TARGETING G-QUADRUPLEX DNA IN PROMOTERS OF CARDIAC FUNCTION-RELATED GENES

A thesis submitted to the Imperial College London
for the degree of Doctor of Philosophy

By

WENHUA ZHOU

March 2012

Section of Molecular Medicine
National Heart and Lung Institute
Imperial College London
London
SW7 2AZ
United Kingdom

Declaration

I hereby declare that this thesis, submitted in to fulfill the requirements for the degree of Doctor of Philosophy of Imperial College London, and entitled "TARGETING G-QUADRUPLEX DNA IN PROMOTERS OF CARDIAC FUNCTION-RELATED GENES", represents my own work and has not been submitted to any institute for any degree or other qualification. Parts of my work described in this thesis were previously published, and were appropriately referenced in the relevant sections herein.

Wenhua Zhou

March 2012

Abstract

G-quadruplexes (G4) are four-stranded DNA secondary structures, which are involved in a diverse range of biological processes. Although the anti-cancer potential of G4s in promoters of oncogenes has been thoroughly investigated, the functions of promoter G4s in non-cancer-related genes are not well understood. The aims of this thesis are (i) to investigate the biological importance of promoter G4s in cardiac function-related genes, and (ii) to explore the feasibility and potential of multi-targeting strategy at the single gene level. These aims have been achieved by revealing the prevalence of G4s in TRRs (transcription regulatory regions) of cardiac function-related genes via bioinformatics approaches, and then investigating potential biological significance of G4s in promoters of human MEF2D and TnIc genes via a combination of biophysical, molecular, and cell biology approaches.

By using bioinformatics approaches, TRRs of cardiac function-related genes were found to be enriched by potential G4-forming sequences. According to these results, G4s from the promoters of human *MEF2D* and *TnIc* were chosen and subjected to biophysical characterisations in solution. By using EMSA, DMS footprinting, CD and smFRET spectroscopy, the formation, thermodynamic stability, and unfolding kinetics of MEF2D and TnIc G4s were investigated. Briefly, in 100 mM K^+, the MEF2DG4 can adopt a hybrid of very stable parallel/anti-parallel G4(s), while the TnIc MNSG4 and -80G4 are able to adopt anti-parallel and parallel structures, respectively, with comparable stability as compared to other oncogene G4s. Subsequently, cooperative regulatory roles of TnIc G4s as enhancers once stabilised by a G4-binding ligand (di-copper complex) were observed *in vitro* mainly by dual luciferase reporter assays in HEK293 cells. Three possible G4-mediated mechanisms have been proposed. This work provides the first indication about the biological significance of promoter G4s in cardiac function-related genes, and the feasibility to target multiple G4s at the single gene level.

Acknowledgements

I would like to thank my supervisors, Dr Liming Ying, and Prof Mike Ferenczi, for providing me the opportunity to do this Ph. D., and for their patient guidance and continuous support in this course. I am also grateful to the National Heart and Lung Institute Foundation for the financial support.

I would also like to thank Dr Nigel Brand, for kindly offering me the opportunity to carry out my research at the Heart Science Centre, Harefield, Imperial College London. It was particularly kind of him to help me as much as he can during my research at the Heart Science Centre. I would also like to thank Prof Ramon Vilar, and his students Kogularamanan Suntharalingam and Beata Klejevskaja at Department of Chemistry, Imperial College London, for providing me the G4-binding ligands.

In particular, I would like to thank my parents. This thesis would not have been possible without their disinterested support and assistance during my Ph. D. study.

Table of Contents

Declaration 2
Abstract 3
Acknowledgements 4
Table of Contents 5
List of Figures 9
List of Tables 12
Abbreviations 13

Chapter 1 Introduction 18
1.1 Structural diversity of G4 DNA 19
1.2 Biological relevance of G4 DNA 24
1.2.1 ILPR 24
1.2.2 MyoD-G4 interaction 25
1.2.3 Telomeric G4 26
1.2.4 G4s in promoters of other genes 31
1.2.5 G4-interacting proteins 37
1.3 Bioinformatics studies on G4s 39
1.4 Methodologies for biophysical characterisation of G4 DNA 41
1.4.1 EMSA 41
1.4.2 DMS footprinting 42
1.4.3 Circular dichroism 43
1.4.4 Single molecule FRET 43
1.4.5 Dual luciferase reporter assays 47
1.5 G4-binding ligands 48
1.5.1 *In situ* protonated G4-binding ligands 49
1.5.2 N-methylated aromatic G4-binding ligands 53
1.5.3 Metallo-organic G4-binding ligands 55
1.5.4 Neutral macrocyclic G4-binding ligands 57
1.5.5 Novel methods to synthesise G4-binding ligands 58
1.6 Off-target effects vs. multi-targeting effects 60
1.7 G4 in the heart 62
1.7.1 G4s in the pPu/pPy regions 62
1.7.2 Regulatory roles of G4s at RNA level 65
1.7.3 Interaction of G4s with TFs 67
1.8 MEF2D and TnIc 69
1.8.1 Human MEF2D 69
1.8.2 Human TnIc 70
1.9 Hypotheses and aims 71
1.10 References 73

Chapter 2 Materials and Methods 93
2.1 Materials 93
2.1.1 General reagents 93
2.1.2 Oligos used in biophysical studies 93
2.1.3 Primers and synthetic oligos used in *in vitro* studies 94
2.1.4 Plasmids used in *in vitro* studies 95
2.1.5 Strains and media 97
2.1.6 Cell line and media 97
2.1.7 G4-binding ligands 97
2.2 Biophysical characterisations 98
2.2.1 Oligo characterisations 98
2.2.2 G4 preparation 99
2.2.3 Electrophoresis mobility shift assay (EMSA) 99
2.2.4 CD spectroscopy and CD melting 100
2.2.5 Single molecule FRET studies 103
2.2.6 Unfolding kinetics of G4s measured by FRET experiments at bulk level 108
2.2.7 Dimethyl sulfate (DMS) footprinting of TnIc MNSG4 and -80G4 110
2.2.8 Ligand-G4 binding affinity assay in solution 114
2.3 Molecular and cell biology studies 116
2.3.1 Agarose gel electrophoresis, purification, and quantification of DNA 116
2.3.2 Polymerase chain reaction (PCR) 117
2.3.3 Restriction enzyme (RE) digestion 118
2.3.4 Plasmid construction via Gateway® recombination cloning 119
2.3.5 DNA sequencing 130
2.3.6 Bacterial transformation 131
2.3.7 Small-scale plasmid purification 131
2.3.8 Large-scale plasmid purification 132
2.3.9 Initiation of the HEK293 cell line from frozen stock 133
2.3.10 Maintenance and passage of the HEK293 cell line 134
2.3.11 Storing the HEK293 cell line in liquid nitrogen 134
2.3.12 Calcium phosphate-mediated transient transfection of the HEK293 cell line 135
2.3.13 G4-binding ligand treatment of HEK293 cell line 136
2.3.14 Dual luciferase reporter assays 137
2.4 Bioinformatics Studies 138
2.4.1 Obtaining gene/transcript sequences and features 138
2.4.2 Redundancy issue 140
2.4.3 Enrichment of G4-forming motifs in different tissues 140
2.4.4 Importance of G4s in different pathways 141
2.5 References 144

Chapter 3 Bioinformatics Studies 145
3.1 Introduction 145

3.2 Enrichment of G4-forming sequences in different tissues 148
3.3 G4 abundance in TRRs of transcripts (the *F* score) 153
3.4 G4 location significance in TRRs of transcripts (the *Q* score) 155
3.5 Correlation between G4 abundance and G4 location significance scores 158
3.6 G4-important pathways 161
3.7 Discussion 164
3.7.1 Tissue-specific enrichment of G4s in TRRs of genes 165
3.7.2 G4-important pathways 166
3.8 References 169

Chapter 4 Biophysical Studies 173
4.1 Introduction 173
4.2 Formation of G4s observed by EMSA 177
4.3 Formation and thermal stability of G4s determined by CD spectroscopy 179
4.4 Conformational analysis G4 by smFRET 184
4.4.1 MEF2DG4 184
4.4.2 TnIc MNSG4 and -80G4 188
4.4.3 TnIc two-repeat MNSG4 192
4.5 G4/duplex competition assay using smFRET 194
4.5.1 MEF2DG4 194
4.5.2 TnIc MNSG4 and -80G4 196
4.6 G4 hybridisation and unfolding kinetics 197
4.6.1 MEF2DG4 197
4.6.2 TnIc MNSG4 and -80G4 199
4.6.3 TnIc two-repeat MNSG4 203
4.7 DMS footprinting of TnIc MNSG4 and -80G4 207
4.8 Stabilising TnIc MNSG4 and -80G4 by G4-binding ligand 209
4.8.1 Stabilising ability of di-copper complex (compound 3) determined by CD melting 209
4.8.2 Binding affinity of di-copper complex to TnIc MNSG4 and -80G4 212
4.9 Discussion 214
4.9.1 Formation of MEF2D and TnIc G4s in solution 214
4.9.2 Conformational polymorphism of MEF2D and TnIc G4s 215
4.9.3 Unfolding thermodynamics and kinetics of MEF2D and TnIc G4s 218
4.9.4 Concerns in thermodynamic and kinetic characterisations of G4 unfolding 223
4.10 References 226

Chapter 5 Molecular and Cell Biology Studies 231

5.1 Introduction **231**
5.2 Alignment of the human *TnIc* -600/+67 bp promoter region **234**
5.3 Alignment of human *TnIc* wild type promoter and constructed promoter **236**
5.4 Comparison of transcription activities of wild type (WT), engineered wild type (E), and MNSG4-mutated (M) human *TnIc* promoters **238**
5.5 Transcription activities of human *TnIc* promoters containing variable number of MNSG4 ..... **240**
5.6 Transcription activities of human *TnIc* promoters with wild type and mutated -80G4 ..... **241**
5.7 G4-binding ligands screened in the HEK293 cell line **244**
5.8 Other possible mechanisms that "activate" the transcription activity of human *TnIc* promoter via G4-binding ligands **246**
5.9 Positive regulatory functions of human *TnIc* promoter G4s **247**
5.10 Discussion **251**
5.10.1 MNSG4 and -80G4 elements are only conserved in the promoters of *TnIc* in the genomes of primate animals **251**
5.10.2 HEK293 cell line vs. rat/mouse cardiomyocytes **252**
5.10.3 Interaction between compound 3 and TnIc G4 elements in living cells **253**
5.10.4 Possible mechanisms of G4-mediated human *TnIc* promoter activation **254**
5.11 References **261**

Chapter 6 Summary and Future Work **265**
6.1 Importance of G4s in cardiac function-related genes **266**
6.2 Biophysical characterisations of human *MEF2D* and *TnIc* promoter G4s **268**
6.2.1 Formation and conformation **268**
6.2.2 Unfolding thermodynamics **269**
6.2.3 Unfolding kinetics **269**
6.3 Regulatory roles of G4s in human *TnIc* promoter **270**
6.4 Future work **272**
6.5 References **275**

List of Figures

Figure 1.1: Schematic representations showing the structural features of G4s **20**
Figure 1.2: Different grooves formed in an example of a G4 ..... **21**
Figure 1.3: Schematic diagram showing three different side loops connecting individual strands in a G4 ..... **22**
Figure 1.4: Proposed roles of G4s at the end of telomeres *in vivo* **28**
Figure 1.5: Proposed regulatory mechanism of a G4 in the NHE III_1 region of *MYC* promoter **33**
Figure 1.6: Schematic representation of the single molecule FRET technique **46**
Figure 1.7: Cationic G4-binding ligands with *in situ* protonation **51**
Figure 1.8: Cationic G4-binding ligands with N-methylation **54**
Figure 1.9: Metallo-organic G4-binding ligands **56**
Figure 1.10: Neutral macrocyclic G4-binding ligands **58**

Figure 2.1: Simplified outline of the home-built dual-channel confocal single-molecule fluorescence microscope setup **104**
Figure 2.2: Molecular mechanism of DMS footprinting to reveal guanines involved in formation of G4 structures **111**
Figure 2.3: Schematic illustration of Gateway® technology **120**
Figure 2.4: Schematic illustration of the protocol to generate *attB*-flanked human TnIc MNS concatemers via three rounds of ligation **122**
Figure 2.5: Ligation products of wild type MNS in a 12 % native polyacrylamide gel **123**
Figure 2.6: Schematic illustration of BP recombination reaction to recombine the MNS concatemers into pDONR221 donor vector in order to create entry vectors **125**
Figure 2.7: Schematic illustration of LR recombination reaction to recombine the MNS concatemers into pGL3-Basic-GW destination vector to give expression vector pGL3-Basic-GW-MNS **126**
Figure 2.8: Construction of luciferase expression vectors driven by human TnIc MNSG4 and -80G4 ..... **128**
Figure 2.9: Schematic illustration of the *Dpn*I-mediated site-directed mutagenesis **130**
Figure 2.10: Schematic illustration of the mechanism of bioluminescent reactions catalyzed by firefly and *Renilla* luciferases **138**

Figure 3.1: Enrichment of G4s in TRRs of genes active in different tissues **149**
Figure 3.2: Distribution of G4 abundance scores (the *F* scores) and corresponding cumulative frequencies (the *CF* scores) **154**
Figure 3.3: Probability distribution and location significance scores (the Q scores) distribution of G4s in TRRs of transcripts **157**
Figure 3.4: Correlation between the location significance (the *CQ* score) and abundance (the *CF* socre) of G4s identified in TRRs of all transcripts

with redundancy 159
Figure 3.5: The normalized percentage of G4s identified in different regions of TRRs of G4-less-important and G4-importance transcripts 160

Figure 4.1: Native and denaturing gel electrophoresis indicating the formation of G4s 178
Figure 4.2: CD spectra and CD melting results of MEF2DG4 and TnIc G4s 182
Figure 4.3: Schematic illustration of the dual-labeled MEF2DG4 systems for single molecule and bulk FRET studies 185
Figure 4.4: Single molecule FRET histograms in MEF2DG4 conformational analysis 187
Figure 4.5: Schematic illustration of the dual-labeled TnIc G4s systems for single molecule and bulk FRET studies 189
Figure 4.6: Single molecule FRET histograms in TnIc MNSG4 conformational analysis 191
Figure 4.7: Single molecule FRET histograms in TnIc -80G4 conformational analysis 192
Figure 4.8: Single molecule FRET histograms in TnIc two-repeat MNSG4 conformational analysis 193
Figure 4.9: Single molecule FRET histograms in MEF2D G4/duplex competition assays and the proposed free energy landscapes 195
Figure 4.10: A model for the competition between MEF2DG4 and duplex in the presence of full length C-rich complimentary strand 198
Figure 4.11: MEF2DG4 unfolding curve and corresponding fitted curve 199
Figure 4.12: TnIc G4s unfolding curves and corresponding fitted curves 200
Figure 4.13: Arrhenius plotting for unfolding of TnIc G4s 202
Figure 4.14: TnIc two-repeat MNSG4 unfolding curve and corresponding fitted curve 204
Figure 4.15: TnIc two-repeat MNSG4 unfolding measured by smFRET and corresponding fitted curves 206
Figure 4.16: DMS footprinting results of TnIc MNSG4 and -80G4 208
Figure 4.17: CD melting of TnIc MNSG4 and -80G4 in the presence of the G4-binding ligand complex 3 211
Figure 4.18: Ligand-binding curves for the ss dual-labeled MNSG4 and -80G4 systems as a function of the ligand concentration 213

Figure 5.1: Alignment of proximal promoters of human, orangutan, mouse and rat *TnIc* genes 234
Figure 5.2: Alignment of the proximal promoter of wild type human *TnIc* genes and the engineered *TnIc* promoter 237
Figure 5.3: Comparison of the sequences of MNSG4 elements in different constructs and their different transcription activities 239
Figure 5.4: Comparison of the transcription activities of human *TnIc* promoters containing variable number of MNSG4 241

Figure 5.5: Comparison of the sequences and transcription activities of human *TnIc* promoters with wild type and mutated -80G4 ..... **243**
Figure 5.6: Structures of G4-binding ligands, and corresponding dose-response curves **245**
Figure 5.7: Antagonistic effects and cytotoxicity of compounds 1 - 3 ..... **247**
Figure 5.8: Transcription activity changes of human *TnIc* promoters in the presence of compound 3 (the di-copper complex) **250**
Figure 5.9: Schematic illustration of the G4/Sp1-mediated mechanism of human *TnIc* promoter activation **256**
Figure 5.10: Schematic illustration of the superhelicity-induced mechanism of human *TnIc* promoter activation **259**

List of Tables

Table 2.1: Details of oligos used in biophysical studies in solution **94**
Table 2.2: Details of primers and synthetic oligos used in *in vitro* studies **95**
Table 2.3: Details of plasmids used in *in vitro* studies **96**
Table 2.4: Bacterial strains used in *in vitro* studies **97**
Table 2.5: Recipes to prepare bacterial media **97**
Table 2.6: Recipes to prepare 12 % native and denaturing gels for EMSA experiments **100**
Table 2.7: Description of components of the home-built dual-channel confocal single-molecule fluorescence microscope setup **104**
Table 2.8: Recipes for preparing 16 % denaturing, 16 % native, and 20 % denaturing polyacrylamide gels for the three rounds of electrophoresis in DMS footprinting experiments **114**
Table 2.9: Recipe to prepare complete growth medium for HEK293 cells **134**
Table 2.10: Recipe to prepare 2 × HBS buffer **136**

Table 3.1: The rankings and scores of G4 enrichment in TRRs of genes active in different tissues **150**
Table 3.2: Comparison of G4 abundance in TRRs of transcripts (the *F* score) **155**
Table 3.3: Comparison of G4 location significance in TRRs of transcripts (the *Q* score) **158**
Table 3.4: CF_o and CQ_o values of TRRs of genes containing well-studied promoter G4s **161**
Table 3.5: KEGG pathways with high G4-importance **163**

Table 4.1: Energetics of MEF2DG4 and TnIc MNSG4 and -80G4 unfolding in solutions determined by CD melting **184**
Table 4.2: Parameters obtained from curve fittings of TnIc MNSG4 and -80G4 unfolding in solution **203**
Table 4.3: Parameters obtained from curve fittings of two-repeat TnIc MNSG4 unfolding measured by bulk and single molecule FRET experiments in solution **207**

