipitate was collected by filtration and washed with 1:1 v/v MeOH:DCM then Et_2O, and dried under vacuum at 40 °C overnight to yield **38** as a white solid (20.3 g, 0.10 mol, 90 %). **^{1}H NMR** (500 MHz, D_2O) δ_H 4.12 (s, 2H, CH_2CO), 3.84 (s, 3H, CH_3), 3.55-3.41

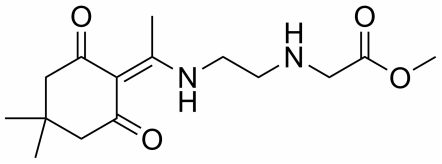
(m, 4H, 2 × NCH_2); **^{13}C NMR** (125.7 MHz, D_2O) δ_C 168.0 (CO), 54.3 (CH_3), 48.3 (CH_2), 44.7 (CH_2), 36.0 (CH_2); ***m/z* (ES$^+$)** 133 $(M+H)^+$, 155 $(M+Na)^+$; **HPLC** t_R = 1.04 min (method 2).

2-Acetyldimedone (39)[83]



To a stirred solution of dimedone (20.0 g, 143 mmol), DCC (29.5 g, 143 mmol, 1 eq) and DMAP (1.74 g, 14.2 mmol, 0.1 eq) in DMF (350 mL) was added glacial acetic acid (8.2 mL, 143 mmol, 1 eq). Stirring was continued for 44 h, then the dicyclohexylurea precipitate was removed by filtration and the filtrate was concentrated *in vacuo* to remove the DMF solvent. EtOAc (150 mL) was added to the residue, which was filtered and washed with 1 M $KHSO_4$ aq (3 × 150 mL), dried ($MgSO_4$), filtered again and concentrated *in vacuo* to afford the crude product as an orange oil which crystallized on standing. Purification by column chromatography (7.5 × 14.5 cm silica, eluting with 5 % EtOAc:DCM) gave **39** as an orange oil (19.1 g, 105 mmol, 73 %). **R_f** 0.67 (5 % EtOAC:DCM); **^{1}H NMR** (250 MHz, $CDCl_3$) δ_H 2.60 (s, 3H, C=C(OH)C$\underline{H_3}$), 2.53 (s, 2H, CH_2CO), 2.35 (s, 2H, CH_2CO), 1.07 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (125.7 MHz, $CDCl_3$) δ_C 202.1 (C), 197.6 (CO), 194.8 (CO), 112.0 (C), 52.1 (CH_2), 46.6 (CH_2), 30.3 (C), 28.2 (CH_3), 27.9 (CH_3); ***m/z* (ES$^+$)** 183 $(M+H)^+$, 205 $(M+Na)^+$; **HPLC** t_R = 3.30 min (method 1).

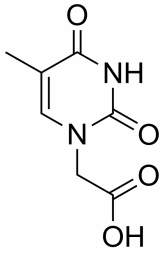
Methyl *N*-(2-[1-(4,4-dimethyl-2,6-dioxocyclohexylidene)-ethylamino]-ethyl)-glycinate (40)[83]



To a stirred solution of **38** (11.5 g, 56 mmol) and DiPEA (20.5 mL, 118 mmol, 2.1 eq) in DCM (200 mL) was added a solution of crude **39** (10.2 g, 56 mmol, 1 eq). Some of **38** remained undissolved, so EtOH (20 mL) was added and the reaction mixture was sonicated to aid dissolution. The mixture was then left to stir at room

temperature for 16 h, before it was filtered (washed with DCM) and concentrated *in vacuo* to an orange oil. This was redissolved in EtOAc (150 mL) and washed with 1 M $KHSO_4$ aq (4 × 25 mL). The aqueous washes were combined, brought to pH 9 with saturated $NaHCO_3$ aq, then back-extracted with EtOAc (4 × 50 mL). The combined organics were washed with brine (50 mL), dried ($MgSO_4$), filtered and concentrated *in vacuo* to afford the crude product as an orange oil. Purification by column chromatography (7 × 16 cm silica, eluting with 0 → 5 % MeOH:DCM) gave **40** as an orange/brown oil (2.54 g, 8 mmol, 15 %). $\mathbf{R_f}$ 0.42 (10 % MeOH:DCM); **^{1}H NMR** (600 MHz, $CDCl_3$) δ_H 13.50 (br s, 1H, Dde-N<u>H</u>), 3.73 (s, 3H, OCH_3), 3.48-3.47 (m, 4H, 2 × NCH_2), 2.94 (t, 2 H, ^3J = 6.0 Hz, CH_2), 2.57 (s, 3H, C=C(NH)C<u>H</u>$_3$), 2.36 (br s, 4H, 2 × CH_2CO), 1.68 (br s, 1H, N<u>H</u>CH_2CO), 1.02 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (125.7 MHz, $CDCl_3$) δ_C 197.1 (CO), 173.3 (C), 172.6 (CO), 107.9 (C), 52.7 (CH_2), 51.8 (CH_3), 50.1 (CH_2), 47.5 (CH_2), 43.3 (CH_2), 29.9 (C), 28.1 (CH_3), 18.0 (CH_3); ***m/z* (ES$^+$)** 297 $(M+H)^+$, 319 $(M+Na)^+$; **HPLC** t_R = 2.33 min (method 1).

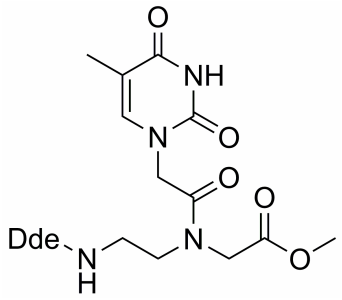
Thymin-1-yl acetic acid (41)[83]



To a stirred suspension of thymine (10 g, 79 mmol) in DMF (240 mL) was added K_2CO_3 (11 g, 80 mmol) followed by methyl 2-bromoacetate (10 mL, 106 mmol). The reaction mixture was left to stir vigorously at room temperature under N_2 (g) for 16 h, and the resulting suspension was filtered and concentrated *in vacuo* to afford a white solid. Cold H_2O at ~ 0 ^oC was added, followed by 1 M $KHSO_4$ aq (5 mL). The mixture was stirred with cooling on ice for 40 min, then the white precipitate was collected by filtration. This was dissolved in a mixture of H_2O (100 mL) and 2 M NaOH aq (50 mL), then refluxed for 3 h. The reaction mixture was again cooled to ~ 0 ^oC, acidified with 1 M $KHSO_4$ aq (120 mL) and stirred at 0 ^oC for 30 min. The resulting precipitate was collected by filtration, washed with water (3 × 50 mL) and dried under vacuum overnight to afford **41** as a white solid (8.3 g, 45 mmol, 57 %).

^{1}H NMR (500 MHz, d_6-DMSO) δ_H 13.14 (br s, 1H, CO_2H), 11.34 (s, 1H, NH), 7.49 (s, 1H, C=CH), 4.36 (s, 2H, CH_2), 1.75 (s, 3 H, CH_3); **^{13}C NMR** (125.7 MHz, d_6-DMSO) δ_C 169.6 (CO), 164.3 (CO), 151.0 (CO), 141.8 (CH), 108.3 (C), 48.4 (CH_2), 11.9 (CH_3); ***m/z* (ES$^+$)** 183 $(M-H)^-$; **HPLC** t_R = 4.11 min (method 2).

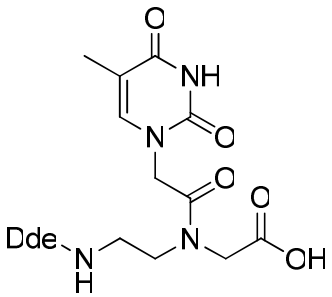
Methyl *N*-[2-(thymin-1-yl)-acetyl]-*N*-(2-[1-(4,4-dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-glycinate (42)[83]



To a stirred solution of **41** (1.74 g, 9.4 mmol, 1.1 eq) in DMF (50 mL) at room temperature was added HOBt.H_2O (1.16 g, 8.6 mmol). Stirring was continued for 10 min, before DCC (1.77 g, 8.6 mmol) was added followed by **40** (2.54 g, 8.6 mmol, 1 eq) as a solution in DMF (40 mL). Stirring was continued for 16 h, then the reaction mixture was filtered and concentrated to dryness *in vacuo* to an oily orange/brown solid. This was redissolved in DCM (270 mL), then washed with 1 M $KHSO_4$ aq (90 mL), 1 M $NaHCO_3$ aq (90 mL) and brine (90 mL), then the organic phase was dried ($MgSO_4$), filtered and concentrated *in vacuo* to afford a sticky off-white solid. This was redissolved in Et_2O (~ 50 mL) and concentrated *in vacuo* to yield an off-white solid which was triturated with 10 % MeOH:EtOAc, collected by filtration and dried *in vacuo* at 40 ^oC overnight to afford **42** as a white solid (2.36 g, 5.1 mmol, 59 %). **^{1}H NMR** (600 MHz, d_6-acetone) two rotamers: δ_H 13.57 and 13.42 (br s, 1H, Dde-N<u>H</u>), 9.98 (br s, 1H, OCNHCO), 7.37 and 7.31 (s, 1H, C=CH), 4.78 and 4.62 (s, 2H, C<u>H$_2$</u>CO_2Me), 4.46 and 4.20 (s, 2H, CH_2C(O)N), 3.90 and 3.70 (s, 4H, 2 × CH_2N), 3.76 and 3.66 (s, 3H, OCH_3), 2.61 and 2.55 (s, 3H, C=C(NH)C<u>H$_3$</u>), 2.34 and 2.30 (br s, 4H, Dde-C<u>H$_2$</u>), 1.83 and 1.82 (s, 3H, C=C(CO)CH_3), 1.00 and 0.99 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (125.7 MHz, d_6-acetone) two rotamers: δ_C 175.5 and 175.2 (CO), 171.6 (CO), 171.3 (CO), 170.1 and 169.4 (CO), 165.8 (C), 153.0 and 152.9 (CO), 143.6 (CH), 110.8 and 110.7 (C), 109.6 and 109.5 (C), 53.8 and 53.3 (CH_3), 51.1 (CH_2), 49.7 and 49.6 (CH_2), 49.5 (CH_2), 49.1 and 48.9 (CH_2), 42.3 and 42.1

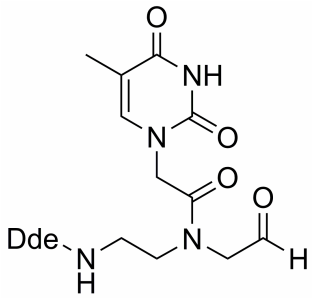
(CH_2), 35.5 (C), 27.5 and 26.7 (CH_3), 18.8 and 18.6 (CH_3), 13.4 (CH_3); ***m/z* (ES$^+$)** 463 $(M+H)^+$, 485 $(M+Na)^+$; **HPLC** t_R = 2.51 min (method 1).

***N*-[2-(thymin-1-yl)-acetyl]-*N*-(2-[1-(4,4-dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-glycine (43)[83]**



To **42** (750 mg, 1.6 mmol) was added MeOH (7.5 mL) and 2 M Cs_2CO_3 aq (7.5 mL). The suspension was stirred at room temperature for 4.5 h, then the MeOH was removed *in vacuo* and the resulting slurry was acidified to pH 1 with 2 M HCl aq The solvents were then removed *in vacuo* to afford a white solid which was sonicated with H_2O (5 mL), filtered and dried overnight *in vacuo* at 40 ^oC to give **43** as a white solid (486 mg, 1.1 mmol, 67 %). $\mathbf{R_f}$ 0.13 (10 % MeOH:DCM + 1 % AcOH); **^{1}H NMR** (500 MHz, d_6-acetone) two rotamers: δ_H 13.56 and 13.40 (br s, 1H, Dde-NH), 9.96 (br s, 1H, OCNHCO), 7.37 and 7.32 (s, 1H, C=CH), 4.78 and 4.64 (s, 2H, CH_2CO_2Me), 4.38 and 4.20 (s, 2H, $CH_2C(O)N$), 3.90 and 3.70 (s, 4H, 2 × CH_2N), 2.61 and 2.56 (s, 3H, C=C(NH)CH_3), 2.33 and 2.30 (br s, 4H, Dde-CH_2), 1.83 and 1.82 (s, 3H, C=C(CO) CH_3) and 1.00 and 0.99 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (125.7 MHz, d_6-acetone) two rotamers: δ_C196.9 and 196.7 (CO), 173.6 (CO), 171.2 and 170.8 (CO), 168.6 and 167.8 (CO), 164.8 (C), 151.5 and 151.4 (CO), 142.5 and 142.4 (CH), 108.6 (C), 107.9 and 107.7 (C), 52.9 (2 x CH_2), 49.4 (CH_2), 48.1 and 46.9 (CH_2), 41.4 (CH_2), 30.2 (C), 28.3 (CH_3), 17.8 and 17.7 (CH_3), 12.5 and 12.4 (CH_3); ***m/z* (ES$^+$)** 449 $(M+H)^+$, 471 $(M+Na)^+$; **HPLC** t_R = 2.46 min (method 1).

_N_-(2-[1-(4,4-Dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-2-(5-methyl-2,4-dioxo-3,4-dihydro-2_H_-pyrimidin-1-yl)-_N_-(2-oxo-ethyl)-acetamide (44)



By reduction of Weinreb amide 45

To a stirred solution of **45** (50 mg, 0.10 mmol) in dry THF (3 mL) at room temperature under N_2 (g) was added a 1.0 M solution of LiAlH(O-*t*-Bu)$_3$ in THF (0.3 mL, 0.30 mmol, 3 eq). After stirring for 95 min, the reaction mixture was added dropwise to stirred 5 % w/w aq $KHSO_4$ (1.2 mL), then extracted with EtOAc (4 × 8 mL). The organic extracts were combined and washed with saturated $NaHCO_3$ aq (3 × 5 mL), brine (3 × 3 mL) then dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford crude **44** as a white solid (22 mg). HPLC shows approximately 65 % conversion of **44** to aldehyde **45**. An analytically pure sample of **44** (3.8 mg) was obtained by preparative HPLC (method 3).[†] $\mathbf{R_f}$ = 0.32 (10 % MeOH:DCM, spot turns orange using 2,4-dinitrophenylhydrazine dip); **^{1}H NMR** (600 MHz, d_6-acetone) two rotamers: δ_H 13.62 and 13.44 (br s, 1H, Dde-NH), 9.94 (br s, 1H, OCNHCO), 9.72 and 9.49 (s, 1H, CHO), 7.39 and 7.32 (s, 1H, C=CH), 4.84 and 4.69 (s, 2H, CH_2-CHO), 4.52 and 4.23 (s, 2H, CH_2CO), 3.95-3.85 (m, 2H, CH_2N), 3.67 (br s, 2H, CH_2N), 2.61 and 2.55 (s, 3H, C=C(NH)CH_3), 2.42-2.26 (m, 4 H, Dde-CH_2), 1.82 (s, 3H, C=C(CO)CH_3) and 1.00 and 0.99 (s, 6H, C(CH_3)$_2$); ***m/z* (ES$^+$)** 433 $(M+H)^+$, 455 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{21}H_{29}O_6N_4$ $(M+H)^+$: calcd 433.20816, found 433.20713; **HPLC** t_R = 2.36 min (method 1).

By reduction of S-benzyl thioester 46

To a stirred suspension of **46** (50 mg, 0.09 mmol) and 10 wt % Pd/C (24 mg, 25 mol % Pd) in anhydrous THF (0.75 mL) under N_2 (g) was added triethylsilane (6.0 eq, 0.54 mmol, 86 μL) and acetone (dried over molecular sieves, 43 μL). A further portion of 10 wt % Pd/C was added after 2h 10 min (72 mg, 75 mol % Pd), and

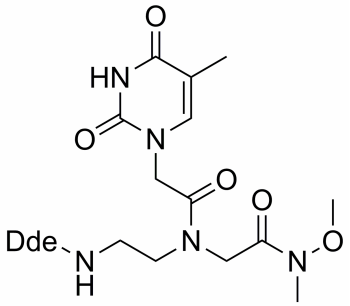
[†] Only partial characterization was achieved as a result of the poor stability of this compound and the small amount isolated. Preparative HPLC peformed by Dr Ivan Trkulja.

further triethylsilane (86 μL, 0.54 mmol, 6 eq) was added after another 2h. Stirring was continued for 16 h, before the reaction mixture was filtered through celite and concentrated *in vacuo* to afford an oily yellow solid which was triturated with Et_2O (2 × 15 mL) and filtered to afford crude **44** as a pale brown solid (20 mg). R_f and ^1H NMR as for Method A above. **HPLC** t_R = 2.36 min (method 1), approximate conversion = 83 %.

By acetal deprotection of 50

50 (12 mg, 24 μmol) was stirred in 90 % TFA in H_2O (0.5 mL) for 1 h at a temperature maintained between 0 and -10 ^oC (dry ice/acetone bath), then blown down with N_2 (g) to a viscous oil which was basified with saturated $NaHCO_3$ aq (1.5 mL) at – 5 ^oC and extracted with EtOAc (10 mL then 2 × 5 mL). The combined organics were dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford crude **44** as a white solid (11 mg). **HPLC** t_R = 2.35 min (method 1), approximate conversion = 90 %.

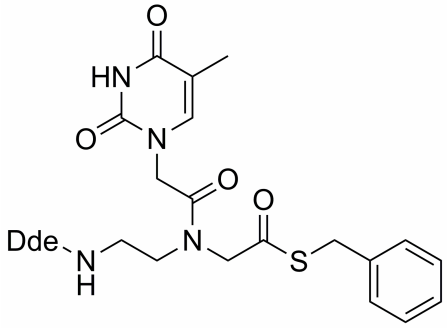
_N_-(2-[1-(4,4-Dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-_N_-[(methoxy-methyl-carbam)-methyl]-2-(5-methyl-2,4-dioxo-3,4-dihydro-2_H_-pyrimidin-1-yl)-acetamide (45)



To a stirred solution of **43** (250 mg, 0.56 mmol), HOBt.H_2O (1.6 eq, 0.90 mmol, 121 mg) and *N, O*-dimethylhydroxylamine.HCl (87 mg, 0.90 mmol, 1.6 eq) in DMF (5 mL) was added triethylamine (289 μL, 2.07 mmol, 3.7 eq), followed by EDC.HCl (215 mg, 1.12 mmol, 2.0 eq) with DMF washings (3.5 mL). The reaction mixture was left to stir at room temperature for 20 h, before it was filtered and concentrated *in vacuo* to afford a yellow oil. This was dissolved in EtOAc (50 mL), then washed with 1 M $KHSO_4$ aq (10 mL). The aqueous phase was back-extracted with EtOAc (2 × 10 mL), and the combined organics were washed with saturated $NaHCO_3$ aq (10

mL), brine (10 mL) and dried (Na_2SO_4), then filtered and concentrated *in vacuo* to afford an oily solid. This was resuspended in Et_2O (30 mL) and concentrated *in vacuo* to afford **45** as a white solid (251 mg, 92 %). The crude product was used in subsequent reactions without purification. An analytical sample (20 mg) was obtained by column chromatography (2 × 14 cm silica, eluting with 10 % MeOH:DCM). $\mathbf{R_f}$ = 0.33 (10 % v/v MeOH:DCM); **mp** 198-200 ^oC; **IR** ν_{max}/ cm^{-1} (neat) 2930 (w), 1668 (s), 1566 (s), 1462 (m), 725 (m); **^{1}H NMR** (500 MHz, d_6-acetone) two rotamers: δ_H 13.55 and 13.44 (br s, 1H, Dde-NH), 9.94 (br s, 1H, OCNHCO), 7.37 and 7.28 (s, 1H, C=CH), 4.77 and 4.60 (s, 2H, CH_2-CON(OMe)Me), 4.55 and 4.38 (s, 2H, CH_2CO), 3.88-3.86 (m, 2H, CH_2N), 3.81 and 3.75 (s, 3H, $NOCH_3$), 3.69-3.65 (m, 2H, CH_2N), 3.20 and 3.12 (s, 3H, $N(OMe)CH_3$), 2.61 and 2.55 (s, 3H, $C=C(NH)CH_3$), 2.33 and 2.30 (br s, 4H, Dde-CH_2), 1.83 (s, 3H, $C=C(CO)CH_3$), 1.00 and 0.99 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (90.6 MHz, $CDCl_3$) two rotamers: δ_C 207.4 (CO), 198.2 (CO), 174.3 and 173.7 (CO), 168.7 and 168.6 (CO), 167.4 (C), 164.2 and 164.1 (CO), 151.2 (CO), 141.4 and 141.1 (CH), 110.9 and 110.7 (C), 108.6 and 108.4 (C), 61.9 and 61.6 (CH_3), 53.6 and 53.0 (CH_2), 50.5 (CH_2), 48.9 (CH_2), 48.5 and 48.4 (CH_2), 47.7 and 47.6 (CH_2), 41.5 and 41.2 (CH_2), 32.8 (C), 30.2 (CH_3), 28.4 (CH_3), 18.1 and 18.0 (CH_3), 12.5 (CH_3); ***m/z* (ES$^+$)** 492 $(M+H)^+$, 514 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{23}H_{34}O_7N_5$ $(M+H)^+$: calcd 492.24527, found 492.24425; **HPLC** t_R = 2.57 min (method 1).

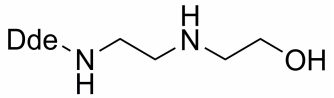
((2-[1-(4,4-Dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-[2-(5-methyl-2,4-dioxo-3,4-dihydro-2*H*-pyrimidin-1-yl)-acetyl]-amino)-thioacetic acid *S*-benzyl ester (46)



To a stirred suspension of **43** (368 mg, 0.8 mmol) and benzyl mercaptan (194 µL 1.6 mmol, 2 eq) in anhydrous DMF (9 mL) under N_2 (g) at room temperature was added

DMAP (11 mg, 0.08 mmol, 0.1 eq) followed by DCC (186 mg, 0.90 mmol, 1.1 eq). The reaction mixture was stirred at room temperature under N_2 (g) for 22 h, before it was filtered and concentrated to dryness *in vacuo* to afford an oily solid. This was suspended in EtOAC (45 mL) and washed with 1 M $KHSO_4$ aq (2 × 7 mL). The combined aqueous phases were back-extracted with EtOAc (2 × 45 mL), then the combined organics were washed with brine (14 mL), dried (Na_2SO_4), filtered and concentrated *in vacuo* to give a glassy solid. This was sonicated with Et_2O (10 mL) and decanted from the solvent to give the crude target as a white solid. Purification by column chromatography (3 × 15 cm silica, 2 → 5 % MeOH:EtOAc, crude material pre-adsorbed onto silica from DCM) afforded **46** as a white solid (306 mg, 0.55 mmol, 67 %). $\mathbf{R_f}$ = 0.27 (5 % MeOH:EtOAc); **mp** 156-158 ^oC; **IR** ν_{max}/ cm^{-1} (neat) 2953 (w), 1669 (s), 1565 (s), 1455 (m), 1332 (m); **^{1}H NMR** (500 MHz, d_6-acetone) two rotamers: δ_H 13.58 and 13.42 (br s, 1H, Dde-NH), 9.98 (br s, 1H, OCNHCO), 7.37-7.24 (m, 6H, CH and Ar-H), 4.84 and 4.68 (s, 2H, CH_2COS), 4.64 and 4.44 (s, 2H, CH_2CO), 4.25 and 4.16 (s, 2H, CH_2Ar), 3.95-3.89 (m, 2H, CH_2N), 3.75-3.68 (m, 2H, CH_2N), 2.59 and 2.54 (s, 3H, C=C(NH)CH_3), 2.30 (br s, 4 H, Dde-CH_2), 1.82 (s, 3H, C=C(CO)CH_3), 1.00 and 0.99 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (62.9 MHz, $CDCl_3$) two rotamers: δ_C 198.1 (CO), 195.5 and 194.7 (CO), 173.9 and 173.8 (CO), 168.0 (CO), 167.3 (C), 164.4 (CO), 151.3 and 151.2 (CO), , 141.1 and 140.7 (CH), 136.4 and 136.1 (C), 128.7 (CH), 128.5 (CH), 127.6 and 127.4 (CH), 110.7 and 110.6 (C), 108.2 and 108.1 (C), 57.6 (CH_2), 55.6 (CH_2) 52.7 (CH_2), 48.0 and 47.8 (CH_2), 47.6 and 47.5 (CH_2), 41.1 and 40.5 (CH_2), 33.3 and 33.0 (CH_2), 29.9 (C), 28.1 (CH_3), 17.8 and 17.6 (CH_3), 12.2 (CH_3); ***m/z*** **(ES$^+$)** 555 $(M+H)^+$, 577 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{28}H_{35}O_6N_4S$ $(M+H)^+$: calcd 555.22775, found 555.22718; **HPLC** (λ_{254}) t_R = 3.24 min (method 1).