Abbreviations

°C	degree Celsius
ΔG	Gibbs free energy change
$\Delta G_{(310K)}$	Gibbs free energy change at 37 °C
ΔH	enthalpy change
ΔS	entropy change
ΔT_m	melting temperature change
A	adenine
AP-2	activator protein-2
APD	avalanche photodiode
APS	ammonium persulfate
ATP	adenosine triphosphate
att	attachment
BCL-2	B-cell lymphoma 2
bHLH	basic helix-loop-helix
bp	base pairs
C	cytosine or carbon atom
CBF-A	CarG-box binding protein A
CC	click chemistry
*ccd*B	control of cell death gene
CD	circular dichroism
CF score	cumulative frequency of *F* score
ChIP	chromatin immunoprecipitation
CNBP	cellular nucleic acid binding protein
cpm	counts per minute
CQ score	cumulative frequency of *Q* score
csv	comma-separated values
DCC	dynamic combinatorial chemistry
DEAE	diethylaminoethanol
DMEM	Dulbecco Modified Eagle's Medium
DMS	dimethyl sulfate
DMSO	dimethyl sulfoxide
DNase I	deoxyribonuclease I
dNTP	deoxyribonucleotide triphosphate
ds	double-stranded
EDTA	ethylenediamine tetraacetic acid
EGFR	epidermal growth factor receptor
EMSA	electrophoretic mobility shift assays

F score	G4 abundance score
G	guanine
G-overhang	guanine-rich overhang
G-tetrad	guanine-tetrad
G-tract	guanine-tract
G4P	G4 formation potentials
G4(s)	G-quadruplex(es)
GIST	gastrointestinal stromal tumors
GRSDB	G-Rich Sequence Database
HAND	heart, autonomic nervous system, neural crest-derived
HC	hemicyanine tetrapeptide conjugate
HCM	hypertrophic cardiomyopathy
HEK293	human embryonic kidney 293
HIF-1	hypoxia-inducible factor 1
hnRNP A1	heterogeneous nuclear ribonucleoprotein A1
HNRNPK	heterogeneous nuclear ribonucleoprotein K
Hop1	homolog of presenilin 1
HPLC	high-performance liquid chromatography
IC_{50}	half maximal inhibitory concentration
ICT	isothermal titration calorimetry
IGF-2	insulin-like growth factor-2
ILPR	insulin linked polymorphism region
K	kelvin
K_a	association constant
K_b	binding constant
kb	kilobase
K_d	dissociation constant
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIT	c-Kit, V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
KLFs	kruppel-like zinc finger factors
KRAS	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
Ku70/86	ATP-dependent DNA helicase 2, subunit 1/2
LB	Luria-Bertani
LDH	lactate dehydragenase
MADS	MCM1, agamous, deficiency, serum response factor
MAPK	mitogen-activated protein kinase
MAZ	Myc-associated zinc finger protein
MCK	muscle-specific creatine kinase
MEF2	myocyte enhancer factor-2

MF3	muscle factor 3
MI	myocardial infarction
miR	microRNA
MNS	minisatellite
mRNA	messenger RNA
MYC	c-MYC, V-myc myelocytomatosis viral oncogene homolog
MyoD	myogenic differentiation 1
N	any nucleotide or nitrogen atom
NFAT	nuclear factor of activated T cells
NHE	nuclease-hypersensitivity element
NHS esters	N-hydroxysuccinimide esters
NM23-H2	nucleoside diphosphate kinase
NMR	nuclear magnetic resonance
No.	number
NRAS	neuroblastoma RAS viral oncogene homolog
O	oxygen atom
$O.D._{260/280}$	optical density at 260/280 nm
PAGE	polyacrylamide gel electrophoresis
PARP-1	poly (ADP-ribose) polymerase 1
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDGF	platelet-derived growth factor
PDGFR	platelet-derived growth factor receptor
PNA	peptide nucleic acid
PNK	polynucleotide kinase
polyG	poly-guanine
POT1	protection of telomeres 1 homolog
pPu/pPy	polypurine/polypyrimidine
psi	pounds per square inch
Q score	G4 location significance score
RAP1	repressor activator protein 1
RB1	retinoblastoma 1
rDNA	ribosomal DNA
RE	restriction enzyme
Rluc	*Renilla* luciferase
RNA	ribonucleic acid
rpm	rounds per minute
rRNA	ribosomal RNA
SAR	structure activity relationship

scFv	single-chain antibody fragment
SDS	sodium dodecyl sulphate
smFRET	single molecule fluorescence resonance energy transfer
sMtCK	sarcomeric mitochondrial creatine kinase
Sp1	specificity protein 1
ss	single-stranded
T	thymine
TAE	tris-acetate-EDTA buffer
Taq	DNA polymerase from *Thermus aquaticus*
TBA	thrombin binding aptamer
TBE	tris-borate-EDTA buffer
TBP/TEBP	telomere end-binding protein
TE	tris-EDTA buffer
$^{tel}EC_{50}$	half maximal effective concentration to inhibit telomerase extension
TEMED	Tetramethylethylenediamine
TERF2IP	telomeric repeat binding factor 2, interacting protein
TERT	telomerase reverse transcriptase
TF(s)	transcription factor(s)
TGF-β	transforming growth factor beta
TGP	tetrahymena G-DNA binding protein
T_m	melting temperature
TmPyP4	*meso*-5,10,15,20-Tetrakis-(N-methyl-4-pyridyl)porphine, Tetratosylate
TnIc/TNNI3	troponin cardiac isoform
TnIf/TNNI2	troponin fast skeletal muscle isoform
TnIs/TNNI1	troponin slow skeletal muscle isoform
TnT/TNNT1	slow skeletal muscle troponin T
TopoI	topoisomerase I
TRAP	telomeric repeat amplification assay
TRF/TERF	telomeric repeat binding factor
TRR	transcription regulatory region
TSS	transcription start site
T-loop	telomere loop
VEGF	vascular endothelial growth factor
VNTR	variable number of tandem repeats
U	units enzyme
Up1	proteolitic portion of hnRNP A1 retaining binding activity
uqTBP25	unimolecular quadruplex telomere-binding protein 25
UTR	untranslated region
UV-Vis	ultraviolet-visible

v/v	volume/volume
VDR	vitamin D receptor
VT-CD	Variable Temperature-CD
WRN	Werner syndrome
ZIC-1	Zic family member 1

Chapter 1 Introduction

Although it is clear that the majority of human genomic DNA adopts the iconic double helical structure (also known as B-form or duplex DNA) proposed by Watson and Crick [1], other non-B-form DNA secondary structures formed within certain sequences in the human genome have also been revealed in many *in vivo* and *in vitro* studies. These include left-handed DNA (Z-DNA) formed in $d(CpGp)_3$ DNA fragments [2, 3], A-DNA formed in dehydrated DNA samples, DNA-RNA hybrids, and double-stranded (ds) RNA [4, 5], DNA triplexes (H-DNA and sticky DNA) formed between one polypurine and two polypyrimidine strands [7, 8], cruciforms formed in inverted repeat DNA [9, 10], slipped (hairpin) DNA formed in direct repeat DNA [11, 12], G-quadruplex DNA (G4 DNA) and i-motifs (i-tetraplexes) formed in oligonucleotides with guanine and cytosine tracts respectively [13], and so on. These diverse non-B-form DNA structures are involved in many biological processes, and their formation and stability are dependent on DNA sequences, topology, interaction with DNA-binding proteins and DNA modifications (reviewed in [12, 14-16]). As 98 % of the human genome is non-coding, more and more interest has been focused on investigating possible biological functions of these transient and dynamic non-B-form DNA structures.

Unlike other non-B-form DNA, G-quadruplexes, or simply G4s are relatively stable under near-physiological conditions. The accumulated distribution of G4s in gene promoter regions throughout the human genome also makes them particularly important as compared to other non-B-form DNA structures [17]. Thus, since the discovery that G4s can form in the G-rich regions from human telomeric oligonucleotides in the late 1980s [18], there have been a growing number of studies focusing on the structures, dynamics, and biological functions of these fascinating DNA structures.

Long before the proposal of DNA double helix, the phenomenon that guanosine and its analogues can readily form a gel in concentrated solution was observed by Bang [19] at the beginning of the twentieth century. However, this unusual physical phenomenon remained unclear until the first G-tetrad (guanine-tetrad) structure

formed by guanylic acids was solved by Gellert and Davies using X-ray diffraction techniques 50 years later [20]. Inspired by the proposal of the double helix DNA model [1], interest in nucleic acids intensified during the 1960s and 1970s and self-assembling G-tetrad structures from both ribose and deoxyribose guanine homopolymers were identified and resolved in solution [21, 22] Subsequent X-ray diffraction studies and modelling on polyG (poly-guanine) further clarified the four-stranded hydrogen-bonding scheme of the G-tetrad and showed that it spontaneously adopts a four-stranded structure in solution under near-physiological conditions [23, 24]. However, little attention was paid to the potential biological significance of this non-canonical DNA structure until 1987, when Henderson and his colleagues first observed that under native conditions the G-rich telomeric sequence can spontaneously adopt a compact structure with faster mobility upon electrophoresis in native polyacrylamide gels in contrast to linear DNA with the same length [18]. This compact structure is now known as the G-quadruplex DNA or, more simply, G4.

1.1 Structural diversity of G4 DNA*

Generally, as proposed by Gellert and Davies in 1962 [20], the core structure of G4 DNA is relatively simple, which is composed of stacked G-tetrads with a cation located in the centre. In each G-tetrad plane, four guanines occupy each corner and form a G-tetrad plane, which is maintained by Hoogsteen hydrogen bonds between N1, N7, O6, and N2 of each guanine base (Figure 1.1 a). The formation of G4 simply requires four repeats of G-tracts (guanine-tracts), and each guanine in a G-tract can interact with guanines on other G-tracts via Hoogsteen bonds to form a G-tetrad. Two or more layers of G-tetrad can be stacked together to form a G4 structure. For example, the sequences of most studied stable intramolecular G4s can be generalized as $G_{\geq 3}(N_{1\sim 7}G_{\geq 3})_{\geq 3}$. Despite the simplicity of the core structure and forming sequences, the actual structures of G4s are related to a range of variables, such as strand stoichiometry, strand polarity, base geometry, sugar conformations, combination of side loops, and the nature of cations involved. Thus, actual structures of G4s are very varied, with even several distinct conformations able to form from the same G4-forming sequence.

* Part of this Section was adopted from my previous publication [6].

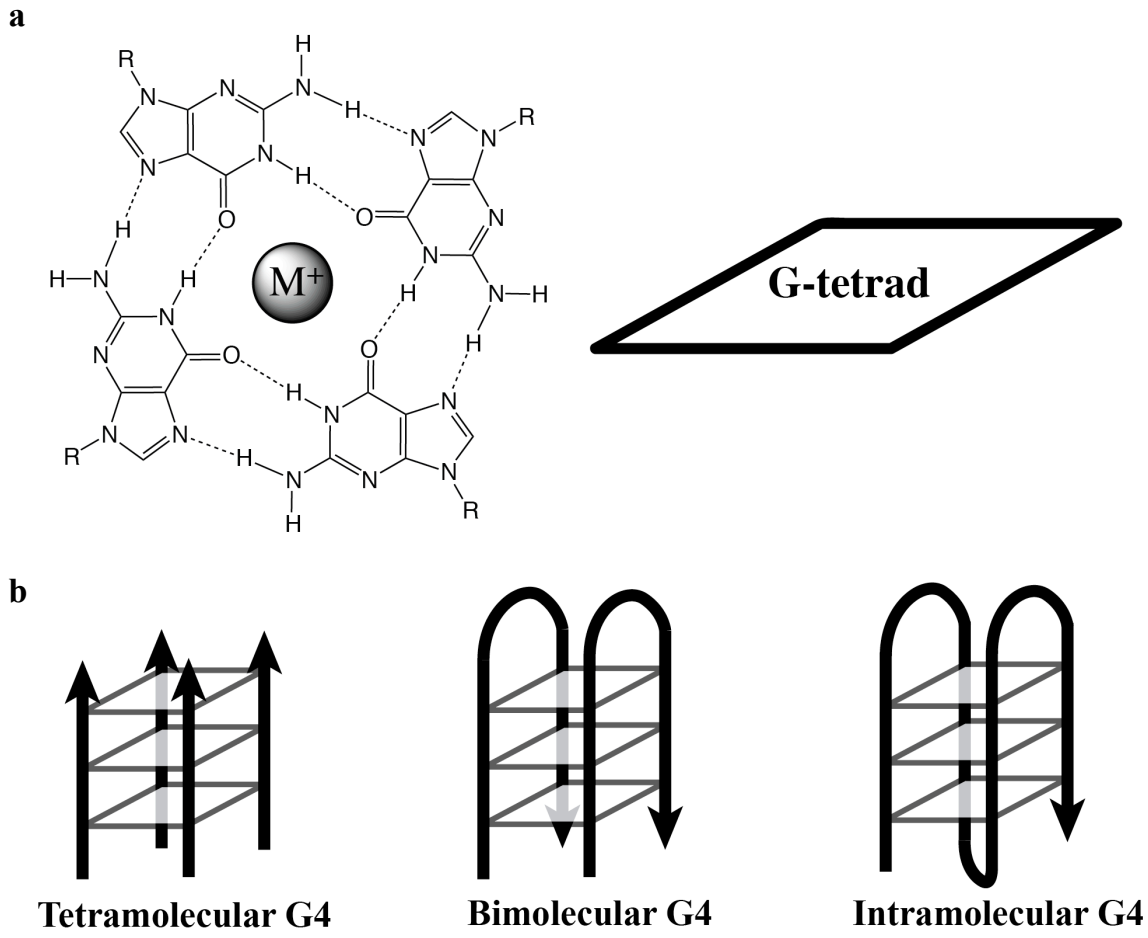


Figure 1.1: Schematic representations showing the structural features of G4s.
(a) A G-tetrad motif highlighting the Hoogsteen hydrogen bonds between N1, N7, O6, and N2 of each guanine base and a cation in the centre stabilising the G-tetrad (left), and a schematic G-tetrad plane (right). **(b)** Schematic diagram of G4 structures composed of four DNA strands (tetramolecular G4, left), two DNA strands (bimolecular G4, middle), and one DNA strand (intramolecular G4, right). **(c)** Schematic diagram showing G4 structures classified based on strand polarity, which represents (left to right direction) a parallel G4, an alternating anti-parallel G4, a G4 with two groups of adjacent parallel strands (overall anti-parallel), and a anti-parallel G4 with three parallel strands and one anti-parallel strand respectively. Arrows indicate 5' to 3' direction of DNA molecules. This Figure was taken from Zhou *et al.*, 2011 [6].

G4s can be formed by the coming together of four DNA strands (tetramolecular G4), two DNA strands (bimolecular G4) or within a single DNA strand (intramolecular G4) (Figure 1.1 b). Although *inter*molecular G4s were found to be likely involved in telomere-telomere association, chromosome pairing and recombination, in most cases, the *intra*molecular G4 is the most biologically relevant form. In terms of considering the function of G4s upon gene transcription, a G4 structure located upstream or downstream of a gene promoter could therefore act in *cis* upon the linked promoter region and modulate promoter output by interacting with certain *trans*-acting factors such as G4-binding TFs (transcription factors) or other proteins. Particular examples of intermolecular and intramolecular G4s will be given later.

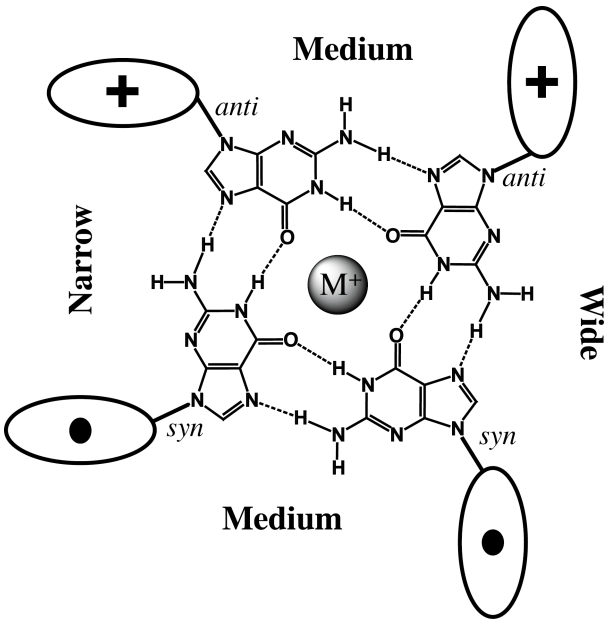


Figure 1.2: Different grooves formed in an example of a G4.
The different width of a groove is determined by combination of sugar geometry (*syn* or *anti* represents different glycosidic bond angles), and strand polarity (plus symbol or filled circles indicates different direction of each strand). This Figure was taken from Zhou *et al.*, 2011 [6].

Strand polarity is another origin of structural heterogeneity of G4 DNA. In the case of strand polarity, G4s can also be classified into parallel, anti-parallel, and parallel/anti-parallel hybrid structures, which are characterised by unique CD spectra (circular dichroism, will be discussed later) (Figure 1.1 c). At the same time, the conformation of deoxyribose (C2'-endo or C3'-endo conformation) on the backbone and glycosidic bond torsion angle (*syn* or *anti*) also related to the overall structure and stacking

energy of G4s. In anti-parallel B-form DNA with *anti* glycosidic bond torsion angle geometry, one minor and one major groove are found. However, in G4s, especially those with anti-parallel structures, due to different combinations of *syn* and *anti* glycosidic bond torsion angles, four grooves between individual strands can be wide, medium or narrow (Figure 1.2). This may provide diverse binding sites for different G4-binding proteins.

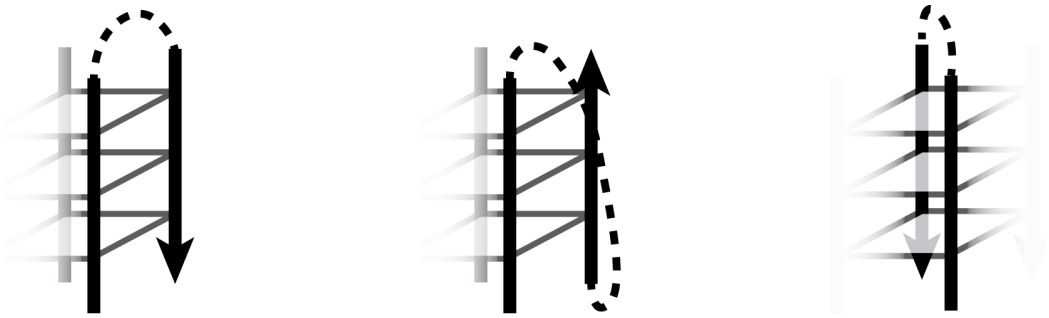


Figure 1.3: Schematic diagram showing three different side loops connecting individual strands in a G4.
This Figure was taken from Zhou *et al.*, 2011 [6].