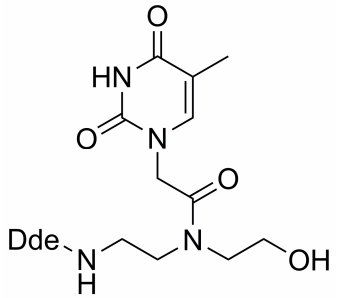
2-(1-[2-(2-Hydroxy-ethylamino]-ethylidene)-5,5-dimethyl-cyclohexane-1,3-dione (47)



To a stirred solution of 2-(2-aminoethylamino)ethanol (2.4 mL, 24 mmol) in DCM (30 mL) under N_2 (g) was added a solution of Dde-OH (**39**, 4.4 g, 24 mmol, 1 eq) in

DCM (10 mL + 20 mL washings). Stirring was continued at room temperature for 16 h. The reaction mixture was then separated from the layer of water which had formed, and concentrated *in vacuo* to afford a yellow oil which solidified on standing. Purification by column chromatography (7 × 16 cm silica, eluting with 10 % MeOH:DCM) afforded **47** as a yellow solid (6.4 g, 24 mmol, 100 %). $\mathbf{R_f}$ = 0.36 (10 % MeOH:EtOAc); **IR** ν_{max}/ cm^{-1} (neat) 3316 (w), 2924 (w), 2867 (w), 1623 (m), 1565 (s); **^{1}H NMR** (500 MHz, $CDCl_3$) δ_H 3.66 (t, 2H, ^3J = 5.0 Hz, $C\underline{H_2}OH$), 3.50-3.46 (m, 2H, Dde-NHC$\underline{H_2}$), 2.97 (t, 2H, ^3J = 6.0 Hz, Dde-$NHCH_2C\underline{H_2}$), 2.85 (t, 2H, ^3J = 5.0 Hz, $C\underline{H_2}CH_2OH$), 2.57 (s, 3H, CH_3), 2.36 (br s, 4 H, 2 × Dde-C$\underline{H_2}$), 1.02 (s, 6H, $C(CH_3)_2$); **^{13}C NMR** (90.6 MHz, $CDCl_3$) two rotamers: δ_C 197.9 (CO), 173.2 (C), 108.0 (C), 61.0 (CH_2), 52.8 (CH_2), 50.9 (CH_2), 47.2 (CH_2), 43.2 (CH_2), 30.0 (C), 28.2 (CH_3), 18.2 (CH_3); ***m/z*** **(ES$^+$)** 269 $(M+H)^+$, 291 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{14}H_{25}O_3N_2$ $(M+H)^+$: calcd 269.18550, found 269.18597; **HPLC** t_R = 6.01 min (method 2).

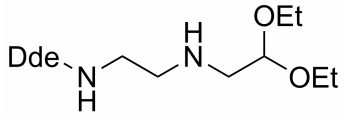
_N_-(2-[1-(4,4-Dimethyl-2,6-dioxo-cyclohexylidene)-ethylamino]-ethyl)-_N_-(2-hydroxy-ethyl)-2-(5-methyl-2,4-dioxo-3,4-dihydro-2_H_-pyrimidin-1-yl)-acetamide (48)



To a stirred suspension of **47** (250 mg, 0.93 mmol) and **41** (172 mg, 0.93 mmol, 1 eq) in DCM (2.1 mL) at -5 ^oC under N_2 (g) was added DiPEA (324 µL, 1.86 mmol, 2 eq) at -5 ^oC and PyBOP (434 mg, 0.93 mmol, 1 eq). Cooling was removed and stirring was continued at room temperature under N_2 (g) for 2 h. The reaction mixture was then washed with 1 M $KHSO_4$ aq (5 mL), 1 M $NaHCO_3$ aq (5 mL), and brine (5 mL) before it was dried ($MgSO_4$), filtered and concentrated *in vacuo* to afford an off-white solid. Purification by column chromatography (3 × 16 cm silica, eluting with 10 % MeOH:DCM) afforded **48** as a white solid (60 mg, 0.14 mmol, 15 %). $\mathbf{R_f}$ = 0.33 (10 % MeOH:DCM); **mp** 168-169 ^oC; **IR** ν_{max}/ cm^{-1} (neat) 3325 (w),

3236 (w), 2930 (w), 1704 (m), 1687 (s), 1565 (s); **^{1}H NMR** (500 MHz, d_6-acetone) two rotamers: δ_H 13.55 and 13.42 (br s, 1H, Dde-NH), 9.94 (br s, 1H, OCNHCO), 7.36 and 7.31 (s, 2H, C=CH), 4.78 and 4.72 (s, 2H, CH_2CO), 3.93-3.54 (m, 8H, 4 × CH_2), 2.61 and 2.55 (s, 3H, C=C(NH)CH_3), 2.30 (br s, 4 H, Dde-CH_2), 1.82 (s, 3H, C=C(CO)CH_3), 0.99 (s, 6H, C(CH_3)$_2$); **^{13}C NMR** (90.6 MHz, acetone-d_6) two rotamers: δ_C 180.7 (CO), 156.9 and 155.3 (CO), 151.2 (C), 147.8 (CO), 124.7 and 124.6 (CH), 91.6 (C), 89.9 (C), 41.4 (CH_2), 34.3 (CH_2), 31.9 (CH_2), 30.8 (CH_2), 27.6 (CH_2), 22.5 (CH_2), 11.9 (C), 9.2 (CH_3), -0.9 (CH_3) and -6.8 (CH_3); ***m/z* (ES$^+$)** 435 $(M+H)^+$, 457 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{21}H_{31}O_6N_4$ $(M+H)^+$: calcd 435.22381, found 435.22349; **HPLC** t_R = 2.40 min (method 1).

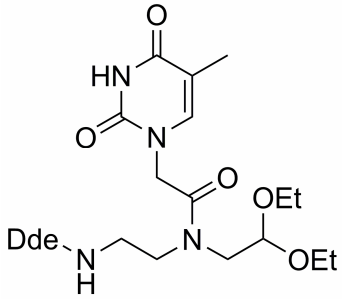
2-(1-[2-(2,2-Diethoxy-ethylamino]-ethylidene)-5,5-dimethyl-cyclohexane-1,3-dione (49)



A mixture of 1,2-diaminoethane (0.74 mL, 11 mmol), bromoacetaldehyde diethyl acetal (1.7 mL, 11 mmol, 1 eq) and K_2CO_3 (3 g, 22 mmol, 2 eq) in MeCN (15 mL) was heated at 130 ^oC in the microwave (pressure = 5 bar) for 30 min. This reaction was repeated × 5. The reaction mixtures were then combined and filtered, washing with DCM. The filtrate was concentrated *in vacuo* to a brown oil which was resuspended in DCM (70 mL) and a white precipitate was removed by filtration, washing with more DCM (30 mL). To the resulting dark brown solution was added a solution of Dde-OH (**39**, 10 g, 55 mmol, 1 eq) in DCM (50 mL), and the reaction mixture was stirred at room temperature under N_2 (g) for 16 h. The mixture was then concentrated *in vacuo* to a dark brown oil which was purified by column chromatography (7.5 × 14 cm silica, eluting with 5 → 10 % MeOH:DCM) to afford **49** as an orange/brown oil (5.6 g, 16 mmol, 30 %). $\mathbf{R_f}$ = 0.41 (10 % MeOH:DCM); **IR** ν_{max}/ cm^{-1} (neat) 2955 (w), 2868 (w), 1636 (m), 1570 (s), 1457 (m); **^{1}H NMR** (500 MHz, $CDCl_3$) δ_H 13.39 (br s, 1H, Dde-NH), 4.52 (t, 1H, ^3J = 5.5 Hz, CH(OEt)$_2$), 3.66-3.60 (m, 2H, OCH_2), 3.50-3.43 (m, 2H, OCH_2), 3.41 (q, 2H, ^3J = 6.0 Hz, Dde-NHCH_2), 2.87 (t, 2H, ^3J = 6.0 Hz, CH_2NHCH_2CH(OEt)$_2$), 2.70 (d, 2H, ^3J = 5.5 Hz, CH_2CH(OEt)$_2$), 2.50 (s, 3H, C=C(NH)CH_3), 2.28 (br s, 4H, 2 × Dde-

C<u>H</u>$_2$), 1.13 (t, 6H, ^3J = 7.0 Hz, 2 × OCH$_2$C<u>H</u>$_3$), 0.95 (s, 6H, C(CH$_3$)$_2$); **^{13}C NMR** (62.9 MHz, $CDCl_3$) δ_C 196.8 (CO), 173.4 (C), 107.9 (C), 102.0 (CH), 62.6 (CH_2), 52.8 (CH_2), 51.8 (CH_2), 48.0 (CH_2), 43.5 (CH_2), 30.0 (C), 28.2 (CH_2), 18.1 (CH_3), 15.4 (CH_3); ***m/z* (EI)** 340 $(M)^+$; **HRMS (EI)** for $C_{18}H_{32}O_4N_2$ $(M)^+$: calcd 340.23566, found 340.23593; **HPLC** t_R = 0.69 min (method 4).