Bimolecular and intramolecular G4s can have three types of side loops connecting individual strands: lateral (L), external (E) (also known as propeller-type loop), and diagonal (D) loop (Figure 1.3). Recent studies have shown that the length, sequences, and type of side loops can provide extra constrains to the overall structure and stability to the G4. Linkers with lengths shorter than two nucleotides cannot adopt diagonal loops, while one nucleotide is enough for an external loop, such as the G4s found in promoters of human *BCL-2* (B-cell lymphoma 2) [25, 26], *VEGF* (Vascular endothelial growth factor) [27], *MYC* (V-myc myelocytomatosis viral oncogene homolog) [28, 29], *KIT* (V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog) [30], and other genes. Short loops can impose topological constraints to the formation of G4, leading to parallel structures. For example, a systematic study showed that parallel intramolecular G4 is the only possible fold when three single nucleotide loops are present, while mixtures of parallel and anti-parallel G4s may exist when longer loops are present [31]. So far, the actual relationship between loop length and G4 stability is still not fully understood. For example, for G4s with total loop length less than five, an increase in loop length corresponds with a significant

decrease in G4 stability, while for G4s with a total loop length of more than five, such a relationship is not apparent [32]. Another study showed that in G4-forming sequences containing a single non-nucleosidic linker in each loop, G4 with medium loop length (GGG-p$(CH_2)_8$O) is more stable than those with shorter (GGG-p$(CH_2)_3$O) or longer loops (GGG-p$(CH_2CH_2)_8$O) [33]. Although generally the increase in G-tract length can increase the stability of intramolecular G4s [34], while the increase in loop length might decrease the stability of intramolecular G4s by facilitating the competition with C-rich complimentary strands [35], G4s with long side loops can still be relatively thermally stable. For example, in the presence of 100 mM K^+, the melting temperatures of two intramolecular G4s, both with a 12 nucleotides loop, from the promoter of the human *KRAS* (V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog) gene are 55 and 72 ºC, respectively [36].

Monovalent cations, especially K^+, can effectively interact with O6 carboxyl oxygen within the G4 core to form cation-dipole interactions with eight guanines. This interaction can reduce the electronic repulsion between central oxygen atoms and further enhance the stability of most G4 DNA structures [37]. The general trend of stabilising effects due to G4-associated cations is $K^+ >> Na^+ > Rb^+ > Cs^+ > Li^+$, suggesting that physiological conditions are more favourable for the formation of G4s. Besides the stabilising effects, the nature of cations can also influence the adopted structure of G4s, with more parallel G4s found in K^+ solution than in Na^+ or NH_4^+ solution [38]. For example, in Na^+ solution, human telomeric G4 can adopt an anti-parallel "basket" form with two lateral loops and one diagonal loop, while in K^+ solution, a parallel crystal structure has been observed [39].

Thus, even from the same forming sequence, structures of G4 are often heterogeneous with different biological relevance. Different conformational arrangements may be in dynamic equilibrium with each other [40]. In the examples of human telomeric G4 forming sequence (d[A$(GGGTTA)_3$GGG]), several distinct intramolecular G4s have been revealed. In 2002, a unique parallel "propeller" form of human telomeric G4 with three trinucleotide external loops was originally resolved by X-ray crystallography [39], while subsequent biophysical studies proposed that the "propeller" form may not be the major form in solution, where at least seven different G4 structures may exist from the same sequence in potassium solution [41]. Later, high resolution NMR (nuclear magnetic resonance) studies revealed two other anti-

parallel G4s (Form 1 with one TTA external loop and two TTA lateral loops, and Form 2 with three TTA lateral loops) in equilibrium [42]. More recently, a novel basket-type two G-tetrad layers G4 (Form 3 with one diagonal and two lateral loops) from the telomeric G4-forming sequence with the highest thermal stability has been revealed by NMR [43]. Similarly, the *MYC* promoter G4-forming sequence is known to adopt two different anti-parallel structures, the so-called "basket" and "chair" types. Of these two, the chair form is more kinetically favoured and biologically relevant compared with the other type [28]. Thus, due to this heterogeneous property, it is still impossible to predict actual G4 structure just based on the DNA sequence.

1.2 Biological relevance of G4 DNA

Since the discovery that human telomere sequence can adopt G4 structures in solution, G4-forming sequences have been identified in various positions within the human genome. Moreover, distinct biological roles of these novel structures have also been revealed.

1.2.1 ILPR

G4 found in the insulin linked polymorphism region (ILPR) is one of the G4s with potential biological functions that have attracted particular interest since the early days of G4 study. The ILPR, located at -363 bp relative to the transcription start site in the human insulin coding sequence, is composed of G-rich variable number of tandem repeats (VNTR), or minisatellites with consensus sequence ACAGGGGT(G/C)(T/C)GGGG [44, 45]. The number of ILPR repeats has also been found to be associated with insulin-dependent diabetes mellitus [46, 47]. A decade later, Hammondkosack *et al.* found that the ILPR is highly nuclease-hypersensitive in supercoiled plasmids and chromatin assembled *in vitro*, and subsequent fine mapping and electron microscopic analyses revealed a highly structured conformation within the G-rich strand which proved be a G4 structure [48-50]. Soon, the correlation between the ILPR, especially the variant *a* (predominant ILPR variant), and downstream insulin transcription activity was revealed. Kennedy *et al.* have shown

that the insulin promoter containing long ILPR (class III, ~175 repeats) has greater transcription activity as compared to the promoter with short ILPR (class I, ~40 repeats), possibly via a mechanism involving interaction between ILPR G4s and the insulin promoter binding factor Pur-1/MAZ (MYC-associated zinc finger protein) [51, 52]. Catasti *et al.* also found that single or double mutations that destabilise the ILPR G4 can lower the transcription activity [53]. Subsequently, the intramolecular G4 from ILPR ($(ACAGGGGTGTGGGG)_2$, a_2) was found to be able to bind with insulin ($K_d = 0.17 \pm 0.03$ μM [54]) and IGF-2 (insulin-like growth factor-2) with high affinity [55], which provided further circumstantial evidence for the biological significance of ILPR G4s [56, 57]. Although a variety of in solution experiments revealed a mixture of stable intramolecular parallel and anti-parallel G4s from the ILPR a_2 fragment, in which the anti-parallel form is more biologically relevant due to the higher stability and preference in insulin-binding [54, 57], the folding behaviours of ILPR G4s might be more complicated. For example, Basu *et al.* demonstrated that when long ILPR fragment (a_4) was included, two G4s can unfold cooperatively, possibly through G4-G4 interaction [58]. Although the molecular mechanism is still not fully understood, the G4-forming ILPR clearly indicated the structural diversity and corresponding potential biological relevance of G-rich minisatellites, which used to be regarded widely as meaningless fragments in the human genome.

1.2.2 MyoD-G4 interaction

Almost at the same time when the ILPR G4s were under investigation, specific interactions between the myogenic determination factor MyoD (myogenic differentiation 1) and G4-forming muscle gene elements were identified. Promoters and enhancers of many muscle-specific genes have been shown to contain a high incidence of G-rich elements, which can readily adopt inter- or intramolecular G4s in solution. In 1991, Santoro *et al.* observed that skeletal muscle DNA-binding proteins MF3 (muscle factor 3) and MyoD can bind to some single-stranded muscle gene elements [59]. Further studies found that the homodimeric form of MyoD is able to bind to various G4s from muscle-specific G-rich regulatory sequences, including a tetramolecular G4 from the human *MCK* (muscle-specific creatine kinase) gene enhancer fragment [60], bimolecular G4s from the human *CKMT2* (sarcomeric

mitochondrial creatine kinase), mouse *MCK*, and alpha-7 integrin promoters [61]. The MyoD homodimer binds to these G4s more specifically as compared to its designated cognate binding site the E-box, which usually conforms to the consensus sequence CANNTG [60, 61]. In muscle differentiation, heterodimers of the bHLH (basic helix-loop-helix) proteins MyoD and E12/47 bind to E-box sequences in the promoters of various muscle-specific genes and activate their transcription, while the homodimers of MyoD can bind to the E-box less avidly with much less activating effects. The interaction between G4s and homodimeric MyoD may well explain the function of the MyoD homodimer. Indeed, Fry *et al.* demonstrated recently that the presence of muscle-specific bimolecular G4s dramatically elevates the transcription activity driven by MyoD and E47 [62]. These authors proposed a model that the G4s formed within promoters of muscle-specific genes may provide localized docking sites for homodimeric MyoD to increase local concentration of MyoD protein upon the promoter. This enrichment of homodimeric MyoD could lead to a localized association with constitutive E-proteins in order to form functional heterodimers, which bind to adjacent E-boxes subsequently and activate downstream promoter activity. Although still very controversial and unratified experimentally, this model is intriguing, since it is not uncommon to find that G4 forming sequences and E-boxes are adjacent to each other in the promoters of many muscle and cardiac function-related genes. Thus, these findings may indicate a potential role for homodimers of MyoD, as well as other bHLH TFs, in regulating transcription of muscle-specific or cardiac-specific gene expression via G4-mediated mechanisms.

1.2.3 Telomeric G4

Since being discovered by Henderson and his colleagues in 1987 [18], telomeric G4s have attracted much attention as potential anti-cancer targets. The telomere contains a region of repetitive DNA at the end of chromosomes of eukaryotic organisms, which maintains the genome stability and protects internal genetic materials from exposure to the DNA damage response or fusion with neighbouring chromosomes (reviewed in [63]). Telomeres are composed of highly conserved tandem repeats of G-rich sequences containing three or four guanines, and hence several telomeric repeats can adopt G4 structures readily in solution. Telomeres may have varied length in different

species, and all telomeres have a single-stranded G-overhang (guanine-rich overhang) protruded from the 3' end of chromosomal DNA. For example, the 3' single-stranded overhangs of ciliated protozoans telomeres composed of TTTTGGGG repeats (ciliates *Oxytricha* and *Stylonychia*) are very short (16 nt 3' G-overhang [64]), while in human telomeres (TTAGGG) double-stranded G-rich repeats can extend for 9 - 15 kb with a 50 ~ 300 bp single-stranded protrusion [65]. The protection effects of telomeres are achieved by assembling a nucleoprotein complex known as shelterin, and forming unique structures (T-loop, telomere loop) in the telomere regions. Shelterin complex consists of many DNA binding proteins, including double-stranded telomeric repeat binding factors TRF1 and TRF2 (telomeric repeat binding factor, also known as TERF), TRF2-interacting factor RAP1 (telomeric repeat binding factor 2, interacting protein, also know as TERF2IP), single-stranded DNA binding protein POT1 (protection of telomeres 1 homolog). These factors, by interacting with other factors transiently, generate a lasso-like protective structure, the T-loop, by invasion of the 3' single-stranded G-overhang into the double stranded TTAGGG repeats (reviewed in [63, 65]). However, the percentage of telomeres that adopt a T-loop structure is low [66].

In normal cells, 50 ~ 200-bases loss of telomere will happen in each round of cell replication due to the so-called "end replication problem", which induces apoptosis when critical shortening of the telomeric DNA is reached [67, 68]. In cancerous cells, the G-overhang is extended by the activation of a reverse transcriptase enzyme, telomerase, which recognizes the single-stranded G-overhang by its own RNA template [69]. Elevated expression of telomerase has been observed in approximately 85 % of cancers [70], which can bypass normal telomere shortening and confer immortality upon a cancerous cell. Although the detailed mechanism is still unclear, targeting and stabilisation of telomeric G4s can result in impeded telomerase activity, and several possible G4-mediated mechanisms have been proposed (Figure 1.4). Thus, telomeric G4s are regarded as potential targets in anti-cancer therapies (reviewed in [71, 72]).

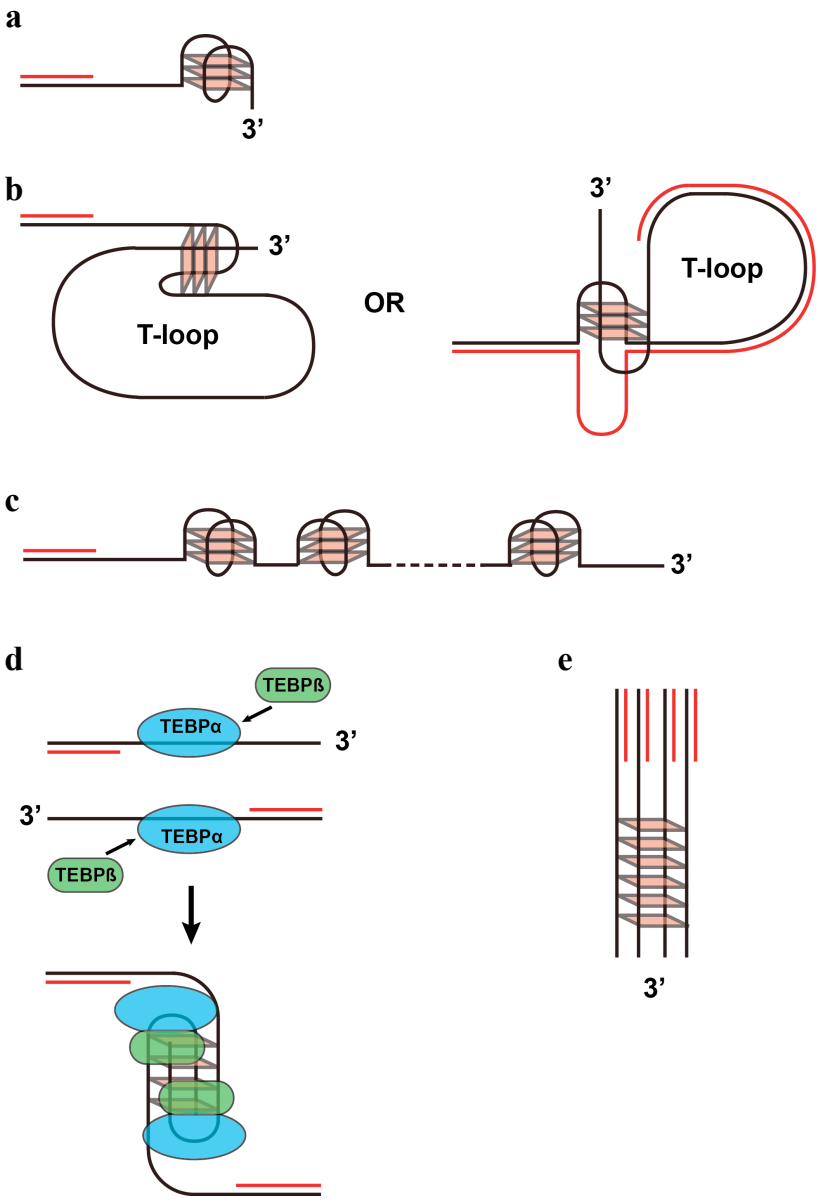


Figure 1.4: Proposed roles of G4s at the end of telomeres *in vivo*.
(a) Formation of one telomeric intramolecular G4 at the G-overhang of telomere. **(b)** Folding back of G-rich G-overhang to adopt intramolecular G4 and leave a T-loop structure. Left, a "3+1" assembly of intramolecular G4 with a single-stranded T-loop [73]; and right, a "2+2" assembly of intramolecular G4 with a duplex T-loop and a looped out C-rich strand [74]. **(c)** Consecutive formation of intramolecular G4s at the end of 3'-overhang [75]. Telomeric G4s as illustrated in **(a)** ~ **(c)** may provide protective capping for telomere ends *in vivo* (reviewed in [76, 77]). **(d)** A model of G4-mediated chromosome pairing in ciliated protozoans. TEBPα (telomere end-binding protein) specifically binds to single-stranded G-overhang, and recruits TEBPβ [78]. Binding of TEBPβ promotes the formation of an anti-parallel bimolecular G4 by pairing two telomeric G-overhangs [79]. The phosphorylation of TEBPβ was found to lead the G4 unfolding later [80]. **(e)** Formation of a tetramolecular G4 by the G-overhang of 4 homologous chromatids during meiosis [81].

Necessary for rational drug design and functional investigation, the structures of human telomeric G4 under physiological conditions have been extensively studied. In 1993, the structure of an intramolecular telomeric G4 in Na^+ solution was resolved with three stacked G-tetrads, which are connected by two lateral and one central diagonal TTA loops [82]. Due to the higher intracellular concentration of K^+, structures of telomeric G4 in K^+ are considered to be more biologically relevant than those in Na^+. In 2002, the iconic propeller-type (external loop) parallel-stranded structure of an intramolecular telomeric G4 was published in crystalline state [39], which indicates a straightforward path for telomeric G4 folding/unfolding. However, subsequent studies have revealed other diverse conformations of telomeric G4 in K^+ solution [41, 83-86]. Between 2006 and 2007, several pieces of independent work have shown that two hybrid types of G4s (hybrid-1, and hybrid-2) are predominant in K^+ solution even in presence of high concentration of Na^+ ions, rather than the propeller type determined by crystallography [42, 87-90]. The two hybrid types of G4s are distinct from one another with TTA loops adopting an external-lateral-lateral arrangement in the hybrid-1 structure, and a lateral-lateral-external arrangement in the hybrid-2 structure. Although both structures are in equilibrium with each other, the hybrid-2 is the favoured formation in solution, possibly due to its novel T:A:T triple capping at one end [88]. In addition, the structural polymorphism and equilibrium of human telomeric G4s in solution may provide a means to interact with diverse G4-binding proteins in the cell.

Although Zahler *et al.* demonstrated that telomeric G4s prevent telomerase activity in solution [91], whether telomeric G4s exist *in vivo* and what biological processes they are involved in are key questions that need to be addressed. Studies on telomeric G4s in ciliated protozoa provide the best direct evidence for the biological functions of G4s *in vivo*. Although the telomeric G-overhang is much shorter in ciliates than that in human telomeres, each macronucleus in ciliates contains up to 10^8 telomeres [92], which makes it feasible to target and observe the telomeric G4s if specific antibodies or ligands are available. Indeed, by using a single-chain antibody fragment (scFv) probe specific to the anti-parallel G4 from the *Stylonychia* telomeric G-overhang, Lipps *et al*. found by immunostaining that the anti-parallel G4 is only absent in the replication bands [93]. The presence of telomeric G4s at the end of non-replicating chromosomes, but not in replicating ones implied that there might be a certain

mechanism to regulate folding and unfolding of these telomeric G4s. Another circumstantial evidence also indicated the function of telomeric G4 in ciliated protozoans. For example, nanochromosomes extracted from macronuclei were linked with each other via their ends [94] in solution studies, showing telomeric G-overhang can adopt anti-parallel bimolecular G4s [95]. Thus, it is very likely that these telomeric G4s are involved in association of macronuclear DNA in ciliates, leading to long DNA aggregation. Furthermore, a group of telomeric G4-binding proteins were identified, such as βTBP/TEBPβ (telomere end-binding protein β) [79], TGP1/3 (tetrahymena G-DNA binding protein 1/3) [96, 97], and TERT (Telomerase reverse transcriptase) [98], which also suggested the biological functions of telomeric G4s *in vivo*.