_N_-(2,2-Diethoxy-ethyl)-_N_-(2-[1-(4,4-dimethyl-2,6-dioxo-cyclohexylidene)-ethyl-amino]-ethyl)-2-(5-methyl-2,4-dioxo-3,4-dihydro-2_H_-pyrimidin-1-yl)-acetamide (50)

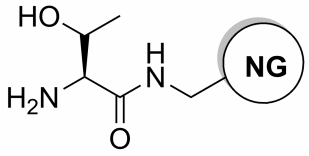


To a solution of **41** (381 mg, 2.1 mmol, 1.1 eq) in DMF (10 mL) under N_2 (g) was added HOBt.H_2O (254 mg, 1.9 mmol, 1 eq) followed by DCC (388 mg, 1.9 mmol, 1 eq) and a solution of **49** (640 mg, 1.9 mmol, 1 eq) in DMF (5 mL). The reaction mixture was heated at 60 ^oC in the microwave for 30 min, then concentrated *in vacuo* to afford a brown oil which was resuspended in Et_2O and again concentrated *in vacuo* to afford an oily off-white solid. This was redissolved in DCM (60 mL) and washed with saturated $NaHCO_3$ aq (20 mL) and brine (20 mL), then dried (Na_2SO_4), filtered and concentrated *in vacuo* to afford the crude product as an off-white solid. Purification by column chromatography (4.5 × 14 cm silica, eluting with 5 % MeOH:DCM, crude product dissolved in MeOH/DCM and pre-adsorbed on silica by evaporation) afforded **50** as an off-white solid (443 mg, 0.87 mmol, 46 %). $\mathbf{R_f}$ = 0.61 (10 % MeOH:DCM); **mp** 178-181 ^oC; **IR** ν_{max}/ cm^{-1} (neat) 3153 (w), 2974 (w), 1695 (s), 1659 (s), 1562 (s); **^{1}H NMR** (500 MHz, $CDCl_3$) two rotamers: δ_H 13.62 and 13.46 (br s, 1H, Dde-N<u>H</u>), 8.08 (br s, 1H, OCNHCO), 7.32 and 7.05 (s, 1H, C=CH), 4.72 and 4.61 (t, 1H, ^3J = 5.5 Hz, C<u>H</u>(OEt)$_2$), 3.81-3.52 (m, 8H, 2 ×OC<u>H</u>$_2$ and 2 × NC<u>H</u>$_2$), 3.49 and 3.41 (d, 2H, ^3J = 5.0 Hz, C<u>H</u>$_2$CH(OEt)$_2$), 2.60 and 2.57 (s, 3H, C=C(NH)C<u>H</u>$_3$), 2.36 (br s, 4H, 2 × Dde-C<u>H</u>$_2$), 1.94 and 1.93 (s, 3H, C=C(CO)CH_3), 1.25 and 1.17 (t, 6H, ^3J = 7.0 Hz, 2 × OCH$_2$C<u>H</u>$_3$), 1.03 (s, 6H, C(CH$_3$)$_2$); **^{13}C NMR**

(62.9 MHz, $CDCl_3$) two rotamers: δ_C 198.3 (CO), 174.2 and 173.8 (CO), 168.3 and 167.3 (C), 164.4 (CO), 151.4 and 151.3 (CO), 141.4 and 141.2 (CH), 110.7 and 110.6 (C), 108.3 (C), 101.2 and 100.6 (CH), 64.6 (CH_2), 53.0 (CH_2), 51.5 and 50.4 (CH_2), 48.4 (CH_2), 47.9 and 47.6 (CH_2), 47.3 (CH_2), 41.0 and 40.4 (CH_2), 31.0 (CH_3), 30.1 (C), 28.3 (CH_2), 18.1 and 18.0 (CH_3), 15.4 (CH_3), 12.5 (CH_3); ***m/z* (ES$^+$)** 507 $(M+H)^+$, 529 $(M+Na)^+$; **HRMS (ES$^+$)** for $C_{25}H_{39}O_7N_4$ $(M+H)^+$: calcd 507.28133, found 507.28072; **HPLC** t_R = 2.95 min (method 3).

6.6.2 Resin Capture of a Thymine PNA Aldehyde (Chapter 5.3)

NovaGel HL-based threonyl resin[188]



Aminomethyl NovaGel HL (200 mg, 0.76 mmol/g, 0.15 mmol) was swollen with DMF over 10 min. Meanwhile, DIPEA (132 µL, 0.76 mmol, 5 eq) was added to a solution of Fmoc-Thr(*t*-Bu)-OH (32 mg, 0.76 mmol, 5 eq) and TBTU (239 mg, 0.74 mmol, 4.9 eq) in DMF (1 mL), and the reaction mixture was shaken for 5 min. The solution of activated protected threonine was then added to the swollen resin and shaken at room temperature for 2h. The resin was washed with DMF (5 ×), THF (5 ×) and DCM (5 ×), then dried *in vacuo* at 40 ^oC overnight. Completion of the coupling was verified using a qualitative ninhydrin test. The resin was then swollen in DMF for 10 min and shaken with 20 % v/v piperidine in DMF (2 mL) for 40 min (repeated once for 20 min). The resin was washed with DMF (3 ×), THF (5 ×) and DCM (5 ×) then shaken with 80 % v/v TFA in DCM (2.5 mL) for 30 min (repeated once). The resin was washed with DCM (5 ×) and dried *in vacuo* at 40 ^oC. Assuming quantitative conversion, new loading N_L = 0.71 mmol/g.

Purification of 3-phenylpropionaldehyde by capture-release

NovaGel HL-based threonyl scavenging resin (150 mg, 85 µmol based on 0.71 mmol/g loading) was swollen with DMF for 30 min then heated at 60 ^oC in the microwave for 1 h with an equimolar mixture of 3-phenylpropionaldehyde (6.6 µL, 50 µmol) and 3-phenylpropan-1-ol (6.8 µL, 50 µmol) in MeOH (1.5 mL) and DiPEA

(10 μL). HPLC (method 1) of the resulting supernatant showed the presence of 3-phenylpropan-1-ol (t_R = 2.98 min) only and no 3-phenylpropionaldehyde (t_R = 3.20 min). The resin was washed with MeOH (5 ×) and stored in a refrigerator for 16 h. The resin was then swollen with DMF for 20 min, washed with DCM (5 ×) and MeOH (5 ×) and heated at 60 ^oC in the microwave for 30 min with 60:40:1 v/v/v MeCN:H_2O:TFA (1 mL, repeated three times). HPLC (method 1) of the cleavage cocktail showed the presence of 3-phenylpropionaldehyde with *trace* impurities.

Purification of aldehyde 44 by capture-release

Capture[188]

To the deprotected NovaGel HL-based threonyl scavenging resin (120 mg, 85 μmol based on 0.71 mmol/g loading) was added a solution of crude aldehyde **2** (18 mg, ~ 27 μmol based on 65 % purity) in a mixture of anhydrous MeOH/DCM/DMF/AcOH (86/9/6/1 v/v, 2 mL). The mixture was shaken at room temperature for 1 h, then the resin was filtered and washed with DCM (3 ×).

Release[188]

The resin was shaken with a mixture of AcOH/H_2O/DCM/MeOH (10/5/5/80 v/v, 2 mL) for 20 min (repeated once for 1 h). The resin was then washed with DMF (5 ×), DCM (5 ×) and Et_2O (5 ×) and dried *in vacuo* at 40 ^oC. All cleavage filtrates and washes were combined and concentrated *in vacuo* to afford a brown oil which was triturated with Et_2O and filtered to give 2 mg of a white solid. HPLC (method 1) shows target aldehyde **44** (t_R = 2.36 min) with reduced concentration of Weinreb amide impurity, but other impurities remain.

6.6.3 Synthesis and Templated Terminal Extension of a PNA Oligomer (Chapter 5.4)

Probe **P13** was synthesized as described as above (Chapter 6.2.2).

P13: ***m/z* (MALDI-TOF MS)**; for $C_{127}H_{166}N_{57}O_{41}$ $(M+H)^+$: calcd 3146.26, found 3146.07; **HPLC** t_R = 10.60 min (method 6).

Templated extension protocol (Figure 5.4)

A sample of acetal **50** (1.6 mg, 3.2 mmol) was hydrolyzed in 90 % TFA:H_2O (300 μL) for 1 h at a temperature maintained between 0 and -10 ^oC, then blown down with N_2 (g) to a yellow oil. Et_2O (1 mL) was added and blown down with N_2 (g) × 2 to afford a white solid which was dissolved in H_2O (316 μL) to give a concentration of approximately 10 mM of crude aldehyde **44**. Mass spectrometric and HPLC analysis confirmed the presence of aldehyde **44** (see Chapter 6.5.1) and a spot of the aqueous solution on a silica TLC plate turned orange on treatment with 2,4-DNP dip (confirming the presence of aldehydic functionality). This solution could not be stored, and was used directly in templated extension reactions.

For templated extension, the conditions described in Chapter 6.3.3(i) were employed, but using 6.4 μL of the solution of aldehyde **44** (approximately 10 mM) prepared as described in place of the nucleobase aldehyde solutions. One reaction was performed in which the aldehyde was added after the initial hybridization at 80 ^oC (Figure 5.4a) and one in which the aldehyde was present from the start (Figure 5.4b). A control reaction was performed in which water was used in place of the DNA solution (Figure 5.4c). For MALDI-TOF analysis, +**44** incorporation resulted in a mass increase of +416.

REFERENCES

1. J. M. Berg, J. L.Tymoczko and L. Stryer, in *Biochemistry*, 5th edn., W. H. Freeman and Company, New York, 2002, ch. 5, pp. 117-142.
2. J. D. Watson and F. H. C. Crick, *Nature*, 1953, **171**, 737-738.
3. C. A. Hutchison III, *Nucleic Acids Res.*, 2007, **35**, 6227-6237.
4. F. Sanger and A. R. Coulson, *J. Mol. Biol.*, 1975, **94**, 441-448.
5. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe and M. Smith, *Nature*, 1977, **265**, 687-695.
6. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U. S. A.*, 1977, **74**, 560-564.
7. W. Fiers, R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert and M. Ysebaert, *Nature*, 1978, **273**, 113-120.
8. F. Sanger, S. Nicklen and A. R. Coulson, *Proc. Natl. Acad. Sci. U. S. A.*, 1977, **74**, 5463-5467.
9. The Nobel Foundation, *The Nobel Prize in Chemistry 1980,* <http://nobelprize.org/nobel_prizes/chemistry/laureates/1980/>, accessed 5th March 2011.
10. C. Brückler, Ph.D. Thesis, University of Edinburgh, 2006.
11. L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood, *Nature*, 1986, **321**, 674-679.
12. M. Kircher and J. Kelso, *Bioessays*, 2010, **32**, 524-536.
13. D. Butler, *Nature*, 2001, **409**, 747-748.
14. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson,

R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F.

Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen, *Nature*, 2001, **409**, 860-921.

15. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad,

B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu, *Science*, 2001, **291**, 1304-1351.

16. S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg and J. C. Venter, *PLoS Biol.*, 2007, **5**, e254.
17. M. Wadman, *Nature*, 2008, **452**, 788.
18. Y.-H. Rogers and J. C. Venter, *Nature*, 2005, **437**, 326-327.
19. M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T.

Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner,. P. Yu, R. F. Begley and J. M. Rothberg, *Nature*, 2005, **437**, 376-380.

20 E. R. Mardis, *Trends Genet.*, 2008, **24**, 133-141.

21. R. D. Mitra, J. Shendure, J. Olejnik, E. Krzymanska-Olejnik and G. M. Church, *Anal. Biochem.*, 2003, **320**, 55-65.

22. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. S. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. D. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church and C. A. Reid, *Science*, 2010, **327**, 78-81.

23. T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss and Z. Xie, *Science*, 2008, **320**, 106-109.

24. D. Pushkarev, N. F. Neff and S. R. Quake, *Nat. Biotechnol.*, 2009, **27**, 847-850.

25. D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs and J. M. Rothberg, *Nature*, 2008, **452**, 872-876.

26. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians,

R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and S. Turner, *Science*, 2009, **323**, 133-138.

27. C. S. Chin, J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarskis, P. Peluso, E. E. Paxinos, Y. Yamaichi, S. B. Calderwood, J. J. Mekalanos, E. E. Schadt and M. K. Waldor, *N. Engl. J. Med.*, 2011, **364**, 33-42.
28. J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid and H. Bayley, *Nat. Nanotechnol.*, 2009, **4**, 265-270.
29. E. Y. Chan, *Mutat. Res., Fundam. Mol. Mech. Mutagen.*, 2005, **573**, 13-40.
30. D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. H. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin and J. A. Schloss, *Nat. Biotechnol.*, 2008, **26**, 1146-1153.
31. M. Eisenstein, *Nat. Biotechnol.*, 2010, **28**, 994.
32. N. Pourmand, M. Karhanek, H. H. Persson, C. D. Webb, T. H. Lee, A. Zahradnikova and R. W. Davis, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 6466-6470.
33. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y.

Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L.

Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F Zacharia, F. S. Collins, K. Kennedy, R. Jamieson and J. Stewart, *Nature*, 2007, **449**, 851-861.

34. L. Feuk, A. R. Carson and S. W. Scherer, *Nat. Rev. Genet.*, 2006, **7**, 85-97.
35. K. A. Frazer, S. S. Murray, N. J. Schork and E. J. Topol, *Nat. Rev. Genet.*, 2009, **10**, 241-251.
36. M. Ehrlich and R. Y. Wang, *Science*, 1981, **212**, 1350-1357.
37. S. Kriaucionis and N. Heintz, *Science*, 2009, **324**, 929-930.
38. M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind and A. Rao, *Science*, 2009, **324**, 930-935.
39. X. Ke, M. S. Taylor and L. R. Cardon, *Eur. J. Hum. Genet.*, 2008, **16**, 506-515.
40. A. C. Syvanen, *Nat. Genet.*, 2005, **37**, S5-S10.
41. M. Eichelbaum, M. Ingelman-Sundberg and W. E. Evans, *Annu. Rev. Med.*, 2006, **57**, 119-137.
42. J. C. Barrett and L. R. Cardon, *Nat. Genet.*, 2006, **38**, 659-662.
43. S. Kim and A. Misra, *Annu. Rev. Biomed. Eng.*, 2007, **9**, 289-320.
44. P. Y. Kwok, *Annu. Rev. Genomics Hum. Genet.*, 2001, **2**, 235-258.
45. J. Perkel, *Nat. Methods*, 2008, **5**, 447-453.
46. C. S. Carlson, T. L. Newman and D. A. Nickerson, *Curr. Opin. Chem. Biol.*, 2001, **5**, 78-85.
47. J. Ragoussis, *Annu. Rev. Genomics Hum. Genet.*, 2009, **10**, 117-133.
48. M. Schena, in *Microarray Analysis*, John Wiley and Sons Inc., Hoboken, New Jersey, 2003, ch. 1, pp. 1-25.
49. F. J. Steemers, W. Chang, G. Lee, D. L. Barker, R. Shen and K. L. Gunderson, *Nat. Methods*, 2006, **3**, 31-33.
50. R. T. Ranasinghe and T. Brown, *Chem. Commun.*, 2005, **44**, 5487-5502.
51. P. M. Holland, R. D. Abramson, R. Watson and D. H. Gelfand, *Proc. Natl. Acad. Sci. U. S. A.*, 1991, **88**, 7276-7280.
52. K. J. Livak, *Genet. Anal.*, 1999, **14**, 143-149.
53. M. P. Johnson, L. M. Haupt and L. R. Griffiths, *Nucleic Acids Res.*, 2004, **32**, e55.

54. L. Bonetta, *Nat. Methods*, 2005, **2**, 305-312.

55. N. Thelwell, S. Millington, A. Solinas, J. Booth and T. Brown, *Nucleic Acids Res.*, 2000, **28**, 3752-3761.

56. D. Whitcombe, J. Theaker, S. P. Guy, T. Brown and S. Little, *Nat. Biotechnol.*, 1999, **17**, 804-807.

57. C. R. Newton, A. Graham, L. E. Heptinstall, S. J. Powell, C. Summers, N. Kalsheker, J. C. Smith and A. F. Markham, *Nucleic Acids Res.*, 1989, **17**, 2503-2516.

58. A. Solinas, L. J. Brown, C. McKeen, J. M. Mellor, J. Nicol, N. Thelwell and T. Brown, *Nucleic Acids Res.*, 2001, **29**, e96.

59. D. J. French, C. L. Archard, T. Brown and D. G. McDowell, *Mol. Cell. Probes*, 2001, **15**, 363-374.

60. D. J. French, C. L. Archard, M. T. Andersen and D. G. McDowell, *Mol. Cell. Probes*, 2002, **16**, 319-326.

61. N. Ben Gaied, J. A. Richardson, D. G. Singleton, Z. Zhao, D. French and T. Brown, *Org. Biomol. Chem.*, 2010, **8**, 2728-2734.

62. J. A. Richardson, M. Gerowska, M. Shelbourne, D. French and T. Brown, *ChemBioChem*, 2010, **11**, 2530-2533.

63. V. Lyamichev, A. L. Mast, J. G. Hall, J. R. Prudent, M. W. Kaiser, T. Takova, R. W. Kwiatkowski, T. J. Sander, M. de Arruda, D. A. Arco, B. P. Neri and M. A. Brow, *Nat. Biotechnol.*, 1999, **17**, 292-296.

64. J. G. Hall, P. S. Eis, S. M. Law, L. P. Reynaldo, J. R. Prudent, D. J. Marshall, H. T. Allawi, A. L. Mast, J. E. Dahlberg, R. W. Kwiatkowski, M. de Arruda, B. P. Neri and V. I. Lyamichev, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 8272-8277.

65. M. Olivier, *Mutat. Res., Fundam. Mol. Mech. Mutagen.*, 2005, **573**, 103-110.

66. J. Tost and I. G. Gut, *Clin. Biochem.*, 2005, **38**, 335-350.

67. J. Tost and I. G. Gut, *Mass Spectrom. Rev.*, 2002, **21**, 388-418.

68. K. Chughtai and R. M. A. Heeren, *Chem. Rev.*, 2010, **110**, 3237-3277.

69. J. Ragoussis, G. P. Elvidge, K. Kaur and S. Colella, *PloS Genet.*, 2006, **2**, 920-929.

70. L. A. Haff and I. P. Smirnov, *Genome Res.*, 1997, **7**, 378-388.

71. S. Kim, J. R. Edwards, L. Deng, W. Chung and J. Ju, *Nucleic Acids Res.*, 2002, **30**, e85.

72. S. Gabriel, L. Ziaugra and D. Tabbaa, *Curr. Protoc. Hum. Genet.*, 2009, Chapter 2, Unit 2.12.

73. T. Wenzel, T. Elssner, K. Fahr, J. Bimmler, S. Richter, I. Thomas and M. Kostrzewa, *Nucleosides Nucleotides Nucleic Acids*, 2003, **22**, 1579-1581.

74. S. Sauer, D. H. Gelfand, F. Boussicault, K. Bauer, F. Reichert and I. G. Gut, *Nucleic Acids Res.*, 2002, **30**, e22.

75. S. Sauer, D. Lechner, K. Berlin, H. Lehrach, J. L. Escary, N. Fox and I. G. Gut, *Nucleic Acids Res.*, 2000, **28**, e13.

76. S. Sauer and I. G. Gut, *Rapid Commun. Mass Spectrom.*, 2003, **17**, 1265-1272.

77. Z. Fei and L. M. Smith, *Rapid Commun. Mass Spectrom.*, 2000, **14**, 950-959.

78. K. H. Buetow, M. Edmonson, R. MacDonald, R. Clifford, P. Yip, J. Kelley, D. P. Little, R. Strausberg, H. Koester, C. R. Cantor and A. Braun, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 581-584.

79. X. Sun, H. Ding, K. Hung and B. Guo, *Nucleic Acids Res.*, 2000, **28**, e68.

80. T. Blondal, B. G. Waage, S. V. Smarason, F. Jonsson, S. B. Fjalldal, K. Stefansson, J. Gulcher and A. V. Smith, *Nucleic Acids Res.*, 2003, **31**, e155.

81. M. Egholm, T. Bentin and P. E. Nielsen, in *Peptide Nucleic Acids: Protocols and Applications*, ed. P. E. Nielsen, Horizon Bioscience, Wymondham, Norfolk, 2nd edn., 2004, ch. 1, pp. 1-36.

82. P. E. Nielsen, M. Egholm, R. H. Berg and O. Buchardt, *Science*, 1991, **254**, 1497-1500.

83. L. Bialy, J. J. Diaz-Mochon, E. Specker, L. Keinicke and M. Bradley, *Tetrahedron*, 2005, **61**, 8295-8305.

84. C. Avitabile, L. Moggio, L. D. D'Andrea, C. Pedone and A. Romanelli, *Tetrahedron Lett.*, 2010, **51**, 3716-3718.

85. G. Breipohl, J. Knolle, D. Langner, G. OMalley and E. Uhlmann, *Bioorg. Med. Chem. Lett.*, 1996, **6**, 665-670.

86. K. L. Dueholm, M. Egholm, C. Behrens, L. Christensen, H. F. Hansen, T. Vulpius, K. H. Petersen, R. H. Berg, P. E. Nielsen and O. Buchardt, *J. Org. Chem.*, 1994, **59**, 5767-5773.

87. S. A. Thomson, J. A. Josey, R. Cadilla, M. D. Gaul, C. F. Hassman, M. J. Luzzio, A. J. Pipe, K. L. Reed, D. J. Ricca, R. W. Wiethe and S. A. Noble, *Tetrahedron*, 1995, **51**, 6179-6194.

88. A. Ray and B. Norden, *FASEB J.*, 2000, **14**, 1041-1060.

89. H. Orum, P. E. Nielsen, M. Egholm, R. H. Berg, O. Buchardt and C. Stanley, *Nucleic Acids Res.*, 1993, **21**, 5332-5336.

90. C. Thiede, E. Bayerdorffer, R. Blasczyk, B. Wittig and A. Neubauer, *Nucleic Acids Res.*, 1996, **24**, 983-984.

91. P. L. Ross, K. Lee and P. Belgrader, *Anal. Chem.*, 1997, **69**, 4197-4202.

92. P. Jiang-Baucom, J. E. Girard, J. Butler and P. Belgrader, *Anal. Chem.*, 1997, **69**, 4894-4898.

93. S. Ye, X. G. Liang, Y. Yamamoto and M. Komiyama, *Chem. Lett.*, 2003, **32**, 10-11.

94. T. J. Griffin, W. Tang and L. M. Smith, *Nature Biotechnol.*, 1997, **15**, 1368-1372.

95. P. Schatz, J. Distler, K. Berlin and M. Schuster, *Nucleic Acids Res.*, 2006, **34**, e59.

96. O. Bauer, A. Guerasimova, S. Sauer, S. Thamm, M. Steinfath, R. Herwig, M. Janitz, H. Lehrach and U. Radelof, *Rapid Commun. Mass Spectrom.*, 2004, **18**, 1821-1829.

97. B. Boontha, J. Nakkuntod, N. Hirankarn, P. Chaumpluk and T. Vilaivan, *Anal. Chem.*, 2008, **80**, 8178-8186.

98. R. J. Ball, P. S. Green, N. Gale, G. J. Langley, and T. Brown, *Artificial DNA: PNA & XNA*, 2010, **1**, 27-35.

99. P. T. Corbett, J. Leclaire, L. Vial, K. R. West, J. L. Wietor, J. K. Sanders and S. Otto, *Chem. Rev.*, 2006, **106**, 3652-3711.

100. N. Giuseppone and J.-M. Lehn, *Chemistry*, 2006, **12**, 1715-1722.

101. J. D. Cheeseman, A. D. Corbett, J. L. Gleason and R. J. Kazlauskas, *Chemistry*, 2005, **11**, 1708-1716.

102. B. Brisig, J. K. Sanders and S. Otto, *Angew. Chem., Int. Ed.*, 2003, **42**, 1270-1273.

103. E. G. Bardaji, E. Freisinger, B. Costisella, C. A. Schalley, W. Bruning, M. Sabat and B. Lippert, *Chem. Eur. J.*, 2007, **13**, 6019-6039.