Due to limited number of telomeres, evidence showing the presence of telomeric G4s in other organisms, especially in human, is less direct than those in ciliates. Some indirect evidence comes from the response of cancerous cells by applying certain telomeric G4-binding ligands. The folding of telomeric G4 further stabilised by ligands ensures it is inaccessible to telomere binding proteins, such as POT1, and thus disrupts the formation of shelterin complex. For example, a group of telomeric G4-stabilising ligands can cause telomere dysfunction and DNA-damage response when introduced into mouse or human cancerous cells [99-101]. These effects seem to be telomeric G4-mediated, since the ligands delocalized the single-stranded G-overhang binding protein POT1, but not the double-stranded binding protein TRF2. Another in solution experiment proposing the ability of POT1 to disrupt telomeric G4 [102] further supported the telomeric G4-mediated effects caused by those ligands. As vague as the actual biological functions of telomeric G4 in living cells are, the corresponding anti-cancer effects related to those ligands are also difficult to elucidate. Although a series of telomeric G4-binding ligands show promising tumour suppressing effect (reviewed in [71]), the actual mechanism is still far from clear. It is believed that the anti-cancer effects of G4-binding ligands are complicated, which may relate to a combination of G4-mediated and non-G4-mediated processes, such as telomere capping, disrupting shelterin complex, causing DNA-damage response, dislocating the telomerase, and inhibit telomere amplification [103].

1.2.4 G4s in promoters of other genes

The anti-cancer potential of human telomeric G4s sparked the interests in G4s outside human telomeres, which led to the discovery of a group of other G4s with anti-cancer potential with regard to the regulation of expression of cancer-related genes. Gene transcription and translation can be regulated at many different points. Traditionally, these include: the dynamic of higher order chromatin structure in response to methylation, acetylation or other mechanisms that affect nucleosome-packing along the DNA duplex; the role of tissue- or stage-specific TFs in the specific regulation of a gene's transcription; regulation by microRNAs; alternative mRNA splicing, and by mechanisms that modify DNA directly such as methylation at CpG islands. It is widely accepted now that the presence and stabilisation of G4s in gene promoters may also act as a novel mechanism in regulating gene transcription.

MYC. As the most well-studied promoter G4 model, substantial evidence of the existence of G4s in the promoter of *MYC* has been gained, which strongly supports the therapeutic potential of targeting the promoter G4s in cancer therapy. MYC is an important oncoprotein that controls cell proliferation and the induction of apoptosis [104], and the deregulated over-expression of MYC is associated with the initiation and maintenance of around 80 % of human solid tumours (reviewed in [105]). The activation of *MYC* gene transcription is related to a group of nuclease-hypersensitivity elements (NHE) in the promoter region, in which a region called $NHEIII_1$ accounts for up to 95 % of total *MYC* transcription [106]. The 33nt $NHEIII_1$ region contains 6 G-tracts with different length, which can spontaneously adopt diverse G4 structures in solution [107, 108]. In 2002, two models of intramolecular anti-parallel G4s, the "basket" and "chair" types, formed by the five G-tracts in the 5' of $NHEIII_1$ (Pu27) were proposed by Hurley's group. Of these two, the chair type is more kinetically favoured and biologically relevant compared with the other form [28]. By including different regions in the $NHEIII_1$, separate NMR studies have subsequently revealed propeller-type of parallel structures in solution, with loop arrangements 1:2:1 (Myc-2345 [109]), 1:6:1 (Myc-1245 [109]), 1:2:1 (MYC22-G14T/G23T [29]), and 1:2:1 (Myc-1234 [110]). In all these structural studies, some parallel/anti-parallel G4s are very stable in K^+ solution (T_m > 75ºC), possibly due to the capping from 3' flanking sequence [29] or different base conformations of the single nucleotide loops [110]. Besides its structures, the regulatory role of MYC G4 as a silencer has also been

clarified. In an *in vitro* study, the expression of MYC was down-regulated by the cationic porphyrin TMPyP4, which binds to and stabilises the MYC G4s in solution [111]. A molecular mechanism for transcription activation by the MYC G4 has been proposed which involves binding of NM23-H2 (nucleoside diphosphate kinase) to the G4-forming motif, unfolding of the MYC G4 within the protein-DNA complex, and recruiting other TFs [28]. Furthermore, *in vitro* studies also demonstrated that the transcription-induced negative supercoiling could further promote the formation of MYC G4s [112]. Recently, a multifunctional phosphorprotein, nucleolin, and CNBP (cellular nucleic acid binding protein) were found to bind to and stabilise the MYC G4s [113, 114], which further provided circumstantial evidence suggesting the function of MYC G4s in living cells. The *MYC* promoter G4 in the $NHEIII_1$ region is believed to be a promising target in repressing the over-expression of MYC in cancer therapy (reviewed in [115]], and the proposed regulatory mechanism of the *MYC* promoter G4 is described in detail in Figure 1.5).

KIT. At the same time, G4s found in another proto-oncogene, *KIT*, were also investigated as potential anti-cancer drug targets. The product of the *KIT* proto-oncogene is a type III receptor tyrosine kinase whose activation regulates apoptosis, cell differentiation, proliferation, chemotaxis, and cell adhesion [116]. Over-expression and pathologic activation of KIT through gain-of-function mutations are associated with testicular seminoma, mast cell disease, melanoma, acute myeloid leukemia, piebaldism, and GIST (gastrointestinal stromal tumors) (reviewed in [116, 117]). Particularly, GIST is principally driven by constitutive activation of KIT, hence it is considered to be a major molecular target in GIST therapy [118, 119]. However, cell responses of kinase inhibitors targeting KIT in cancer therapies can vary due to different KIT mutations in different cancers. Thus, G4-forming motifs identified in the promoter of *KIT* have been considered to be an alternative target in KIT inhibition. Two G4-forming motifs were identified in the core promoter of *KIT*

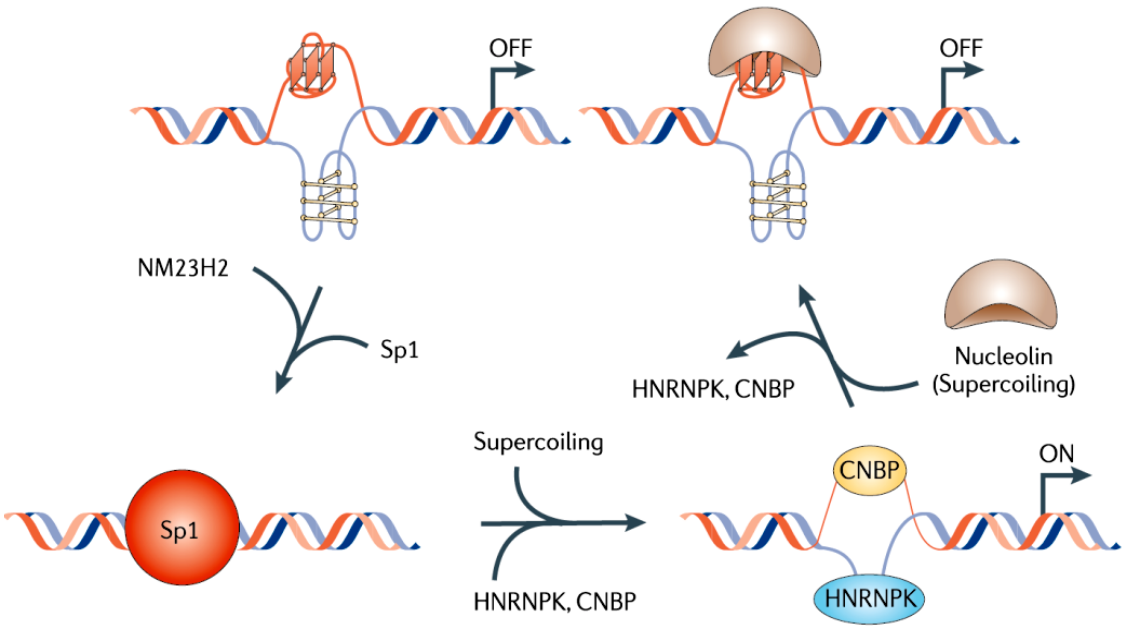


Figure 1.5: Proposed regulatory mechanism of a G4 in the NHE III_1 region of *MYC* promoter.
The formation of G4 can impede the assembly of transcription machinery, and hence repress downstream *MYC* gene transcription [28]. The binding of nucleolin to the G4 could further stabilise the folded structure [113]. The binding of NM23H2 (the non-metastasis 23 protein family) to the G4 would lead to resolution of the folded G4 into duplex DNA, and induce transcription activation by recruiting ancillary transcription activators such as Sp1 [120]. Binding of heterogeneous nuclear ribonucleoprotein K (HNRNPK) and cellular nucleic acid binding protein (CNBP) to the single-stranded DNA could also activate downstream gene transcription [28, 114]. The figure was taken from Balasubramanian *et al.*, 2011 [121].

and designated KIT1 and KIT2. These are located between position -87 and -109 and between -140 and -160, respectively, relative to the TSS (transcription start site) of *KIT*, with a critical Sp1 (specificity protein 1) binding site situated between them [122]. Both KIT1 and KIT2 can adopt stable G4s in solution, and their structures have been resolved. KIT1 G4 exhibits a unique folding topology in which an isolated guanine in the CGCT sequence folds back in G-tetrad core formation, although four G-tracts are present [30]. Furthermore, sequence variations intervening between G-tracts (loop regions) dramatically disrupted the formation of G4 [123], and the sequence of KIT1 was found to be unique in the human genome [124]. Together with its unprecedented structure, the KIT1 G4 could be used as a unique and specific anti-cancer target for KIT in the human genome. In KIT2, a mixture of more conventional propeller-type parallel structures were resolved [125]. The heterogeneous

conformations of the KIT2 G4 with varied sequence and length of loop 2 were identified in a study combining NMR, CD, and UV-Vis (ultraviolet-visible) spectroscopy methods [126]. In the same study, biophysical properties of the KIT2 G4 were found to be conserved across human, chimpanzee, rat, and mouse [126]. In a more recent NMR study, beside a conventional parallel monomeric G4, a novel dimeric G4 was resolved with two parallel monomeric G4s stacked vertically and mediated by a sandwiched A•A non-canonical pair [127]. Although the actual regulatory mechanism of *KIT* promoter G4s is still not clear, accumulating indirect evidence has strongly supported the biological relevance of KIT G4s *in vivo*. For example, *in vitro* experiments shown that several G4 binding ligands that bind to and stabilise both KIT1 and KIT2 G4s can reduce *KIT* mRNA in different cell models [128, 129]. Another biophysical study by using single-molecule fluorescence resonance energy transfer (smFRET) spectroscopy has demonstrated that the KIT2 G4 is able to fold non-duplex structures within an extended DNA duplex [130], which further supports the biological relevance of KIT G4s *in vivo*.

Besides *MYC* and *KIT*, promoter G4s with potential biological functions have been identified in other genes regarded as six hallmarks of cancer.

KRAS. A single amino acid substitution of the proto-oncogene *KRAS* is responsible for the gain-of-function mutation possessing oncogenic potential, which results in various malignancies including lung adenocarcinoma, mucinous adenoma, ductal carcinoma of the pancreas and colorectal carcinoma [131]. Similar to *MYC*, in the critical NHE region close to the TSS of *KRAS*, a G4-forming motif was identified, which can adopt stable parallel and anti-parallel G4s in solution [36, 132]. In the same study, three nuclear proteins that recognize the KRAS G4-forming sequence were identified, which are PARP-1 (poly (ADP-ribose) polymerase 1), Ku70/86 (ATP-dependent DNA helicase 2, subunit 1/2) and hnRNP A1 (heterogeneous nuclear ribonucleoprotein A1). Subsequently, the hnRNP A1 and its derivative Up1 (a proteolitic portion of A1 retaining binding activity) were found to be able to unwind the parallel G4 to ssDNA [132-134], and PARP-1 has been shown to be catalytically activated upon binding to the KIT and telomeric G4s *in vitro* with high affinity (K_d = 65 nM) [135]. Several pieces of indirect evidence also demonstrated that the KRAS G4 acts as a silencer. For example, the stabilisation of KRAS G4 by TMPyP4 can suppress the transcription of *KRAS* (decreased mRNA level) in both human and

mouse cell models [132, 136]. Furthermore, oligonucleotides that mimic the G4 formed in another region of NHE in the *KRAS* promoter can compete with a DNA-protein complex formed between the *KRAS* NHE region and a nuclear extract from pancreatic cancer cells, and covalent modification of KRAS G4 that stabilises the G4 can strongly inhibit the proliferation of the Panc-1 cell line [137].

PDGF-A. Due to their critical roles in angiogenesis, aberrant expression of PDGFs (platelet-derived growth factors) is associated with various abnormalities, especially in cancer development and progression (reviewed in [138]). In the proximal promoter region of the human *PDGF-A* gene, a NHE region was found to control 80 % of basal promoter activity, in which a GC-rich region was identified [139]. The G-rich strand contains five contiguous G-tracts with at least four guanines in each run, and each G-tract is separated by only one nucleotide, which are believed to be able to adopt very stable G4 structures in solution. Indeed, Hurley's group observed the formation of several intramolecular propeller-type parallel G4s within both the single-stranded and the duplex DNA from the *PDGF-A* NHE region, even at very high temperature (up to 100 °C) [140]. In the same study, the authors also demonstrated that the TMPyP4, which stabilises one major form of PDGF-A G4, reduced the transcription activity of *PDGF-A* by 40 % in human pancreatic cancer MIA PaCa-2 cells at 10 μM concentration [140].

RET. The *RET* proto-oncogene encodes a receptor tyrosine kinase. RET gain-of-function mutations are involved in the development of various types of human cancer (reviewed in [141]). In the two adjacent GC boxes involved in determining the basal promoter activity of *RET* [142], five G-tracts has been identified. By analyzing an oligo containing four G-tracts from the 3' of GC boxes, two stable (T_m ~ 80ºC) parallel intramolecular G4s composed of three G-tetrads were revealed, while the C-rich complementary strand adopted a stable i-motif structure [143]. Although the actual structure is still unresolved, based on biophysical studies, a stable low energy model of this GC-rich region has been proposed, in which a parallel G4 (minimum loop 1:3:1), an i-motif (minimum loop 2:3:2), and adjacent duplex containing Sp1 binding sites coexist stably [143].

VEGF. Over-expression of VEGF, which stimulates vasculogenesis and angiogenesis, contributes to diverse diseases, including various types of cancer and

vascular diseases (reviewed in [144]). Five contiguous G-tracts were identified in the pPu/pPy (polypurine/polypyrimidine) tract in the proximal promoter of the human *VEGF-A* gene. Formation of propeller-type parallel intramolecular G4s from part of this region d($G_4CG_3C_2G_5CG_4$) has been determined in solution in the presence of K^+ ions and G4-binding ligands by a variety of methods [145, 146]. Interestingly, this G4-forming region seems to be involved in determining the hypersensitivity of the pPu/pPy tract toward DNase I (deoxyribonuclease I) or S1 nuclease *in vitro* [145]. DMS (dimethyl sulfate) footprinting experiments carried out in solution and *in vitro* both demonstrated that the guanines involved in G4 formation are significantly protected from DMS as compared to flanking guanines [147]. It is also worthwhile to note that this protection is only apparent in the supercoiled plasmid but not in the linearized plasmid, which indicates that negative supercoiling is required for the unwinding of the pPu/pPy tract [147, 148]. Furthermore, suppressed *VEGF-A* transcription has been observed in cancer cell lines when incubated with G4-binding ligands, especially the Se2SAP [146].

RB1. The retinoblastoma protein (RB) is a tumor suppressor protein encoded by *RB1* (retinoblastoma 1) that is normally dysfunctional in most types of cancer [149]. A potential G4-forming motif in the 5' end of the *RB1* gene was discovered in human and mouse in the early 1990s [150]. A decade later, an anti-parallel G4 derived from this region was revealed by using natural and chemically modified oligos [151-153].

BCL-2. BCL-2, as one of the key regulators of apoptosis, functions as an inhibitor of apoptosis and is involved in a number of cancers, including melanoma, breast, prostate, and lung carcinomas (reviewed in [154]). In the *BCL-2* promoter, a GC-rich region was identified in an NHE region located between -1,490 and -1,451 upstream of the P1 promoter of *BCL-2* [155]. In a systematic research carried out by Hurley's group, the G-rich strand of this region, d($G_3CGAG_3CG_5AAG_2AG_3CGCG_3CG_4$), was found to be able to adopt intramolecular G4s with varied structures, in which three major population of G4s are formed at the 5' end, middle, and 3' end of the strand (named as 5'G4, MidG4, and 3'G4, respectively) [156]. In the same study, as determined by DMS footprinting and polymerase stop assays, the MidG4 is the most stable G4 structure in solution, with different loop isomers containing different lengths in loop2 and loop3 [156]. A dual mutant that mimics the most stable MidG4 (G15 to T, and G16 to T) was subject to subsequent NMR study, and a mixed

parallel/anti-parallel G4 was resolved with two lateral, and one external loops (2:7:1) [25, 26]. The interaction of BCL-2 G4s and TMPyP4 has been studied in solution as well [157]. Although the actual biological function of BCL-2 G4s in living cells is still unclear, there is some interesting evidence showing potential roles of G4 in BCL-2 dependent pathways. For example, *BCL-2* mRNA levels were reduced in the presence of a G4-binding ligand, triazine derivative 12459, and correspondingly, apoptosis was induced in A549 cells via mitochondrion-related pathways [158].