104. A. P. Martinez and W. W. Lee, *J. Org. Chem.*, 1965, **30**, 317-318.

105. M. T. Doel, A. S. Jones and N. Taylor, *Tetrahedron Lett.*, 1969, **27**, 2285-2288.

106. Z. Q. Xu, Y. L. Qiu, S. Chokekijchai, H. Mitsuya and J. Zemlicka, *J. Med. Chem.*, 1995, **38**, 875-882.

107. G. N. Parkinson, in *Quadruplex Nucleic Acids*, ed. S. Neidle and S. Balasubramanian, Royal Society of Chemistry, Cambridge, 2006, ch. 1, pp. 1-30.

108. L. A. Loeb and B. D. Preston, *Annu. Rev. Genet.*, 1986, **20**, 201-230.

109. J. Lhomme, J.-F. Constant and M. Demeunynck, *Biopolymers*, 1999, **52**, 65-83.

110. T. J. Matray and E. T. Kool, *Nature*, 1999, **399**, 704-708.

111. M. Egholm, O. Buchardt, L. Christensen, C. Behrens, S. M. Freier, D. A. Driver, R. H. Berg, S. K. Kim, B. Norden and P. E. Nielsen, *Nature*, 1993, **365**, 566-568.

112. J. M. Heemstra and D. R. Liu, *J. Am. Chem. Soc.*, 2009, **131**, 11347-11349.

113. Y. Ura, J. M. Beierle, L. J. Leman, L. E. Orgel and M. R. Ghadiri, *Science*, 2009, **325**, 73-77.

114. M. J. Welsh, B. W. Ramsey, F. Accurso and G. R. Cutting, in *The Metabolic and Molecular Bases of Inherited Diseases*, ed. C. R. Scriver, A. L. Beaudet, W. S. Sly and D. Valle, McGraw-Hill, New York, 8th edn., 2001, vol. 3, ch. 201, pp. 5121-5188.

115. D. C. Gadsby, P. Vergani and L. Csanady, *Nature*, 2006, **440**, 477-483.

116. J. R. Riordan, J. M. Rommens, B. S. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J. L. Chou, M. L. Drumm, M. C. Iannuzzi, F. S. Collins and L. C. Tsui, *Science*, 1989, **245**, 1066-1073.

117. A. E. Shrimpton, I. Mcintosh and D. J. H. Brock, *J. Med. Genet.*, 1991, **28**, 317-321.

118. A. Quint, I. Lerer, M. Sagi and D. Abeliovich, *Am. J. Med. Genet. A*, 2005, **136A**, 246-248.

119. G. He, S. Rapireddy, R. Bahal, B. Sahu and D. H. Ly, *J. Am. Chem. Soc.*, 2009, **131**, 12088-12090.

120. C. D. Mamotte, *Clin. Biochem. Rev.*, 2006, **27**, 63-75.

121. S. Tabor and C. C. Richardson, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 6339-6343.

122. J. M. Butler, P. Jiang-Baucom, M. Huang, P. Belgrader and J. Girard, *Anal. Chem.*, 1996, **68**, 3283-3287.

123. W. W. Grody, G. R. Cutting and M. S. Watson, *Genet. Med.*, 2007, **9**, 739-744.

124. D. H. Farkas, N. E. Miltgen, J. Stoerker, D. van den Boom, W. E. Highsmith, L. Cagasan, R. McCullough, R. Mueller, L. Tang, J. Tynan, C. Tate and A. Bombard, *J. Mol. Diagn.*, 2010, **12**, 611-619.

125. X. Sun and J. K. Lee, *J. Org. Chem.*, 2010, **75**, 1848-1854.

126. F. Pompanon, A. Bonin, E. Bellemain and P. Taberlet, *Nat. Rev. Genet.*, 2005, **6**, 847-859.

127. Y. Ohnishi, T. Tanaka, K. Ozaki, R. Yamada, H. Suzuki and Y. Nakamura, *J. Hum. Genet.*, 2001, **46**, 471-477.

128. J. A. Sanchez, K. E. Pierce, J. E. Rice and L. J. Wangh, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 1933-1938.

129. H. Perry-O'Keefe, X.-W. Yao, J. M. Coull, M. Fuchs and M. Egholm, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 14670-14675.

130. J. M. Lee, H. Cho and Y. Jung, *Angew. Chem., Int. Ed.*, 2010, **49**, 8662-8665.

131. A. T. Krueger and E. T. Kool, *Chem. Biol.*, 2009, **16**, 242-248.

132. S. A. Benner, *Science*, 2004, **306**, 625-626.

133. I. Hirao, *Curr. Opin. Chem. Biol.*, 2006, **10**, 622-627.

134. Y. J. Seo, G. T. Hwang, P. Ordoukhanian and F. E. Romesberg, *J. Am. Chem. Soc.*, 2009, **131**, 3246-3252.

135. C. Switzer, S. E. Moroney and S. A. Benner, *J. Am. Chem. Soc.*, 1989, **111**, 8322-8323.

136. J. A. Piccirilli, T. Krauch, S. E. Moroney and S. A. Benner, *Nature*, 1990, **343**, 33-37.

137. J. D. Bain, C. Switzer, A. R. Chamberlin and S. A. Benner, *Nature*, 1992, **356**, 537-539.

138. M. L. Collins, B. Irvine, D. Tyner, E. Fine, C. Zayati, C. A. Chang, T. Horn, D. Ahle, J. Detmer, L. P. Shen, J. Kolberg, S. Bushnell, M. S. Urdea and D. D. Ho, *Nucleic Acids Res.*, 1997, **25**, 2979-2984.

139. S. A. Benner, *Acc. Chem. Res.*, 2004, **37**, 784-797.

140. M. J. Moser, D. J. Marshall, J. K. Grenier, C. D. Kieffer, A. A. Killeen, J. L. Ptacin, C. S. Richmond, E. B. Roesch, C. W. Scherrer, C. B. Sherrill, C. V. Van Hout, S. J. Zanton and J. R. Prudent, *Clin. Chem.*, 2003, **49**, 407-414.

141. S. C. Johnson, D. J. Marshall, G. Harms, C. M. Miller, C. B. Sherrill, E. L. Beaty, S. A. Lederer, E. B. Roesch, G. Madsen, G. L. Hoffman, R. H. Laessig, G. J. Kopish, M. W. Baker, S. A. Benner, P. M. Farrell and J. R. Prudent, *Clin. Chem.*, 2004, **50**, 2019-2027.

142. T. A. Martinot and S. A. Benner, *J. Org. Chem.*, 2004, **69**, 3972-3975.

143. Z. Yang, D. Hutter, P. Sheng, A. M. Sismour and S. A. Benner, *Nucleic Acids Res.*, 2006, **34**, 6095-6101.

144. Z. Yang, A. M. Sismour, P. Sheng, N. L. Puskar and S. A. Benner, *Nucleic Acids Res.*, 2007, **35**, 4238-4249.

145. P. Sheng, Z. Yang, Y. Kim, Y. Wu, W. Tan and S. A. Benner, *Chem. Commun.*, 2008, **41**, 5128-5130.

146. Z. Yang, F. Chen, S. G. Chamberlin and S. A. Benner, *Angew. Chem., Int. Ed.*, 2010, **49**, 177-180.

147. S. Hoshika, F. Chen, N. A. Leal and S. A. Benner, *Angew. Chem., Int. Ed.*, 2010, **49**, 5554-5557.

148. N. Minakawa, N. Kojima, S. Hikishima, T. Sasaki, A. Kiyosue, N. Atsumi, Y. Ueno and A. Matsuda, *J. Am. Chem. Soc.*, 2003, **125**, 9970-9982.

149. A. T. Krueger, H. Lu, A. H. Lee and E. T. Kool, *Acc. Chem. Res.*, 2007, **40**, 141-150.

150. B. A. Schweitzer and E. T. Kool, *J. Am. Chem. Soc.*, 1995, **117**, 1863-1872.

151. E. T. Kool and H. O. Sintim, *Chem. Commun.*, 2006, **35**, 3665-3675.

152. K. M. Guckian, J. C. Morales and E. T. Kool, *J. Org. Chem.*, 1998, **63**, 9652-9656.

153. A. M. Leconte, G. T. Hwang, S. Matsuda, P. Capek, Y. Hari and F. E. Romesberg, *J. Am. Chem. Soc.*, 2008, **130**, 2336-2343.

154. Y. J. Seo and F. E. Romesberg, *ChemBioChem*, 2009, **10**, 2394-2400.

155. S. Matsuda, J. D. Fillo, A. A. Henry, P. Rai, S. J. Wilkens, T. J. Dwyer, B. H. Geierstanger, D. E. Wemmer, P. G. Schultz, G. Spraggon and F. E. Romesberg, *J. Am. Chem. Soc.*, 2007, **129**, 10466-10473.

156. M. Kimoto, T. Mitsui, R. Yamashige, A. Sato, S. Yokoyama and I. Hirao, *J. Am. Chem. Soc.*, 2010, **132**, 15418-15426.

157. G. T. Hwang, Y. Hari and F. E. Romesberg, *Nucleic Acids Res.*, 2009, **37**, 4757-4763.

158. D. L. McMinn, A. K. Ogawa, Y. Wu, J. Liu, P. G. Schultz and F. E. Romesberg, *J. Am. Chem. Soc.*, **1999**, *121*, 11585-11586.

159. G. Haaima, H. F. Hansen, L. Christensen, O. Dahl and P. E. Nielsen, *Nucleic Acids Res.*, 1997, **25**, 4639-4643.

160. A. B. Eldrup, C. Christensen, G. Haaima and P. E. Nielsen, *J. Am. Chem. Soc.*, 2002, **124**, 3254-3262.

161. A. Sen and P. E. Nielsen, *Biophys. Chem.*, 2009, **141**, 29-33.

162. V. Chenna, S. Rapireddy, B. Sahu, C. Ausin, E. Pedroso and D. H. Ly, *ChemBioChem*, 2008, **9**, 2388-2391.

163. R. F. Borch, M. D. Bernstein and H. D. Durst, *J. Am. Chem. Soc.*, 1971, **93**, 2897-2904.

164. J. Lohse, O. Dahl and P. E. Nielsen, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11804-11808.

165. S. Kawahara and T. Uchimaru, *Eur. J. Org. Chem.*, 2003, **14**, 2577-2584.

166. T. Lan and L. W. McLaughlin, *Biochemistry*, 2001, **40**, 968-976.

167. O. Yaren, M. Mosimann and C. J. Leumann, *Angew. Chem., Int. Ed.*, 2011, **50**, 1935-1938.

168. J. J. Chen, X. Cai and J. W. Szostak, *J. Am. Chem. Soc.*, 2009, **131**, 2119-2121.

169. J. P. Schrum, A. Ricardo, M. Krishnamurthy, J. C. Blain and J. W. Szostak, *J. Am. Chem. Soc.*, 2009, **131**, 14560-14570.

170. Y. Z. Xu, N. B. Karalkar and E. T. Kool, *Nat. Biotechnol.*, 2001, **19**, 148-152.

171. A. H. El-Sagheer, V. V. Cheong and T. Brown, *Org. Biomol. Chem.*, 2011, **9**, 232-235.

172. S. Ficht, C. Dose and O. Seitz, *ChemBioChem*, 2005, **6**, 2098-2103.

173. N. Griesang, K. Giessler, T. Lommel and C. Richert, *Angew. Chem., Int. Ed.*, 2006, **45**, 6144-6148.

174. U. Plutowski, S. R. Vogel, M. Bauer, C. Deck, M. J. Pankratz and C. Richert, *Org. Lett.*, 2007, **9**, 2187-2190.

175. J. T. Goodwin and D. G. Lynn, *J. Am. Chem. Soc.*, 1992, **114**, 9197-9198.

176. Z. Y. J. Zhan and D. G. Lynn, *J. Am. Chem. Soc.*, 1997, **119**, 12420-12421.

177. X. Li, Z-Y. J. Zhang, R. Knipe and D. G. Lynn, *J. Am. Chem. Soc.*, 2002, **124**, 746-747.

178. D. M. Rosenbaum and D. R. Liu, *J. Am. Chem. Soc.*, 2003, **125**, 13924-13925.

179. X. Li and D. R. Liu, *Angew. Chem., Int. Ed.*, 2004, **43**, 4848-4870.

180. R. E. Kleiner, Y. Brudno, M. E. Birnbaum and D. R. Liu, *J. Am. Chem. Soc.*, 2008, **130**, 4646-4659.

181. N. M. Bell, R. Wong and J. Micklefield, *Chem. Eur. J.*, 2010, **16**, 2026-2030.

182. B. W. Bycroft, W. C. Chan, S. R. Chhabra and N. D. Hone, *Chem. Commun.*, 1993, **9**, 778-779.

183. A. Moulin, J. Martinez and J. A. Fehrentz, *J. Pept. Sci.*, 2007, **13**, 1-15.

184. M. C. de Koning, L. Petersen, J. J. Weterings, M. Overhand, G. A. van der Marel and D. V. Filippov, *Tetrahedron*, 2006, **62**, 3248-3258.