HIF-1α. HIF-1 (Hypoxia-inducible factor 1) is a dimer from HIF-1α and HIF-1β, which plays an essential role in cellular and systematic response to hypoxia [159]. It is also an important TF in regulating VEGF expression, and its over-expression is related to many types of cancer [160, 161]. The G-rich strand in the pPu/pPy region in *HIF-1α* promoter (-65 bp to -85 bp) can adopt parallel intramolecular G4 structures in solution with K^+ ions [162]. Although the actual structures and biological roles of this G4-forming sequence in the promoter of *HIF-1α* is still unclear, an interesting example may indicate the general roles of G4s in regulating HIF-1α and HIF-1-regulated proteins. A recent study revealed that two oligos, JG243 and JG244, which adopt parallel intramolecular G4s in solution, can decrease the level of HIF-1α and the related HIF-2α by selectively binding to them (IC_{50} < 2 μM), which then inhibit the expression of HIF-1-regulated proteins such as VEGF, BCL-2, and BCL-X_L, and dramatically repress the growth rates of several cancerous cell lines [163]. Although not directly related to the G4 within the promoter of *HIF-1α*, the high abundance of G4-forming sequences in HIF-1-regulated genes and other proto-oncogenes make it plausible to assume that a group of TFs such as HIF-1 may target these promoter G4s in order to regulate downstream gene expression in cancer-related pathways.

Indeed, more and more proteins and TFs have been found to be able to interact with G4s, which further strengthens the biological relevance of G4s *in vivo*. I discuss some examples in the next section.

1.2.5 G4-interacting proteins

A variety of G4-interacting proteins have been identified in diverse organisms. The existence of these natural proteins, which preferentially interact with G4 DNA rather

than other single-stranded forms or duplex DNA, also indicates the biological importance and prevalence of these fascinating secondary structures in living cells. These G4-interacting proteins are classified into several groups based on their distinct functions. The first group is the G4-binding/stabilising proteins, which can bind to G4 structures specifically and stabilise them upon binding. Two examples are the HIV-1 zinc finger nucleocapsid protein and the uqTBP25 (unimolecular quadruplex telomere-binding protein 25). HIV-1 zinc finger nucleocapsid protein is involved in the compact assembly of the viral nucleocapsid [164], while uqTBP25 was identified in rat hepatocytes, and potentially is involved in telomeric DNA interactions [165]. Proteins that promote formation of intra- or intermolecular G4 structures are classified as the second group, and these include the yeast proteins RAP1 (repressor activator protein 1) [166], Hop1 (homolog of presenilin 1) [167], and the human TOPOI (topoisomerase I) [168]. Another interesting example in this group is an intramolecular G4 aptamer of human thrombin, which can bind to the protein in the nanomolar range and inhibit activity of the protein *in vitro* [169]. Although it is unclear whether this interaction occurs naturally in living cells, it provides an attractive strategy for modifying certain biological functions by interfering with G4 elements. G4-unwinding helicases and destabilising proteins are the other two types of G4-interacting proteins, which further suggests the dynamic properties of G4 structures within living cells. Unlike G4-destabilising proteins, which unwind G4 DNA in a non-catalytic way, all unwinding helicases are energy-dependent, requiring hydrolysis of nucleotide triphosphate. Most of these unwinding helicases and destabilising proteins are involved in maintaining the stability of genome. For example, the family of RecQ helicases which have been identified in many organisms from *Eschericia coli* to human can effectively unwind G4 structures in the Fragile X Syndrome repeat sequence, an X chromosome expansion of the trinucleotide repeat CGG, while malfunctions of WRN (Werner syndrome) and BLM (Bloom syndrome) proteins, two members of the human RecQ family, are related with Werner's and Bloom's syndromes, respectively, due to chromosome instability [170, 171]. In budding yeast, Pif1 DNA helicase is involved in maintaining stability of the human G-rich minisatellite CEB1 in both nuclear and mitochondrial genomes [172]. More recently, the human PIF1 DNA helicase also has been identified to preferentially bind and unwind synthetic G4s [173]. Also, many members of hnRNP family have been identified as G4-destabilising proteins, such as CBF-A (CarG-box binding protein A,

[174]) and UP1 [175, 176], both potentially involved in telomere maintenance. The last group is the G4-selective nucleases, which incise DNA within or close to G4 structures. The yeast *Saccharomyces cerevisiae* KEM1 protein, for example, can cleave G4s [177], and is involved in cellular senescence and telomere shortening [178].

1.3 Bioinformatics studies on G4s

Since G4s have been revealed in many specific regions in the genome, such as telomeres and the promoters of several proto-oncogenes, bioinformatics approaches are taken to investigate whether G4 DNA is an evolutionarily conserved and widespread mechanism for the regulation of diverse biological processes. Meanwhile, the progress of biophysical studies has also enriched our understanding towards the folding pattern of G4s, which underlies the basis of G4-predicting algorithms.

Early bioinformatics studies mainly focused on the sequence composition of potential intramolecular G4s in the human genome. In 2005, two different groups independently examined the human genome by using different G4-predicting algorithms. Both algorithms revealed around 370,000 potential G4-forming motifs in the human genome, although not all of these are predicted to adopt stable G4 conformations [179, 180]. Interestingly, they both claimed that the putative loops involved in G4 formation exhibit significant non-random length and sequence, which may indicate that certain loop length and sequence combination are more favoured for the formation of G4s *in vivo* [179, 180]. For example, the conserved sequence CCGTG was found to be the most commonly occurring sequence in the first loop of G4s [179]. It is also notable that the distribution of these G4s shows non-random properties in the human genome as revealed in several different studies. G4s are under-represented in the coding strand of exonic regions of genes, indicating that the formation of G4 structures within mRNA might be detrimental and therefore, from an evolutionary viewpoint, unlikely to be tolerated [180]. Also, more than 40 % of gene promoters contain one or more potential G4-forming motifs. These G4-forming motifs are particularly enriched in promoter regions, and correlate strongly with the location of nuclease hypersensitive sites [17]. Further study found the biased

distribution of G4s in promoters, which peaked at around -50 bp relative to TSS [181]. Furthermore, accumulation of G4s has been observed in the 5'-ends of 5'-UTRs (untranslated region), within regions immediately downstream of the 3'-ends of genes [181], and within the first intron in the human genome [182]. Distribution of G4s in the promoters of protein-coding genes was analyzed in other species, including prokaryotes [183], yeast [184], chimpanzee [185], chicken [186], and genomes of other animals [185, 186]. The biased distribution of G4s, with a major peak centred around -50 bp and a minor peak between +150 bp and +300 bp, was identified in most warm-blooded but not in cold-blooded animals [186]. Interestingly, G4 abundance differences seem to correlate with different temporally grouped genes in the human genome, in which G4 abundance is higher in evolutionarily older genes as compared to younger genes [187]. All these findings have been regarded therefore as important evidence indicating the biological potential of G4s in the human genome.

Besides the preferred positional distribution in the human genome, G4s also show different distribution based upon gene functions. Eddy *et al.* have analyzed all genes available in the human RefSeq database, and revealed the correlation between G4P (G4 formation potentials) and gene functions. Particularly, high G4P values were identified in proto-oncogenes, while low G4P values exist in tumor suppressor genes [188]. Perhaps the area that is most intriguing though concerns the potential for G4s to regulate downstream gene expression through interacting with TFs or other sequence-specific DNA-binding proteins. By combining the human genome G4 search results and gene expression data from the Gene Atlas dataset, Du *et al.* have shown that the abundance of G4s in the proximal region downstream to the TSS (TSS ~ +500 bp) are positively correlated with the gene expression level, and the asymmetric distribution of G4s in these regions is significant, where the coding strands contain much more G4s than the template strands [189]. The authors proposed a mechanism that abundant G4 on the coding strand may facilitate the separation of two strands during transcription, while the asymmetric distribution of G4 on the template strand reduces the probability to block the progression of RNA polymerase complex [189]. However, in another study combining *in silico* and experimental approaches, transcription levels of most genes with conserved G4s around TSS were down-regulated when treat with high concentration (100 μM) of TMPyP4 in HeLa S3 and A549 cells [185]. Although not enough evidence could support that this is mainly

caused by direct interaction between G4s and the ligand, the authors postulated that the ligand induced formation of G4s that hinder the progression of the RNA polymerase II transcription bubble and abort transcription. A recent bioinformatics study indicated that G4 motifs are enriched within several types of gene regulatory regions, including binding sites for TFs, nuclease hypersensitivity sites, promoter-linked CpG islands, insulator regions between chromosomal domains and so forth. The frequency of G4 DNA within these regions is significantly higher as compared to intergenic regions, introns or within pseudogenes [190]. In a more recent study, zinc-finger TF binding sites were revealed to be associated with G4-forming sequences, which is of potential interest given the large number of zinc finger protein genes within the genome [191].

These observations further supported the rationale that G4 DNA elements may act as molecular switches in concert with DNA-binding proteins to control transcription and to respond to changes in chromatin structure or signalling mechanisms. Taken together, the occurrence and distribution of G4-forming sequences could be far more than random phenomena in the context of evolution.

1.4 Methodologies for biophysical characterisation of G4 DNA*

Biophysical properties of G4s in solution are critical to estimate and understand their corresponding biological significance *in vitro* and *in vivo*. Thus, a group of conventional and novel methods have been used to characterise the formation, thermodynamic properties, and kinetic behaviours of G4s in solution.

1.4.1 EMSA

EMSA (electrophoretic mobility shift assays) of G4-forming oligos in native polyacrylamide gels has been used to detect the formation of G4 structures for over 20 years [18]. In native EMSA experiments, the mobility of DNA oligo is mainly determined by its size and charge. When two oligos have the same length, and therefore possessing similar charges, the relative mobility between them is dependent on their shapes (hydrodynamic diameter). However, in the presence of a DNA

* Part of this Section was adopted from my previous publication [6].

denaturing agent (urea or, less frequently, formamide), the mobility of denatured DNA is determined by its length, and almost completely independent on its base composition and sequence [192]. Thus, as compared to their linear counterparts with the same length, G4 structures migrate much faster in native gels due to their more compact structures. This technique remains the first step to prove the presence of G4 under native conditions. This technique can also be used to separate and purify intermolecular G4 from intramolecular G4 structures for subsequent experiments, for example DMS footprinting.

1.4.2 DMS footprinting

Under alkali conditions, DMS reacts with guanine to methylate it at the N7 position, and subsequent piperidine treatment breaks the N9 glycosidic bond, resulting in a break to the DNA strand at the guanine position. This method was first introduced by Maxam and Gilbert as part of their chemical DNA sequencing method [193], and is now widely used in revealing DNA-protein interactions, as well as in probing RNA structures [194]. Because the N7 of guanines in G-tetrads are involved in Hoogsteen bonding with adjacent guanine bases, these methylation sites are therefore inaccessible to DMS. However, if guanines are involved in normal GC base pairing, or less commonly, GG base pairing, the N7 position is still reactive with DMS. Subsequent denaturing PAGE (polyacrylamide gel electrophoresis) alongside a DNA sequencing reaction can identify all guanines that are protected from or still reactive with DMS, and thus resolve guanines that are directly involved in G4 formation. Thus, DMS footprinting experiments can provide direct evidence to confirm the formation of G4, and reveal guanines involved in G4 formation, as has been shown in determining the folding pattern of *HIF-1α* promoter G4s, for example [162]. In some cases, DMS footprinting can also be used to probe different G4 structures based on their different loop arrangements. For example, DMS footprinting of a 28-mer G-rich strand from the NHE of mouse *Kras* promoter revealed several G4 structures with different loop arrangements (loop: 1:1:9, and 1:1:5) [132]. In Pu27 (a 27mer G-rich strand from human *MYC* $NHEIII_1$ region), different guanine protection patterns from two intramolecular G4s with different mobility in native PAGE supported the proposed “basket” and “chair” models of G4 structures [28]. DMS footprinting can

also reveal G4 formation in plasmid in solution, or even in promoter regions *in vitro*. For example, DMS footprinting experiments on a G4-forming sequence from human *VEGF* promoter have been carried out at three different levels: a single-stranded oligo, a duplex contained in a supercoiled plasmid and in cells, where guanines involved in G4 formation but not those in flanking regions, or in the loops, are all strongly protected against DMS methylation [147, 148].

1.4.3 Circular dichroism

Circular dichroism (CD) spectroscopy is also a routinely used method to investigate the formation and to distinguish parallel/anti-parallel and hybrid structures of G4 DNA empirically. Although the relationship between G4 folding pattern and CD spectrum shape is yet to be clarified, CD spectroscopy is widely used as an empirical method, and CD spectra signatures have been well characterised for several G4 structures in solution and compared with NMR or X-ray crystallography data. For example, CD spectra with a trough at 240nm, a major peak around 260nm, and a minor peak at about 212 nm represent intramolecular parallel G4 structures such as the G4 from the G-rich DNA oligomer (MYC Pu27-Mer) in *MYC* $NHEIII_1$ region [195], whose structure has been resolved by NMR [109]. Similar CD spectra with an additional positive shoulder between 280 nm and 300 nm represent a hybrid parallel/anti-parallel G4 structure, such as the G4 of BCL-2 oligomer (Pu39WT) resolved by CD [156] and NMR [25]. The chair-type anti-parallel G4 structure has a CD spectrum characterised by a major positive band around 290 nm, a minor positive band at 246 nm, and a negative band at 265 nm, such as the G4 structure of TBA (thrombin binding aptamer) [196, 197], whose structure is resolved by X-ray crystallography and NMR [198]. Together with CD melting and UV absorbance at 295nm, these techniques can be used to obtain many of the key stability parameters for G4 structures, such as T_m, ΔH, and ΔG^0, especially in the human telomeric G4s [89, 199-201].

1.4.4 Single molecule FRET

The smFRET (fluorescence resonance energy transfer) technique has proved to be a unique approach for resolving the structural heterogeneity of DNA secondary structures such as G4 based on the fluorescent signals from individual immobilized or freely diffusing molecules. FRET is the non-radiative dipole-dipole energy transfer from a donor fluorophore to an acceptor fluorophore when they come within 10 ~ 100 Å of each other, and the FRET efficiency depends on the fluorophores' spectral overlap, separation and orientation [202]. The relationship between efficiency of energy transfer and distance between donor and acceptor are described in Eq. 1.1 and 1.2 [203].

$$E_{\mathrm{FRET}} = \frac{1}{1+(\frac{R}{R_0})^6} \qquad \text{Eq. 1.1}$$

where R is the donor-acceptor distance, and R_0 is separation at which 50 % of the excitation energy is transferred to the acceptor and is known as the Förster distance, which can be calculated as described in Eq. 1.2.

$$R_0 = 9.79 \times 10^3 \sqrt[6]{J\varphi^{\mathrm{D}}\kappa^2 n^{-4}} \qquad \text{Eq. 1.2}$$

where J is the overlap integral, φ^{D} is the quantum yield of the donor in the absence of the acceptor, κ is the orientation factor, and n is the reflective index of surrounding media.

smFRET by detecting freely diffusing molecules with a confocal microscope is the most commonly used single molecule technique in studying G4 structural heterogeneity, thermodynamics, and kinetics in solution. Generally, the donor and acceptor fluorescent intensities (I_{d}, and I_{a}) will be collected from individual molecules, and their FRET efficiencies of individual molecules can be derived from Eq. 1.3

$$E_{\mathrm{FRET}} = \frac{I_{\mathrm{a}}}{I_{\mathrm{a}} + \gamma I_{\mathrm{d}}} \qquad \text{Eq. 1.3}$$

where γ is a correction factor that accounts differential detection efficiencies and quantum yields of the two fluorophores, which is normally assumed to be 1 in calculation of apparent FRET efficiency (E_{app}) [204] or "proximity ratio" (P) [205].

Subsequent statistical analyses applied to the FRET efficiencies of thousands of different molecules can reveal the structural complexity and heterogeneity, which are normally averaged in conventional ensemble measurements. For G4 analysis, short single-stranded oligonucleotides are used which are labelled with spatially separated donor and acceptor fluorophores. Upon folding into G4 structures, the donor and acceptor fluorophores are brought into close juxtaposition. The FRET efficiency histograms can reveal the conformational distribution of G4 structures under different conditions and allow the observer to follow their folding/unfolding process in real-time (mechanism of smFRET is illustrated in Figure 1.6, single molecule techniques are reviewed generally in [206, 207], and specifically in G4 studies in [208, 209]). The most widely used smFRET instrument is composed of a laser emitter, a confocal microscope, serial optics, and high sensitivity APDs (avalanche photodiodes). In most commonly used one-photon excitation confocal microscopy setups, a collimated laser beam is reflected by a dichroic mirror and then passed through a confocal aperture (pinhole) in the microscope. The beam is then focused by the objective to a spot at the confocal plane in the highly diluted solution of interest (normally around 100 pM). The focused laser beam forms an excitation volume around the confocal plane, where fluorophores can be excited. Then, in the collection path, the confocal aperture rejects most of the fluorescent signals from the excitation volume, in order to further reduce the actual observation volume (blue area in Figure 1.6 [206], right, normally around 1 fl) to make sure only one molecule is in the actual observation volume in each measurement. The incident laser beam will excite the donor/acceptor-labeled single molecule freely diffused into the actual observation volume, and the correlated donor/acceptor fluorescence is collected by the same objective. The relative intensities of the donor/acceptor fluorescence are dependent on the inter-dye separation. Then, collected fluorescence is further separated by bandpass filters into donor and acceptor fluorescence based on the wavelength, and recorded by APDs. In each measurement, due to the small size of the actual observation volume, low concentration of molecules in solution, and the short measuring time (normally 1 millisecond), only one molecule would diffuse into the actual observation volume and be monitored.