185. Y. Ohta, S. Itoh, A. Shigenaga, S. Shintaku, N. Fujii and A. Otaka, *Org. Lett.*, 2006, **8**, 467-470.

186. M. Paris, C. Pothion, A. Heitz, J. Martinez and J. A. Fehrentz, *Tetrahedron Lett.*, 1998, **39**, 1341-1344.

187. R. V. Jones, L. Godorhazy, N. Varga, D. Szalay, L. Urge and F. Darvas, *J. Comb. Chem.*, 2006, **8**, 110-116.

188. S. I. Al-Gharabli, S. T. A. Shah, S. Weik, M. F. Schmidt, J. R. Mesters, D. Kuhn, G. Klebe, R. Hilgenfeld and J. Rademann, *ChemBioChem*, 2006, **7**, 1048-1055.

189. K. Sato and M. Sasaki, *Tetrahedron*, 2007, **63**, 5977-6003.

190. G. J. Wells, M. Tao, K. A. Josef and R. Bihovsky, *J. Med. Chem.*, 2001, **44**, 3488-3503.

191. J. Sun, Y. Dong, L. Cao, X. Wang, S. Wang and Y. Hu, *J. Org. Chem.*, 2004, **69**, 8932-8934.

192. X. W. Zhang, J. Rodrigues, L. Evans, B. Hinkle, L. Ballantyne and M. Pena, *J. Org. Chem.*, 1997, **62**, 6420-6423.

193. S. Kitamura, H. Fukushi, T. Miyawaki, M. Kawamura, N. Konishi, Z. Terashita and T. Naka, *J. Med. Chem.*, 2001, **44**, 2438-2450.

194. H. Peng, D. Carrico, T. Van, M. Blaskovich, C. Bucher, E. E. Pusateri, S. M. Sebti and A. D. Hamilton, *Org. Biomol. Chem.*, 2006, **4**, 1768-1784.

195. J. Dimaio and B. Belleau, *J. Chem. Soc., Perkin Trans. 1*, 1989, **9**, 1687-1689.

196. B. J. Min, X. Gu, T. Yamamoto, R. R. Petrov, H. Qu, Y. S. Lee and V. J. Hruby, *Tetrahedron Lett.*, 2008, **49**, 2316-2319.

197. S. S. Jain, F. A. L. Anet, C. J. Stahle and N. V. Hud, *Angew. Chem., Int. Ed.*, 2004, **43**, 2004-2008.

198. N. V. Hud, S. S. Jain, X. H. Li and D. G. Lynn, *Chem. Biodiversity*, 2007, **4**, 768-783.

199. E. D. Horowitz, A. E. Engelhart, M. C. Chen, K. A. Quarles, M. W. Smith, D. G. Lynn and N. V. Hud, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 5288-5293.

200. H. E. Gottlieb, V. Kotlyar and A. Nudelman, *J. Org. Chem.*, 1997, **62**, 7512-7515.

201. W. C. Chan and P. D. White, in *Fmoc Solid Phase Peptide Synthesis*, W. C. Chan and P. D. White, Oxford University Press, Oxford, 2000, ch. 2, pp. 41-76.

202. M. Egholm, T. Bentin and P. E. Nielsen, in *Peptide Nucleic Acids: Protocols and Applications*, ed. P. E. Nielsen, Horizon Bioscience, Wymondham, Norfolk, 2nd edn., 2004, Appendix, p. 308.

203. A. Muratovska, R. N. Lightowlers, R. W. Taylor, D. M. Turnbull, R. A. J. Smith, J. A. Wilce, S. W. Martin and M. P. Murphy, *Nucleic Acids Res.*, 2001, **29**, 1852-1863.

204. I. Wempen, G. B. Brown, T. Ueda and J. J. Fox, *Biochemistry*, 1965, **4**, 54-57.

205. T. Tanaka, C. Tsuda, T. Miura, T. Inazu, S. Tsuji, S. Nishihara. M. Hisamatsu and T. Kajimoto, *Synlett*, 2004, **2**, 243-246.

206. S. E. Klassen, G. H. Daub and D. L. Vanderjagt, *J. Org. Chem.*, 1983, **48**, 4361-4366.

207. H. Rink, *Tetrahedron Lett*., 1987, **28**, 3787-3790.

208. J. J. Diaz-Mochon, L. Bialy and M. Bradley, *Org. Lett.*, 2004, **6**, 1127-1129.

209. R. M. Osterman, B. A. Mckittrick and T. M. Chan, *Tetrahedron Lett.*, 1992, **33**, 4867-4870.

210. *Int. Pat. Appl.*, WO 2005/005378, 2005.

APPENDIX 1

Calculation of the Number of Moles of DNA Template Obtained by Isolation from a Buccal Swab

There is one copy of the CFTR gene in the haploid human genome. The amount of template present for analysis in the DNA derived from a buccal swab thus equates to the number of haploid genomes present. Given that the haploid human genome contains 3×10^9 base pairs with a GC content of approximately 40 %[‡] and the Isohelix kit manufacturer's claim of 2 – 10 μg of DNA per swab, then (neglecting contributions from mitochondrial and any bacterial DNA present):

Formula weight (fwt) of an AT base pair = 615.4 g/mol

Formula weight (fwt) of a GC base pair = 616.4 g/mol

Mean fwt of a base pair = (0.6 × 615.4) g/mol + (0.4 × 616.4) g/mol
= 615.8 g/mol

Fwt of (haploid) genomic DNA = (No. of base-pairs) × (Fwt of a base-pair)
$= (3 \times 10^9) \times 615.8$ g/mol
$= 1.85 \times 10^{12}$ g/mol

Number of moles in 2 μg extracted DNA = (Mass) / (Fwt of genomic DNA)
$= (2 \times 10^{-6}\text{ g}) / (1.85 \times 10^{12}\text{ g/mol})$
$= 1 \times 10^{-18}$ mol
= 1 attomole

Number of moles in 10 μg extracted DNA = (Mass) / (Mass of 1 mole)
$= (10 \times 10^{-6}\text{ g}) / (1.85 \times 10^{12}\text{ g/mol})$
$= 5 \times 10^{-18}$ mol
= 5 attomoles

∴ Range of moles of DNA template isolatable by Isohelix kit (according to the manufacturer) = <u>1 - 5 attomoles (or (1-5) × 10^{-18} moles)</u>

[‡] R. Horton, L. A. Moran, G. Scrimgeour, M. Perry and D. Rawn, in *Principles of Biochemistry*, Pearson Education Inc., Upper Saddle River, New Jersey, 4th edn., 2006, p 590.

APPENDIX 2

Calculation of the Number of Moles of DNA Template Obtained After Asymmetric PCR

Mean formula weight (fwt) of a base pair = 615.8 g/mol (see Appendix 2)

Approximate fwt of a 92 bp (G551D) amplicon = (No. of bp) × (Fwt of a bp)

$= 92 \times 615.8$ g/mol

$= 5.7 \times 10^4$ g/mol

Sample volume used for analysis = 26 μL

For sample FB:

DNA concentration after symmetric PCR = 97 ng/μL (see Table 3.4)

Mass of DNA template present = Concentration × Volume

$= 97$ ng/μL $\times$ 26 μL

$= 2{,}522$ ng

$= 2.5$ μg

Number of moles of templating DNA = (Mass) / (Approx. fwt of G551D amplicon)

$= (2.5 \times 10^{-6}$ g) / $(5.7 \times 10^4$ g/mol)

$= 4.4 \times 10^{-11}$ mol

= <u>44 picomoles</u>

For sample JJ:

DNA concentration after symmetric PCR = 108 ng/μL (see Table 3.5)

Mass of DNA template present = Concentration × Volume

$= 108$ ng/μL $\times$ 26 μL

$= 2{,}808$ ng

$= 2.8$ μg

Number of moles of templating DNA = (Mass) / (Approx. fwt of G551D amplicon)

$= (2.8 \times 10^{-6}$ g) / $(5.7 \times 10^4$ g/mol)

$= 4.9 \times 10^{-11}$ mol

= <u>49 picomoles</u>

APPENDIX 3

Spreadsheets of Raw Peak Table Data for the Genotyping of Clinical Cystic Fibrosis Samples

Presented below are the output from spreadsheets after the input of raw peak table data (for peaks > 5 % relative intensity) into Microsoft Excel. Representative examples for each clinical CF sample are given, together with the formulae used to convert these data into an output genotype (p.190).

Sample CF1:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3966.810891	100	P7	**DF508/N**
2	3967.818546	92.45	P7	
3	3983.029162	12.1	X	
4	4128.043848	81.54	DF508	
5	4128.394273	80.66	DF508	
6	4129.358563	66.62	DF508	
7	4241.923812	48.29	P8	
8	4418.934391	61.01	wt508	
9	4419.905466	52.46	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3917.426522	99.61	P5	**G551D/N**
2	3918.334878	100	P5	
3	3921.89034	26.94	P5	
4	3923.592319	12.22	X	
5	4056.190367	21.15	wt551	
6	4057.775898	13.71	wt551	
7	4068.785866	7.21	G551D	
8	4069.53848	11.2	G551D	
9	4070.81629	9.71	G551D	

Sample CF2:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3965.972546	56.3	P7	**DF508/N**
2	3967.050118	74.44	P7	
3	3967.946466	66.17	P7	
4	3968.982236	44.64	P7	
5	4127.04115	53.29	DF508	
6	4128.027786	76.88	DF508	
7	4129.025653	67.76	DF508	
8	4241.127068	64	P8	
9	4242.133299	96.62	P8	
10	4243.097441	97.09	P8	
11	4418.216332	60.52	wt508	
12	4419.203163	94.08	wt508	

13	4420.102237	100	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	2027.504879	5.97	X	**G551D/N**
2	2049.514874	6.32	X	
3	2083.437109	5.26	X	
4	3917.367294	54.74	P5	
5	3918.110703	49.47	P5	
6	3918.875128	49.83	P5	
7	3920.954942	11.58	P5	
8	3931.639044	7.72	X	
9	4054.21074	100	wt551	
10	4055.048381	98.94	wt551	
11	4058.685684	17.19	wt551	
12	4069.155729	22.81	G551D	
13	4070.256402	25.26	G551D	
14	4071.394952	25.61	G551D	
15	4072.363292	9.83	G551D	