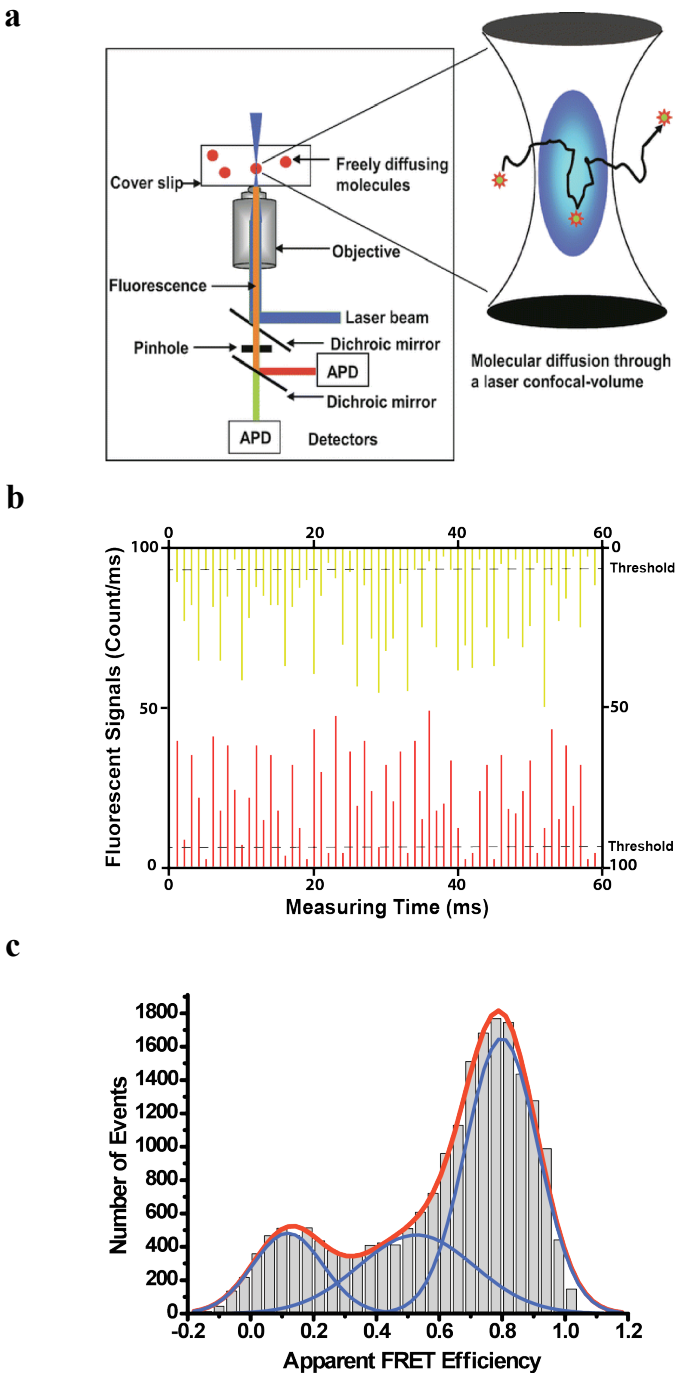


Figure 1.6: Schematic representation of the single molecule FRET technique.
(a) A typical experimental setup for single molecule FRET measurement, left; close-up of the excitation and observation volume (also called confocal volume), right [206]; **(b)** A typical donor-acceptor (green and red, respectively) fluorescent signals trajectory observed in single molecule FRET measurement. Each spike corresponds to the donor and acceptor signals from an individual molecule. When donor and acceptor signals both exceed threshold (dashed line), it will be recorded as a valid FRET event and FRET efficiency from an individual molecule will then be calculated. **(c)** A typical single molecule FRET histogram revealing different subpopulations of G4 structures based on different donor-acceptor distance.

G4-forming motifs from human telomere and *KIT* promoter have been thoroughly investigated by smFRET. For example, Ying *et al.* identified two stable folding G4 structures with high FRET efficiencies in TMR/Cy5-labeled human telomeric G4-forming sequences [205], and subsequent unfolding experiments determined the unfolding thermodynamics of telomeric G4 such as activation enthalpy and entropy of these G4s in the presence of complimentary C-rich oligos [205] or PNA (peptide nucleic acid) trap [210]. Subsequently, Green *et al.* measured the enthalpy and entropy changes of human telomeric G4 induced by a G4-binding ligand, HC (hemicyanine tetrapeptide conjugate), at single molecule level [211]. By using a modified labelling model, Lee *et al.* further revealed that in human telomeric G4-forming sequence, there are two folded states and one unfolded state, all of which contain one short-lived and one long-lived species as being detected by single molecule FRET [86]. Besides human telomeric G4, *KIT* promoter G4s were also subjected to single molecule FRET investigation. For example, differences of unfolding dynamics and kinetics of KIT1 and KIT2 were measured and compared by Fegan *et al.*, in which KIT1 was found to be less kinetically stable but more thermodynamically stable as compared to KIT2 [212].

1.4.5 Dual luciferase reporter assays

Since more and more interest is focused on the transcriptional and translational roles of G4s in the gene promoter and within the 5' UTR of the mRNA, the dual luciferase reporter assay is now widely used in the study of biological consequences of G4 DNA. By putting the firefly luciferase reporter gene under the control of the *MYC* promoter, Hurley's group confirmed that a nuclease hypersensitive site within a negative regulatory element in the *MYC* promoter adopts a G4 structure. As discussed earlier, a nuclease hypersensitive site (NHE) is a short region of chromatin that can be cleaved by many nucleases due to having a loose nucleosomal structure compared to compact chromatin. A single nucleotide (G to A) substitution within this region of the *MYC* promoter could release the transcriptional repression by 3-fold by disrupting the G4 structure, leading to enhanced promoter-luciferase activity [28]. In contrast, the G4 could be stabilised in the presence of the cationic porphyrin TMPyP4, leading to further measurable repression of the linked luciferase reporter. Similar methods were

used by Balasubramanian's group to demonstrate the translation-repressing effects of an RNA G4 in the 5'-UTR of the *NRAS* (neuroblastoma RAS viral oncogene homolog) proto-oncogene [213]. Such methods thus can be used to demonstrate transcription regulatory roles of G4s, by site-directed mutagenesis of key guanines residues, and to evaluate the effects of G4-binding ligands in living cells [128].

1.5 G4-binding ligands

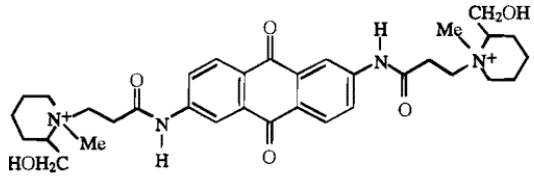
After a decade of investigation about its biological functions *in vitro* and *in vivo* since the discovery of telomeric G4, G4 DNA has attracted exceptional attention from nucleic acid chemists to design and synthesise novel G4-binding ligands that can interact efficiently and selectively with this peculiar DNA structure. So far, most G4-binding ligands are intensively researched as potential anti-cancer agents by targeting cancer-related G4s, especially the human telomeric G4. Compared to the duplex-binding ligands, G4-binding ligands can employ diverse mechanisms to interact with a G4 DNA, due to the polymorphic structural property of this DNA structure. The affinity and selectivity of G4-binding ligands to G4s rather than to other DNA forms are mainly achieved by the π-π stacking and electrostatic interaction, which are also the two main criteria in G4-binding ligand design. The large area of G-tetrad represents one of the most unique features of G4s as compared to other DNA forms. Thus, in most cases, the G4-stabilising effects are rendered by π-π stacking between the G-tetrad and the ligand (usually flat polyaromatic heterocycles), and the electrostatic interaction between positively charged ligand and negatively charged DNA. Although the intercalation binding mode of certain G4-binding ligands (ligand binds to a G4 by intercalating between G-tetrads, while the stabilising monovalent cations in the core of G4 are disrupted) were found in some cases [214, 215], the end-stacking mode (ligand binds to a G4 by stacking on the external face of G4, and the cations in the core of G4 are not disrupted) is by far the major mechanism as proposed by most experimental and modelling studies. This is mainly because the energetic cost to disrupt the inner cations is too great to be compensated by G4-ligand π-π stacking. Another issue in G4-binding ligand design is its selectivity between different G4s. Although general guidelines are yet to be established, some examples from novel G4-binding ligands suggest that the polymorphic features of G4 DNA can also be used as

recognition and binding sites for selective G4-ligand interaction. As mentioned previously, a G4 DNA can adopt four distinct grooves with unique geometry and electrostatic potential according to the combinations of *syn* or *anti* deoxyguanosine conformations. Furthermore, it is common that the sequences and arrangements of G4 loops can vary among different G4s or even among G4s from the same forming motif. Thus, G4 loops, together with grooves, may provide specific recognition sites in designing G4-binding ligands that target individual G4 selectively. More recently, a novel strategy via targeting the central cation channel to induce and stabilise the folding of G4 structures by mimicking the hydrogen bonding and cation-dipole interactions of central monovalent cations has been proposed by Rodriguez *et al.* [216]. However, due to the structural polymorphism, dynamic features of G4s, and correspondingly, insufficient confirmed structural data, the design of G4-binding ligands with high selectivity by employing those unique structural properties of G4 (*i.e.* loops, grooves, cation channel) is still mainly a trial-and-error process. So far, most G4-binding ligands can be grouped into several categories based on the catonic features and binding mechanisms, which are: (1) *in situ* protonated G4-binding ligands, (2) *N*-methylated aromatic G4-binding ligands, (3) metallo-organic G4-binding ligands, (4) neutral G4-binding ligands, and (5) peptide-nucleic acids (PNA) as G4-binding ligands. Individual examples will be given in the following section.

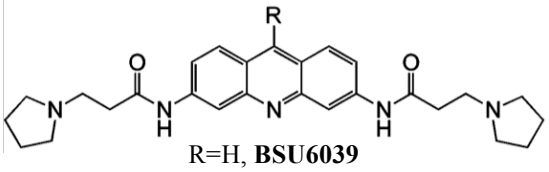
1.5.1 *In situ* protonated G4-binding ligands

Dated from the late 1990s, this group of G4-binding ligands is based on the strategy that a flat aromatic plane is involved in π-π stacking with G-tetrads, and the protonated sidearms retains water solubility. The compound 2,6-diamidoanthraquinone (Figure 1.7), designed together in Neidle's and Hurley's groups, is the first non-nucleoside, small molecule inhibitor that exhibits human telomerase inhibition effects by interacting with human telomeric G4 ($^{tel}EC_{50}$ = 23 μM, determined by a PCR-based amplification of the telomerase primer extension) [217]. Subsequently, a group of 1,4- and 2,6-disubstituted amidoanthracene-9, 10-dione derivatives and 3,6-disubstituted acridine chromophore derivatives have been synthesised as potential telomerase inhibitors (minimum $^{tel}EC_{50}$: 1.3 μM, determined by modified telomeric repeat amplification assays, TRAP) [218-220]. Although the

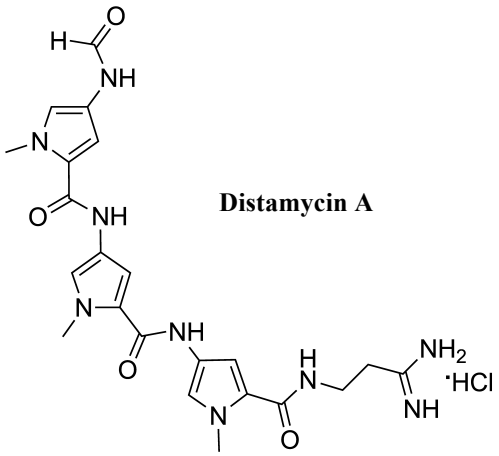
selectivity between G4 and duplex restricted further biological application of these ligands, the original strategy has been proved to be promising and various modifications have been followed subsequently. The compound 3,6-bis-[3-pyrrolidino-propionamide] acridine (BSU6039, $^{tel}EC_{50}$ = 5.2 μM [220], Figure 1.7) generated in previous studies was of particular interest once the crystal structure of this ligand with a bimolecular G4 was resolved [214]. The structure showed that the acridine moiety of the ligand formed a π-π stacking with the end G-tetrad, and specific hydrogen bonds with thymines in one loop [214]. More interestingly, both protonated sidearms of the ligand were shown to position close to the grooves of the G4, and this electrostatic interaction via hydrogen bonds provides further stabilising effects [214]. In accordance with the structural data, a series of 3,6,9-trisubstituted acridine inhibitors derived from BSU6039 have shown dramatically improved performance due to an extra protonated sidearm interacting with an extra groove [221-223]. Among these three-armed acridine derivatives, BRACO-19 (Figure 1.7) has attracted much attention, as it dramatically stabilises the humane telomeric G4 (ΔT_m ~ 20ºC, FRET), binds selectively to G4 DNA rather than duplex DNA (~ 30 times more specific), and significantly inhibits telomerase activity ($^{tel}EC_{50}$ ~ 100 nM, TRAP) [221, 222, 224]. Various *in vivo* investigations have also demonstrated the potential of BRACO-19 as an inhibitor of cancer cell proliferation [223, 225, 226]. Its main biopharmaceutical limitation is the poor cell permeability [227]. In an effort to overcome this, the aniline group at 9-amino substituent in BRACO-19 was replaced by a difluorobenzylamine group, known as compound AS1410 subsequently (Figure 1.7). This modification leads to improved lipophilicity and more favourable ΔT_m (~ 28ºC, FRET) and $^{tel}EC_{50}$ (30 nM, TRAP) as compared to the parent ligand BRACO-19 [228]. More recently, a group of tetrapeptide conjugates derived from BSU6039 and BRACO-19 (3,6-bis peptide acridines) have been reported to bind to G4s more tightly and selectively as compared to duplex DNA, and even among different G4s [229, 230].



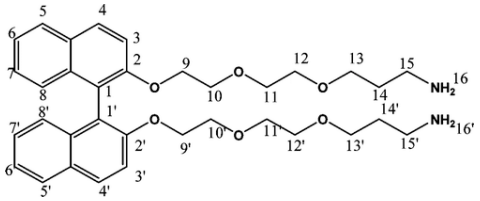
2,6-diamidoanthraquinone



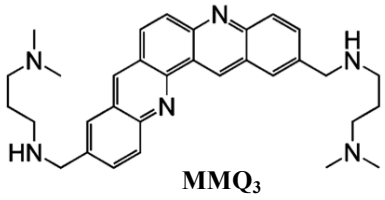
R=H, **BSU6039**
R=NH-*p*-C_6H_4-$N(CH_3)_2$, **BRACO-19**
R=NH-C_2H_2-$C_6H_3F_2$, **AS1410**



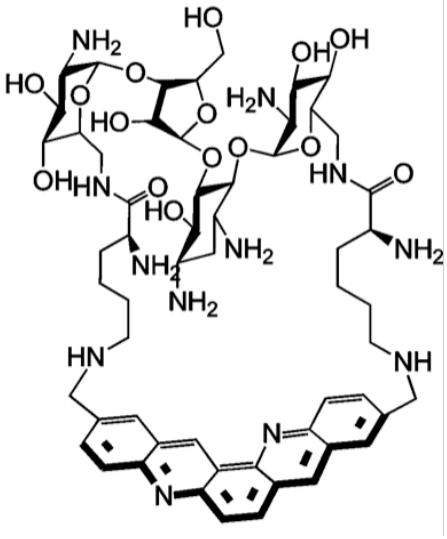
Distamycin A



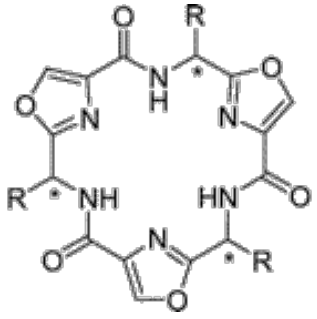
NPA (binaphthyl derivative)



MMQ_3



NCQ



oxazole-based peptide macrocycles

Figure 1.7: Cationic G4-binding ligands with *in situ* protonation.
(Structures of ligands were taken from references [217, 228, 231-235])

Besides their end-stacking mode, G4-binding ligands can also interact with G4s via grooves. The polypyrrole ligand distamycin A (Figure 1.7) has been identified as a DNA binder targeting the minor groove of dA•dT base pairs in duplex DNA [236]. ^{1}H

NMR spectroscopy and isothermal titration calorimetry (ICT) studies have revealed that the distamycin A is able to bind to two opposite grooves of parallel G4 $[d(TGGGGT)]_4$ in a dimeric anti-parallel fashion with a stoichiometry ration of 4:1 [231, 232]. Inhibited telomerase activity and cell proliferation, and induced apoptosis were found in various cancerous cell lines when exposed to a distamycin derivative MEN 10716 [237]. Another dicationic derivative of distamycin A was reported as a groove binder with a ligand-G4 ration of 2:1 with improved affinity ($K_b \sim 1.9 \times 10^6$ M^{-1}) and selectivity between G4 and duplex DNA [238]. More recently, a scissors-shaped binaphthyl derivative (NPA, Figure 1.7) has been found to be a potential telomerase inhibitor by placing its ethoxy chains along with the G4 grooves, although its selectivity and stabilising effects need to be further optimised (ΔT_m < 10 ºC, FRET) [233]. These groove binders suggest a new strategy in designing of G4-binding ligands. Another trend now emerging in G4-binding ligand design is to target multiple structural characteristics (i.e. large G-tetrad, loops, grooves) of G4s to increase the ligand selectivity. For example, the neomycin capped quinacridine series (NCQ, Figure 1.7) binds to the G-tetrad with its flexible pentacyclic quinacridine moiety, and interact with loops with its neomycin motif [239]. Although the stabilising and telomerase inhibition effects of NCQ (ΔT_m = 14ºC, FRET; $^{tel}EC_{50}$ ~ 200 nM, TRAP) are not as good as another pentacyclic quinacridine derivative (MMQ3, Figure 1.7, ΔT_m ~ 20ºC, FRET; $^{tel}EC_{50}$ ~ 28 nM, TRAP [240]), NCQ exhibits a much better selectivity for the loop-containing G4 [239]. This strategy was further followed and has led to the neomycin-perylene conjugate linking two neomycin units to a perylene diimide unit. This compound stacks at the end of human telomeric G4 and interacts with thymine bases in the loops by its perylene moiety, and two neomycin groups insert into the grooves of the G4, although further modification is needed to increase its stabilising ability (ΔT_m = 6 ºC, 1 μM, UV [241]). The oxazole-based peptide macrocycles (Figure 1.7), designed in Balasubramanian's group, interact with G4s in a similar mechanism. The trioxazole macrocyles with three groove-interacting protonated methylamine arms ($-(CH_2)_4NH_2$) was found to preferentially bind to and stabilise the KIT G4s as compared to the human telomeric G4 ($\Delta T_{m\ kit}$ = 10.8ºC, $\Delta T_{m\ tel}$ = 5.2ºC, 1 μM, FRET) [234, 242]. Several G4-binding ligands derived from macrocyclic oligoamides scaffold have been proved to be highly selective to human telomeric G4s, especially the planar trimeric macrocyclic

oligoamide of 8-amino-2-quinoline carboxylic acid, which stabilises human telomeric G4 more strongly than telomestatin [242, 243].

In addition, another G4 structural element, the central cation channel, has been proved to be a potential target in ligand design. Rodriguez *et al.* have designed an anthracene-derived ligand, which induces folding of parallel human telomeric G4 in the absence of cations by hydrogen bonding and cation-dipole interactions by mimicking the central K^+ ions [216].

As seem in above examples, this group of ligands are normally composed of a planar heteroaromatic moiety and a range of different types of substituent sidearms, such as amines [128, 244, 245], and peptides [228, 229]. Although still at very early stage, these compounds proved the strategy to design selective ligands for a particular G4 by targeting multiple structural elements when structural data are available (i.e., loop composition and arrangement, groove dimensions).

1.5.2 N-methylated aromatic G4-binding ligands

This group of ligands are normally quaternized on nitrogen atoms in the aromatic ring, which can increase their water solubility without addition of sidearms, and enhance the π-π stacking due to reduced electron repulsion. The classical example is the tetracatonic porphyrin TMPyP4 and its derivatives. TMPyP4 (Figure 1.8) can interact with G4s through diverse modes. For example, it can stack on the top of MYC G4 [246], intercalate between G-tetrads in inter- and intramolecular G4s [215, 247], or even stack onto the TTA loop of human telomeric G4 through hydrogen bonds without contact with G-tetrad [248]. The diverse binding modes of TMPyP4 may explain its non-selective binding to almost all G4 structures. Due to this property, TMPyP4 has been widely used in biophysical and biological studies about G4s. Several G4-binding ligands are derived from the tetracatonic macrocycles, such as TMPyPz and Se2SAP. Due to more extended porphyrazine moiety, which may completely overlap the G-tetrad, TMPyPz (Figure 1.8) exhibits much improved affinity (K_d = 0.2 ± 0.02 μM, UV-Vis) and selectivity (> 30-fold G4 over duplex DNA, SPR) to human telomeric G4 as compared to TMPyP4 (K_d = 14 μM, UV-Vis [215]) [249]. Se2SAP (Figure 1.8) is a core-modified expanded porphyrin analogue

that binds to G4s more strongly and more selectively (~ 50-fold G4 over duplex DNA, SPR) as compared to TMPyP4 [250]. Furthermore, Se2SAP can convert parallel MYC G4 and anti-parallel human telomeric G4 into a hybrid of parallel/anti-parallel G4 upon binding [250, 251], which indicated its potential to discriminate different G4s based on their different structures. Besides porphyrin derivatives, several quaternized small molecules have also been found to be promising G-binding ligands, such as RHPS4. Reported in 2000 by Stevens' group, RHPS4 (Figure 1.8) was selected from a series of tetra- and pentacyclic acridine molecules, which exhibited inhibitory effects to various cancerous cell lines *in vitro* [252]. Subsequently, RHPS4 was determined to be a potent telomerase inhibitor ($^{tel}EC_{50}$ ~ 300 nM, TRAP) [253]. NMR study showed that this ligand can stack on the external G-tetrads of parallel human telomeric G4 $d(TTAGGGT)_4$ [254]. A variety of *in vitro* studies determined its favourable pharmaceutical properties (i.e. fast cellular uptake, condensation in nucleus), which further confirmed it as a potent telomerase inhibitor [255, 256].

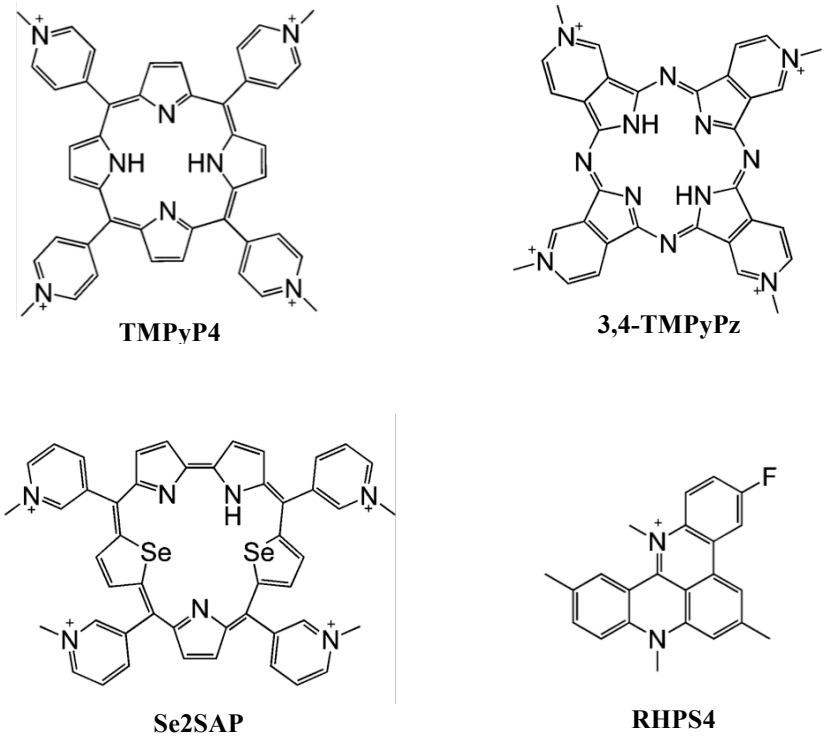
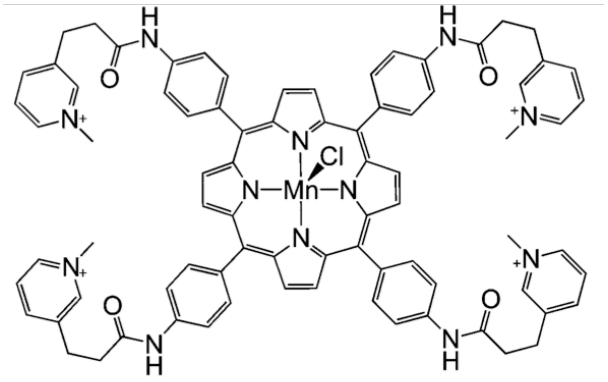


Figure 1.8: Cationic G4-binding ligands with N-methylation.
(Structures of ligands were taken from references [235])

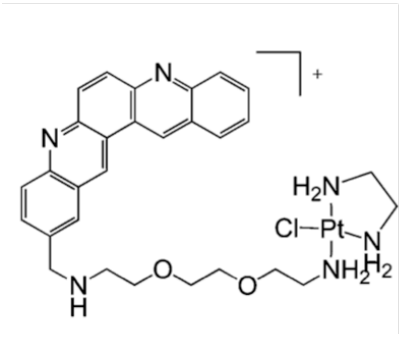
1.5.3 Metallo-organic G4-binding ligands

An alternative approach to develop efficient G4-binding ligands is to insert metallo-organic complexes either in the centre of π-π stacking planar cores, or as sidearms. In the first case, the π-π stacking between the organic planar core and G-tetrad could position the central metal atom of the ligand over the cation channel of the G4, which optimises stacking. On the other hand, the positive charge of the metal atoms in the sidearms of these ligands can form strong interactions with nearby G4 loops or grooves, which further enhance their performance as G4 stabilisers. The first group of metallo-organic G4-binding ligands were derived from the TMPyP4, in which a metal atom, such as Cu [257, 258], Mn, or Ni [259, 260], was inserted in the central cavity of this tetracatonic porphyrin. Although these ligands only showed modest telomerase inhibition ability ($^{tel}EC_{50} \geq 5$ μM, TRAP [259]), subsequent modification dramatically improved their performance. For example, by attaching four sidearms to the aromatic core of Mn(III)-phorphyrin (Figure 1.9), the new ligands can bind to human telomeric G4 strongly and selectively ($^{tel}EC_{50}$ = 580 nM, TRAP; 10^4-fold binding to G4 over duplex DNA), which is believed to be caused by π-π stacking between the organic planar core and the external G-tetrad, electrostatic interactions between the central Mn(III) ion and the carbonyl oxygens of guanine bases, and interaction between flexible cationic sidearms and G4 grooves and loops [261].

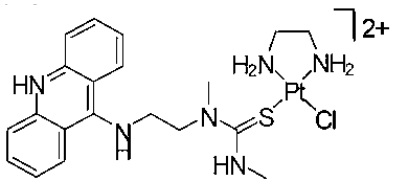
Since most G4-binding ligands described previously stabilise the human telomeric G4 via weak, reversible non-covalent interactions, another advantage of these metallo-organic ligands, especially platinum-complexed ligands, is to form covalent linkage with telomeric G4 and thus irreversibly stabilise it. By attaching a metal complex moiety to the central aromatic core of metallo-organic ligand, a novel platinum-quinacridine hybrid Pt-MPQ (Figure 1.9), comprising of a platinum moiety and a pentacyclic quinacridine G4-binding ligand MPQ, has been reported to dramatically stabilise the anti-parallel human telomeric G4, even under denaturing conditions. This interaction is proposed to be driven by stacking of MPQ and platinum moieties on opposite external G-treads, which preferentially platinate the G2/G22 residues [262]. Another cytotoxic platinum-acridine conjugate, PT-ACRAMTU (Figure 1.9), can bind to human telomeric G4 by its acridine moiety via π-π stacking, and platinate the N7 of adenine in the TTA loop by its platinum moiety [263], and thus cause irreversible trapping of the human telomeric G4.



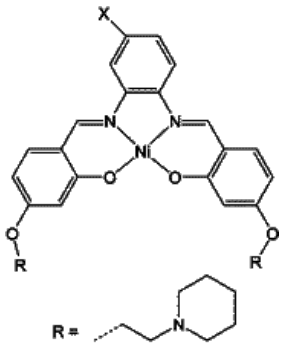
Mn(III)-porphyrin



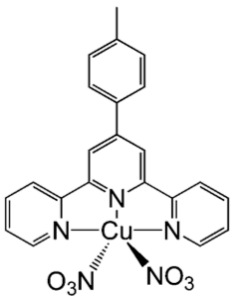
Pt-MPQ



PT-ACRAMTU



alkylamine-substituted
nickel(II)-salphen complexes
X = H/F



Cu-ttpy

Figure 1.9: Metallo-organic G4-binding ligands.
(Structures of ligands were taken from references [261, 263-265])

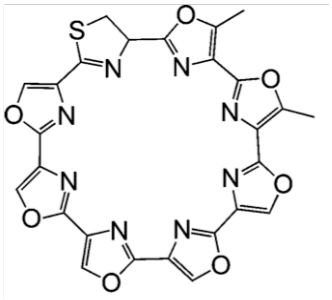
A group of simple synthetic ligands containing metal ions in the centre of planar core have been reported recently, and soon attracted much attention due to their easy synthesis and more importantly, promising performance. For example, two alkylamine-substituted nickel(II)-salphen complexes (Figure 1.9) has been reported to dramatically stabilise human telomeric G4 in solution ($\Delta T_{m\ tel}$ ~ 33ºC, 1 μM, FRET; $^{tel}EC_{50}$ ~ 120 nM, TRAP) with high selectivity (50-fold binding to G4 over duplex DNA) [265]. More recently, a family of terpyridine metallo-organic complexes containing Cu, Pt, Zn, and Ru have been designed and investigated, among which the

Cu-ttpy complex (Figure 1.9) exhibited best features. The Cu-ttpy, the product of a remarkably simple one-step synthesis process, has the strongest affinity and selectivity for G4s ($\Delta T_{m\ tel}$ = 16ºC, 1 μM, FRET; 22-fold binding to G4 over duplex DNA, FID) [264]. X-ray and modelling studies shows that the Cu^{2+} ion is important in offering the good performance of Cu-ttpy by perfectly stacking on the external G-tetrad, placing Cu^{2+} metal over the cation channel of G4, and impeding interaction with duplex DNA by a pyramid shape of $Cu(NO_3)_2$ [264]. Although its performance is not as good as Cu-ttpy, the Pt-ttpy has attracted further investigation due to its selective covalent linking properties [266]. Based on the Cu- and Pt-ttpy complexes, Vilar's group has designed dimetallic complexes, in which the second metal is coordinated to a dipicolyl amine linked to terpyridine by a three-atom spacer [267]. These dimetallic complexes show significantly improved selectivity as compared to the monometallic complexes reported previously (over 100-fold G4/duplex DNA for dicopper complex). The authors proposed that as compared to the monometallic complexes, the dimetallic substituent can further interact with DNA's backbone by electrostatic/metal-phosphate interactions [267].

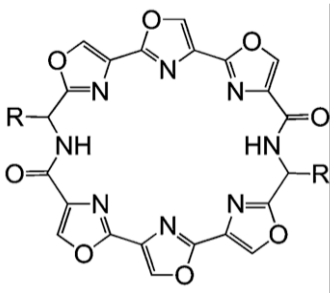
1.5.4 Neutral macrocyclic G4-binding ligands

The most important G4-binding ligand in this category is the telomestatin (Figure 1.10), a polyheteroaromatic 24-membered ring molecule. Since its isolation from *Streptomyces anulatus* 3533-SV4 in 2001 by Shin-ya *et al.* [268], it has been studied extensively mainly due to its exceptional properties, such as great stabilisation effect ($\Delta T_{m\ tel}$ = 24ºC), high selectivity to human telomeric G4 (> 70-fold, [269]), and strong telomerase inhibition ($^{tel}EC_{50}$ ~ 5 nM, [268]). This high efficiency of telomestatin is believed to be related to its perfect overlap with G-tetrad when forming π-π stacking, and the absence of significant groove interaction could further enhance its selectivity [270]. Due to these in solution observations, telomestatin was subject to various *in vitro* studies. For example, short-term and long-term treatment of telomestatin in different cancerous cell lines can induce telomere shortening and a growth arrest [271], inhibit the telomere maintenance, and cause apoptosis [272]. Following studies showed that telomestatin exerts its anti-cancer ability not only by inhibiting telomere elongation, but also through the uncapping of proteins, such as TRF2 [273], POT1

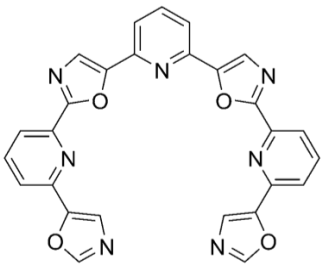
[101, 274], in the shelterin complex. One main drawback of this ligand is the difficulties in its synthesis, which requires an arduous multistep synthetic pathway [275], although some optimised methods have been reported [276, 277]. Meanwhile, the impressive anti-cancer activity of telomestatin has prompted efforts to develop oxazole-containing macrocyclic analogues (Figure 1.10), such as HXDV [278, 279], TOxaPy [280], and so on.



Telomestatin



R = $CH(CH_3)_2$: **HXDV**



TOxaPy

Figure 1.10: Neutral macrocyclic G4-binding ligands.
(Structures of ligands were taken from references [263, 280])

1.5.5 Novel methods to synthesise G4-binding ligands

Although most G4-binding ligands have been obtained by following traditional approaches of drug discovery, i.e. structure activity relationship (SAR), due to the insufficient structural data and their extreme dynamic properties of G4s, more efficient drug discovery rationale are needed. This can be potentially addressed by novel, high throughput approaches such as click chemistry, and dynamic combinatorial chemistry. Click chemistry (CC), first introduced by Sharpless *et al.* in 2001, describes a chemistry approach that generates large oligomers by joining small,

simple units together through carbon-heteroatom links [281, 282]. By following this powerful strategy, diverse drug-like compounds have been synthesised (reviewed in [281, 283, 284]), including G4-binding compounds with anti-cancer potential. For example, by using click chemistry employing the Cu(I)-catalyzed Huisgen reaction, a new class of ligands with non-polycyclic pharmacophore containing 1,4-triazole units have been prepared, which show promising G4-stabilising effects, affinity and selectivity [285, 286]. Compared to the CC approaches, which generate different G-tetrad-interacting core structures, the dynamic combinatorial chemistry (DCC) approaches are mainly focused on introducing different sidearms to a core structure in order to screen and optimise the binding properties of ligands to G4s. For example, by using partial combinatorial selection of tetrapeptides conjugated to a hemicyanine core (HC), Schouten *et al.* successfully synthesised a new class of G4-binding ligands with high affinity and selectivity [287]. Subsequently, they synthesised a group of ligands based on an oxazole-based peptide macrocycle core platform, together with libraries of side chains, which further proved DCC as a powerful approach in G4-binding ligands design in the absence of detailed G4 structural data [245].

In summary, discriminating and targeting specific G4s via small synthetic G4-binding ligands is still very challenging at present, which is due to: (1) structural polymorphism and dynamic properties of G4s, (2) similar major recognition sites (G-tetrad) in G4s, (3) limited precise structural data of G4s available, (4) weak abundance and instability of G4s in the human genome as compared to duplex DNA, (5) inaccessibility (unable to form) when bound up in chromatin, (6) obstacles in synthesis of natural G4-binding ligands, and (7) poor pharmaceutical properties of some synthetic compounds. Furthermore, as has been illustrated recently, anti-cancer mechanisms of G4-binding ligands targeting human telomeric G4s are found to be much more complicated than simply binding to telomeres and inhibiting telomerase activity (reviewed in [71, 288], and critically discussed in [103]). However, as more and more design rationales and synthesis approaches are being explored, combinations of recognition sites of G4s (G-tetrads, grooves, loops, cation channels, etc.) are being tested and investigated, and diverse *in vitro* and *in vivo* experimental models are being established, cellular and molecular principles that may guide synthesis of G4-binding ligands with high affinity, selectivity (between G4 and

duplex DNA, and even among different G4s), and good biological behaviours *in vivo* are possible in the close collaboration between chemists and biologists.

1.6 Off-target effects vs. multi-targeting effects

The correlations between G4s and gene functions, the roles of promoter G4s in regulating expression of oncoproteins, and the importance of telomeric G4 in maintaining telomere stability have been thoroughly investigated in recent decades, which leads to the application of several G4-binding ligands as potential anti-cancer drugs (as discussed in Section 1.5, and reviewed in [71, 235, 289]). As mentioned earlier, there are indeed some encouraging successes in designing specific G4-binding ligands, such as Mn(III) porphyrin (10^4-fold, G4 vs. duplex DNA) [261], and dimetallic terpyridine-based ligand (over 100-fold) [267]. Furthermore, the potential of these ligands is further strengthened by the accumulating evidence indicating the existence of G4 *in vivo*, such as G4-induced guanine-protection revealed by *in vitro* DMS footprinting [147, 148], and the telomeric G4 in the macronucleus of ciliated protozoa [93] (reviewed in [72]). However, although it becomes clear that G4 is functioning in living cells, discriminating and targeting specific G4s via small synthetic ligands (i.e. selectivity of G4-binding ligands) is still very challenging. At present, the selectivity (G4 vs. duplex) of most ligands is within the range between 10- and 100-fold (reviewed in [235]), while even fewer ligands have been reported to exert satisfactory selectivity between different G4s.