Sample CF3:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3966.321635	100	P7	**DF508/N**
2	3967.050265	92.13	P7	
3	3968.105076	72.62	P7	
4	3979.321628	9.24	X	
5	3980.25447	10.95	X	
6	4127.084941	66.33	DF508	
7	4128.09606	70.71	DF508	
8	4128.95638	65.37	DF508	
9	4132.521468	10.61	DF508	
10	4241.054795	33.61	P8	
11	4242.107886	40.66	P8	
12	4242.595098	43.94	P8	
13	4243.308503	35.66	P8	
14	4419.43804	41.62	wt508	
15	4420.042289	41.27	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3916.457	65.31	P5	**G551D/N**
2	3917.453125	100	P5	
3	3918.23475	97.45	P5	
4	3919.521072	65.14	P5	
5	3923.308244	10.38	X	
6	3930.334713	7.82	X	
7	4054.032003	71.77	wt551	
8	4054.46543	93.88	wt551	
9	4055.530693	84.52	wt551	
10	4060.872586	11.57	X	
11	4068.624407	55.1	G551D	
12	4069.64724	75.17	G551D	

13	4070.335856	70.07	G551D
14	4072.115661	45.92	G551D

Sample CF4:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3966.90709	97.8	P7	**DF508/DF508**
2	3967.373764	94.81	P7	
3	3968.019263	83.78	P7	
4	3972.98992	8.82	X	
5	3980.078698	10.24	X	
6	3981.754657	8.82	X	
7	3982.591526	8.03	X	
8	3983.541533	7.25	X	
9	3984.87579	5.04	X	
10	4037.699634	5.04	X	
11	4128.173998	70.39	DF508	
12	4134.270038	6.93	X	
13	4135.144132	7.25	X	
14	4135.951949	6.46	X	
15	4242.638694	38.74	P8	
16	4245.619725	13.7	P8	
17	4256.469973	5.98	X	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3917.758854	100	P5	**N/N**
2	3919.091798	93.31	P5	
3	3985.793224	6.42	X	
4	4055.244417	89.37	wt551	
5	4055.845885	84.97	wt551	

Sample CF5:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3967.650682	100	P7	**N/N**
2	3968.59487	89.36	P7	
3	3970.649816	42.4	P7	
4	4036.081811	9.27	X	
5	4037.571882	9.12	X	
6	4242.256111	19.3	P8	
7	4243.771475	19.46	P8	
8	4244.674436	15.2	P8	
9	4418.717174	18.85	wt508	
10	4419.763378	27.81	wt508	
11	4420.738527	29.64	wt508	
12	4421.433853	27.05	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	2000.330403	9.11	X	**G551D/G551D**
2	2011.976627	7.97	X	
3	2015.812968	7.97	X	

4	2017.765048	7.29	X
5	2030.411167	6.15	X
6	2046.891048	12.53	X
7	2047.894361	15.26	X
8	2048.943374	34.17	X
9	2049.953232	40.55	X
10	2050.919032	27.79	X
11	2056.216034	9.11	X
12	2058.712358	8.66	X
13	2060.59043	11.16	X
14	2065.837476	9.11	X
15	2089.302499	8.43	X
16	2110.259203	7.52	X
17	2130.952018	7.97	X
18	2188.935372	5.92	X
19	2217.307016	6.6	X
20	2271.413417	6.61	X
21	2344.390998	6.15	X
22	2377.673951	6.15	X
23	2496.183939	5.24	X
24	2739.692873	5.24	X
25	3493.581087	5.01	X
26	3877.198144	5.47	X
27	3916.648971	100	P5
28	3917.44937	94.53	P5
29	3919.810669	38.5	P5
30	3921.492569	16.63	P5
31	4068.451383	16.4	G551D
32	4069.695505	15.72	G551D

Sample CF6:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3966.981282	100	P7	**N/N**
2	3982.63414	8.55	X	
3	4035.476547	8.12	X	
4	4242.069333	35.06	P8	
5	4419.699533	59.94	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3916.712815	81.73	P5	**G551D/N**
2	3917.832442	100	P5	
3	3918.762163	90.99	P5	
4	3922.923315	11.01	X	
5	4054.952408	33.29	wt551	
6	4055.674562	31.92	wt551	
7	4056.879448	23.03	wt551	
8	4069.229655	14.77	G551D	
9	4070.899652	13.89	G551D	

Sample CF7:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3965.271816	100	P7	**DF508/N**
2	3966.141573	95.8	P7	
3	3967.288353	75.37	P7	
4	3970.936195	11.11	P7	
5	3991.10737	5.41	X	
6	4033.799095	7.21	X	
7	4035.649593	8.41	X	
8	4108.108563	5.71	X	
9	4108.908582	6.01	X	
10	4124.304301	8.11	DF508	
11	4125.373278	24.32	DF508	
12	4126.250146	31.23	DF508	
13	4127.812577	26.13	DF508	
14	4129.218446	17.42	DF508	
15	4239.240033	33.93	X	
16	4240.279205	45.34	P8	
17	4241.032609	43.85	P8	
18	4243.318807	22.22	P8	
19	4245.198725	8.41	P8	
20	4309.429831	6.91	X	
21	4417.521421	27.33	wt508	
22	4419.024048	25.83	wt508	
23	4419.615081	24.93	wt508	
24	4420.607306	13.51	wt508	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3918.368568	100	P5	**G551D/N**
2	3923.58215	6.05	X	
3	3987.592872	5.01	X	
4	4054.323249	12.84	wt551	
5	4055.362795	15.21	wt551	
6	4056.207267	13.92	wt551	
7	4069.160193	5.79	G551D	
8	4070.19426	5.64	G551D	
9	4071.201838	5.45	G551D	

Sample CF8:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3965.75527	81.59	P7	**N/N**
2	3966.751143	100	P7	
3	3967.531055	98.71	P7	
4	3968.600941	72.98	P7	
5	3972.53592	9.27	X	
6	4035.440709	8.75	X	
7	4037.660848	9.27	X	
8	4240.805064	23.17	P8	
9	4241.852771	27.16	P8	
10	4242.747324	28.19	P8	

11	4243.795387	22.14	P8
12	4245.871481	7.85	P8
13	4418.133485	8.37	wt508
14	4419.089855	13.9	wt508
15	4419.702991	14.16	wt508
16	4420.761652	11.33	wt508
17	4421.838635	8.75	wt508

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3875.717426	6.95	X	**G551D/N**
2	3876.647004	7.87	X	
3	3917.592608	95.71	P5	
4	3917.942364	100	P5	
5	3919.447771	80.49	P5	
6	3921.131645	34.42	P5	
7	4054.620928	36.47	wt551	
8	4055.134477	36.47	wt551	
9	4055.627643	35.55	wt551	
10	4056.318533	28.4	wt551	
11	4058.346363	14.3	wt551	
12	4059.936464	7.05	X	
13	4068.277435	11.85	G551D	
14	4069.474072	15.32	G551D	
15	4071.146326	15.12	G551D	
16	4072.508146	8.79	G551D	

Sample CF9:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3967.19636	78.15	P7	**DF508/DF508**
2	3968.161302	69.04	P7	
3	3972.225868	14.9	X	
4	4128.586172	100	DF508	
5	4129.541257	84.94	DF508	
6	4138.809581	11.26	X	
7	4241.884773	53.15	P8	
8	4243.058207	56.62	P8	
9	4245.287982	31.13	P8	
10	4255.254261	8.61	X	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3917.532034	75	P5	**N/N**
2	3920.152722	29.8	P5	
3	4053.597879	80.14	wt551	
4	4054.465376	100	wt551	
5	4055.457146	87.84	wt551	

Sample CF10:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3965.470194	5.82	P7	**DF508/DF508**

2	3967.175935	21.87	P7
3	3968.112052	22.27	P7
4	4128.442136	98.31	DF508
5	4129.198417	100	DF508
6	4130.070862	85.74	DF508
7	4241.896098	19.49	P8
8	4242.450061	21.87	P8
9	4243.84993	18.39	P8

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3914.008647	11.04	P5	**N/N**
2	3915.576769	20.93	P5	
3	3917.067715	16.03	P5	
4	3917.790575	12.51	P5	
5	3918.794531	7.85	P5	
6	4051.386989	46.2	wt551	
7	4052.266685	81.44	wt551	
8	4053.284947	100	wt551	
9	4053.837413	72.37	wt551	
10	4054.186478	70.65	wt551	
11	4054.896142	52.66	wt551	
12	4057.50089	16.03	wt551	

Sample CF11:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3969.08479	10.89	P7	**DF508/DF508**
2	4128.364824	97.88	DF508	
3	4129.170679	100	DF508	
4	4241.708599	9.08	P8	
5	4243.449123	12.95	P8	

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	2026.990876	5.91	X	**N/N**
2	3914.73384	11.83	P5	
3	3915.70913	20.5	P5	
4	3916.712561	19.05	P5	
5	3917.676266	12.22	P5	
6	4050.798617	22.34	wt551	
7	4051.827167	64.52	wt551	
8	4052.814675	100	wt551	
9	4053.558776	89.62	wt551	
10	4056.768344	19.58	wt551	
11	4057.57472	8.67	wt551	

Sample CF12:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3966.129543	9.32	P7	**DF508/DF508**
2	3967.197871	11.41	P7	
3	3968.182063	6.92	P7	

4	4126.281945	34.13	DF508
5	4127.271572	79.34	DF508
6	4128.274615	100	DF508
7	4129.258145	90.68	DF508
8	4130.24975	53.02	DF508
9	4131.21924	31.79	DF508
10	4132.251173	15.3	DF508
11	4241.366167	6.92	P8
12	4242.351448	10.07	P8
13	4243.325138	8.83	P8

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3915.827013	33.29	P5	**N/N**
2	3916.952678	44.38	P5	
3	3917.763194	40.31	P5	
4	3918.476601	27.81	P5	
5	4053.946749	100	wt551	
6	4054.696042	91.01	wt551	

Multiplex CF3:

Index	Centroid Mass	Relative Intensity	*LOOKUP Formula*	*GENOTYPE*
1	3915.374977	10.6	P5	**DF508/N**
2	3916.333324	31.85	P5	**G551D/N**
3	3917.319613	36.52	P5	
4	3918.273589	29.15	P5	
5	3919.239299	18.7	P5	
6	4052.368543	32.26	wt551	
7	4053.442433	80.78	wt551	
8	4054.406778	100	wt551	
9	4055.339504	86.18	wt551	
10	4056.252441	55.38	wt551	
11	4067.50691	8.57	G551D	
12	4068.401288	19.9	G551D	
13	4069.352807	24.99	G551D	
14	4070.296081	23.9	G551D	
15	4071.478928	13.92	G551D	
16	4072.485329	5.77	G551D	
17	4126.465614	14.23	DF508	
18	4127.505489	39.74	DF508	
19	4128.462542	52.78	DF508	
20	4129.364879	43.74	DF508	
21	4130.343941	29.2	DF508	
22	4131.241096	16	DF508	
23	4241.724751	6.81	P8	
24	4242.634432	9.45	P8	
25	4243.454839	7.79	P8	
26	4417.630675	10.7	wt508	
27	4418.857024	35.38	wt508	
28	4419.623937	51.33	wt508	
29	4420.571047	45.82	wt508	

30	4421.464937	31.53	wt508

LOOKUP formula used to identify peaks:

=LOOKUP(B2,{1500,3912,3922,3962,3972,4049,4059,4064,4074,4123,4133,4240,4252,4414,4424},{"X","P5","X","P7","X","wt551","X","G551D","X","DF508","X","P8","X","wt508","X"})

IF formula used to translate the peak list into a ΔF508 genotype:

=IF(AND(ISNUMBER(MATCH("DF508",D:D,0)),ISNUMBER(MATCH("wt508",D:D,0))),"DF508/N",IF(ISNUMBER(MATCH("DF508",D:D,0)),"DF508/DF508","N/N"))

IF formula used to translate the peak list into a G551D genotype:

=IF(AND(ISNUMBER(MATCH("G551D",D:D,0)),ISNUMBER(MATCH("wt551",D:D,0))),"G551D/N",IF(ISNUMBER(MATCH("G551D",D:D,0)),"G551D/G551D","N/N"))

APPENDIX 4
Publication

Permission to reproduce the following publication in print form only has been granted by the copyright holder Wiley-VCH.