Thus, when taking into account the instability of G4s that occur transiently in the human genome, their much lower abundance as compared to adjacent duplex DNA, and the structural similarity between different G4s, G4-binding ligands with ideal selectivity and affinity to a particular G4 structure would be extremely difficult to achieve, and off-target effects are inevitable. Off-target effects have limited the therapeutic potential of these ligands *in vivo*. On the other hand, off-target effects may lead to multi-targeting effects in some cases, which in turn become advantageous for G4-binding ligands, and may enhance their therapeutic potential in some cases. For instance, a small G4-binding ligand, the naphthalene diimide derivative 1, has been shown to inhibit GIST cell growth by targeting two G4s: (1) it stabilises the *KIT* promoter G4s and down-regulates KIT expression level; and (2) it stabilises telomeric

G4s, disrupts shelterin complex, and efficiently arrests tumor cell proliferation (GIST) [290]. In another example, a ligand from the 2,4,6-triamino-1,3,5-triazine derivatives series, ligand 12459, was found to stabilise telomeric G4 (inhibit telomerase activity in TRAP assays), and hence induce cell senescence [291, 292]. However, subsequent in cell studies revealed that this ligands can alter the hTERT splicing pattern, and induce over-expression of -β hTERT (inactive hTERT form) by targeting G4-forming sequences in the intron six of -β hTERT transcript, which results in inhibition of telomerase activity [293]. Furthermore, the G4-binding fluoroquinolone derivative Quarfloxin (CX-3543), which was optimised and selected to specifically bind to *MYC* promoter G4, can repress up to 85 % of *MYC* gene transcription in tumor tissues (reviewed [294]). Later, the anti-cancer activity of this ligand was found to involve inhibiting rRNA (ribosomal RNA) biogenesis by targeting G4s in the rDNA (ribosomal DNA), and selectively disrupting nucleolin/rDNA G4 complexes [295]. Besides these limited examples, several pieces of evidence also imply that targeting multiple similar G4s in a particular pathway are feasible. Firstly, evidence from bioinformatics studies suggests that G4-forming sequences correlate with gene functions. For example, G4 abundance is higher in proto-oncogenes than that in tumor suppressor genes [188]. Secondly, composition of G4-forming sequences also shows certain level of conservation in a gene function-related pattern. For example, in cancer-related genes, promoter G4s have similar composition ($G_3N_{1\text{-}3}G_3N_{2\text{-}9}G_3N_1G_3$) [296]. Furthermore, most of promoter G4s identified in cancer-related genes all act as repressors, and their stabilisation by synthetic ligands can inhibit downstream gene expression, such as those identified in *MYC* [28, 111], *KIT* [290], *KRAS* [132, 136], *PDGF-A* [140] and *VEGF* [146]. Last but not least, structural dynamics property of G4s can also be used in designing ligands that efficiently interact with multiple G4s. Most G4-forming sequences are able to adopt several different structures that are in equilibrium. Due to this dynamic property of G4s, it is very likely that different G4-forming sequences may adopt similar structures (i.e. parallel, anti-parallel, etc.). Thus, certain G4-binding ligands can recognize similar sites (G-tetrads, or long loops) of these G4s even though they are formed by different G4-forming sequences, and preferentially stabilise them. For example, Se2SAP can convert parallel MYC G4 (less biologically relevant [28]) and anti-parallel human telomeric G4 into a hybrid of parallel/anti-parallel G4 upon binding [250, 251]. This structurally selective binding property of G4-binding ligands is worth further investigation, since stabilisation of

both biologically-related and unrelated G4s can both impede downstream gene transcription in proto-oncogenes. Thus, it is feasible to group G4s in promoters of genes with similar functions, or more precisely, those involved in the same or related pathways, and characterise their structural similarities, which can be used to guide the design of multi-targeting G4-binding ligands.

1.7 G4 in the heart*

Early bioinformatics studies clearly indicated that G4s are widely distributed in the human genome [17, 179-182]. However, up to date, almost all G4 researches have been focussed on their potential anti-cancer activities, such as telomeric G4 and promoter G4s in proto-oncogenes. Moreover, extensive studies towards these cancer-related G4s also have contributed great knowledge in this field, which have established the fundamentals to investigate the role of G4s in other genes, such as those involved in regulating cardiac functions. Since possible biological functions of G4s have been identified in some heart-related genes, such as *PDGF-A*, *VEGF-A* and *HIF-1α*, G4 research also has relevance to cardiovascular science. There are several points of interest about G4s and cardiovascular gene regulation.

1.7.1 G4s in the pPu/pPy regions

The first point is the transcription regulation function of G4s identified in the pPu/pPy regions or GC-rich regions of several gene promoters, such as *PDGF-A*, *VEGF-A* and *HIF-1α*.

The PDGF family is a group of multifunctional proteins composed of four different isoforms, PDGF-A, -B, -C, and -D. PDGF family members are key players in the regulation of embryonic development, angiogenesis, and in promoting cell differentiation and proliferation, while excessive expression of PDGFs has been associated with diverse human disorders, including atherosclerosis, balloon injury-induced restenosis, pulmonary hypertension, organ fibrosis and tumorigenesis (reviewed in [297]). From a cardiovascular perspective, PDGF-A and its receptor PDGFRα (platelet-derived growth factor receptor α) are particularly important in

* Part of this Section was adopted from my previous publication [6].

hypertension, atrial fibrillation and chronic rejection of cardiac allografts. For example, in angiotensin II-treated rat models, increased mRNA and protein levels of PGDF-A correlated with blood pressure and arterial wall hypertrophy, suggesting the direct roles of elevated blood pressure and production of PDGF-A in progression of vascular hypertrophy [298, 299]. More recently, arterial smooth muscle cells were found to be responsive to pro-proliferative effects of PDGF-AA, a homodimer of PGDF-A [300]. Liao *et al.* found that administration of PDGF-AA can increase the susceptibility to atrial fibrosis and artrial fibrillation in normal mouse hearts, whereas neutralizing PDGFR-α by specific antibody could attenuate the susceptibility in a pressure-overloaded model of heart failure [301]. Furthermore, PDGF-A, together with PDGF-C and -D, are involved in mediating profibrotic and proarteriosclerotic effects in rejection of rat cardiac allografts [302].

The VEGF family is a group of pluripotent cytokines that includes VEGF-A, -B, -C, -D, -E, and placenta growth factor. These factors bind to tyrosine kinase receptors VEGFR-1, -2, and -3 [303, 304]. VEGF-A is conventionally known as a critical mediator of angiogenesis during development and in collateral vessel formation in the adult, while more recently, alternative-splicing isoforms of VEGF-A were found to be anti-angiogenic (reviewed in [305]). Expression of VEGF-A protein is correlated with neovascularization in animal models [306], and can be up-regulated by hypoxia [307]. A particular interest has been their potential manipulation in order to regulate the formation of *de novo* blood vessels in the infarcted myocardium and for the vascularization of stem cell or other tissue-engineered patches for clinical repair [308].

HIF-1 is a heterodimeric TF that can sense hypoxia stress in myocardium. HIF-1 is composed of two subunits encoded by separate genes: *HIF-1α*, and *HIF-1β*. HIF-1β is constitutively expressed in myocardium under normoxic and hypoxia conditions, while HIF-1α level is usually very low under normoxic conditions but is rapidly up-regulated in response to hypoxia [309]. Under normoxia conditions, HIF-1α protein is rapidly degraded by the p53-dependent ubiquitin-proteasomal pathway [310]. However, in the presence of hypoxic stress, HIF-1α protein is rapidly accumulated due to the arrest of proteasomal-mediated degradation. The elevated level of HIF-1α allows the formation of functional HIF-1 dimers [310]. The functional HIF-1 dimers bind to their gene targets and subsequently recruit multiple transcriptional co-

activators via the HIF-1α subunit. The role of HIF-1α in neovascularisation associated with tumor development and cardiac hypertrophy has been studied in detail. Neovascularisation is essential in the process of the tumor development to support excessive tissue growth. In tumor angiogenesis, the expression of p53 has been found to inversely correlate with levels of HIF-1α and VEGF isoforms, both of which are critical components in angiogenesis [310]. It is well known that myocardial hypertrophic growth is initially compensatory ("physiological hypertrophy") but can decline into a pathological hypertrophy rapidly, which leads to impaired cardiac functions. It has been proposed recently that the extent of coronary angiogenesis determines the process of the hypertrophic heart declining into a pathological phenotype and heart failure [311]. In the process of angiogenesis, HIF-1 has been shown to regulate diverse angiogenic factors such as VEGFs and their receptors (reviewed in [312]).

These studies indicate that the *PDGF*, *VEGF* and *HIF-1α* genes could be potential targets in regulating vascular and cardiac functions, while G4s identified in the pPu/pPy regions in the promoters of these genes could also be attractive.

As mentioned previously, G4-forming sequences that are able to adopt G4 structures have been identified in the GC-rich NHE region in the promoter of *PDGF-A* [139], pPu/pPy regions in promoters of *VEGF-A* [145, 147, 148] and *HIF-1α* [162], while all these regions are critical in regulating downstream gene expression. Furthermore, repressor-like functions of G4s have been proved in cell models by stabilising them via G4-binding ligands [140, 146]. Several studies have shown that the pPu/pPy region of *VEGF* can adopt compact G4 structures in supercoiled plasmids and chromatins as determined by nuclease hypersensitivity assays, in solution and *in vitro* DMS footprinting, and chromatin immunoprecipitation (ChIP) experiments [145, 147]. In particular, ChIP experiments revealed that both G4-containing pPu/pPy regions from the *VEGF* and *HIF1-α* promoters were able to interact with RNA polymerase II, the Sp1 TF and nucleolin [147]. Nucleolin, which can bind G4s with a nanomolar-range dissociation constant [313, 314], is an abundant nucleolar protein involved in chromatin remodelling [315], while chromatin remodeling is the first step in regulating cardiac development (reviewed in [316]). On the other hand, pPu/pPy tracts, which may potentially adopt G4 structures, tend to be over-represented in the proximal promoter regions of many mammalian genes that lack a TATA-box,

including *VEGF*, *HIF-1α*, *EGFR* (epidermal growth factor receptor), *TGF-β* (transforming growth factor beta), *MYC*, *KRAS*, and *PDGF-A* [317, 318]. Putting all the evidence together, it is possible that G4-mediated mechanisms are involved in regulating the expression of genes involved in cardiac function, and thus may provide specific targets in order to manipulate transcription to an activating or repressing scenario, as appropriate.

1.7.2 Regulatory roles of G4s at RNA level

Another point of interest is the regulatory roles of G4s at the RNA level. The searches for the mechanisms that are responsible for the physiological and pathological processes of heart have been exhaustive. Now, rather than governed by few key proteins, it is known that these mechanisms are much more complicated than previously expected, and that cardiac functions are regulated at several different levels, including at the mRNA level. VEGF-A has long been regarded as a family of pro-angiogenic and pro-permeability vasodilator peptides, while the discovery of anti-angiogenic splice isoforms of VEGF-A_{xxx}b family in recent years has lead to a total re-evaluation of VEGF-A biology. For example, VEGF-A_{165} was found to chronically increase vascular permeability *in vivo* [319], stimulate angiogenesis *in vivo* [320], and mediate pathological retinal neovascularization *in vivo* [321]. However, an isoform with an alternative splice-site selection in exon 8, the VEGF-A_{165}b variant, was found to inhibit these pro-angiogenic effects mediated by VEGF-A_{165} [322-324]. So far, 14 subtypes of VEGF-A with diverse biological functions derived from alternative splicing have been revealed (reviewed in [305, 325]), which reflects the complexity and importance of regulation at the mRNA level in this cardiac function-related gene. Besides alternative splicing, recent studies have also revealed a potentially important role for a family of tiny regulatory RNAs, known as microRNA (miR) in controlling various aspects of cardiac functions in physiological and pathological conditions. For example, Rooij and colleagues found that miR-23a, miR-23b, miR-24, miR-195, and miR-214 were over-expressed in the progression of cardiac hypertrophy, and *in vitro* studies showed that they are able to provoke cardiac hypertrophy in neonatal cardiomyocytes models [326]. In the same study, the authors also found that some other microRNAs, such as miR-150 and-181b, were down-regulated in the same

hypertrophic model, which can reduce cardiac myocyte size when transfected into cells [326]. Although their mechanisms are still unclear, there is emerging evidence showing important roles of these microRNAs in regulating expression of diverse genes that govern heart functions (reviewed in [327, 328]). All these results indicate the importance of regulation at the mRNA level in cardiac function-related genes.

As compared to DNA G4s, although considerably fewer studies have been reported towards RNA G4s, it does not mean that RNA G4s are irrelevant in biological processes. In fact, G4s are believed to be more preferred at RNA level for some reasons. Firstly, RNA G4 is generally more stable than DNA G4. For example, Sacca *et al.* found that the substitution of thymine with uracil results in a moderate or strong stabilisation of G4 structures, although the effects of substitution of 2'-H of the sugar moiety with a hydroxy group are complicated [329]. Another systematic study showed that RNA G4s are more thermodynamically stable than corresponding DNA G4s [330]. Secondly, as a single-stranded structure, formation of RNA G4 does not need to compete with adjacent duplex, or to overcome the chromatin unfolding. Together with the evidence showing that almost all single-stranded G4-forming sequences can readily adopt G4 structures under near-physiological conditions, it is reasonable to assume that much less stringent conditions are needed for the formation of RNA G4s. There is accumulating evidence supporting the notion of biological importance of RNA G4s.

As mentioned previously, bioinformatics studies have revealed the accumulation of G4-forming sequences in the 5'-UTR, 3'-UTR, and the first introns in pre-mRNAs [181, 182], which are important in regulating RNA transcription, processing, and protein translation. Furthermore, G4-forming sequences are also widely identified around RNA processing sites and in mature mRNA. Kostadinov *et al.* established a G-Rich Sequence Database (GRSDB) from 1,310 human and mouse genes, in which 14.3 % RNA G4-forming sequences were identified around alternative splicing sites [331]. While in a model plant species *Arabidopsis thaliana*, ~ 43,000 G_2 (G4 with two G-tetrads) RNA G4-forming sequences were identified, in which 80 % of them were located in genetic regions suggesting the prevalence of G4 formation in mature mRNA [332]. Besides bioinformatics studies, emerging *in vitro* observations also indicated that RNA G4s are involved in regulating gene expression. In 2007, Balasubramanian's group showed that a conserved G4 formed in the 5'-UTR of the

NRAS proto-oncogene mRNA can down-regulate translation with and without G4-binding ligand in a cell-free translation system [213, 333]. The authors proposed a mechanism in which the formation of G4 in the 5'-UTR of mRNA can block the progression of the ribosomal complex, and thus down-regulate translation. Indeed, subsequent studies showed that mRNA 5'-UTR G4s identified in a variety of genes all down-regulate translation, such as those identified in *BCL-2*, *ZIC-1* (Zic family member 1), and TRF-2, for example (reviewed in [334]). Besides the 5'-UTR, G4s are also found around alternative splicing sites, and are involved in processes of alternative splicing. For example, as mentioned previously, the stabilisation of RNA G4s by ligand in the intron 6 of *hTERT* lead to the production of inactive -β hTERT transcript via impeding the splicing machinery [293]. More recently, Marcel *et al.* showed that the destabilisation and stabilisation of an RNA G4 in the intron 3 of a pre-mRNA of tumor suppressor gene *TP53* can lead to distinct alternative splicing isoforms [335]. Interestingly, several RNA G4s located downstream to the *P53* cleavage site were found to be critical in maintaining *P53* pre-mRNA 3' processing under stress conditions [336]. Meanwhile, *VEGF-A* is extremely rich in G4-forming sequences in both introns and exons, and it would be intriguing to investigate if RNA G4 is also involved in the alternative splicing of this gene under different stresses that may induce cardiac hypertrophy. Although the detailed mechanism is yet to be fully investigated, all these studies indicate that RNA G4s are no less likely to occur, or no less important than DNA G4s.

1.7.3 Interaction of G4s with TFs

Another aspect of G4 functions is the interaction of G4s with TFs, especially tissue-specific ones. For some TFs that recognize GC-rich region, their binding sites might well be contained within the G4-forming sequences. For example, it has been observed widely that G4-forming sequences often overlap with binding sites of C_2H_2 zinc finger TFs such as Sp1. In 2008, two independent studies demonstrated a clear link between the Sp1 binding site (GGGCGG) and G4-forming motifs in regions proximal and upstream to the TSS [182, 337]. Subsequently, genome-wide analysis have revealed that in gene promoters in human, chimpanzee, mouse and rat, the co-localization of G4-forming motifs and zinc finger TFs binding sites are highly

conserved, which may indicate a co-regulatory role between G4s and these TFs [191]. More recently, Balasubramanian's group found that besides its canonical duplex binding site, Sp1 can bind to the anti-parallel two-tetrad KIT G4 in solution with a comparable affinity [338]. Although correlations between G4s and binding sites of zinc finger TFs other than Sp1 are weak at genome level [182], correlation in cardiac function-related genes might be different. A subset of closely-related zinc finger proteins, the KLFs (kruppel-like zinc finger factors), bind a consensus C-rich sequence of CACC or CACCC. KLF proteins are involved in normal and pathological cardiac processes, such as cardiovascular remodeling and hypertrophy, and many show particular tissue-specific expression [339-341]. Given that the overall amino-acid similarity within the zinc finger motifs that comprise the DNA-binding domain of Sp1 and other members of the Sp1-like/KLF family (a minimum of 66.7 %, reviewed in [342]), it is also worthwhile to investigate possible interactions between G4s and KLFs. On the other hand, the close juxtaposition of G4 with a TF binding site also raises the possibility that G4s act as micro-environments for the recruitment of tissue-restricted factors, such as the mechanism proposed in the MyoD example described in Section 1.2.2 [60-62]. Although in cardiac muscle, key TFs such as the myogenic bHLH factors typified by MyoD are not present, two unique and largely cardiac-restricted bHLH proteins, HAND1 and HAND2 (HAND, heart, autonomic nervous system, neural crest-derived protein), have been found to be important in the process of heart development. HAND1 and HAND2 are involved in left-right patterning of the early cardiac tube, with mutation experiments in chick or other organisms indicating that their correct expression is absolutely essential for normal looping of the cardiac tube and development of a four-chambered heart [343]. Also, HAND1 appears to maintain the balance between cardiac myocyte differentiation and proliferation as over-expression of the mouse *Hand1* gene is associated with abnormal elongation of the developing heart tube and exaggerated looping [344]. It is perplexing though that very few gene targets of HAND proteins have been identified, which may suggest that HAND proteins function more as co-factors in conjunction with other TFs. It would be interesting to investigate the affinity of HAND proteins for G4 DNA in myocardial genes. HAND1 homodimers have been identified through immunoprecipitation and mammalian two-hybrid studies, and their affinity for DNA is at least an order of magnitude lower than that of HAND1/E-factor heterodimers [345], which echoes the observations made for MyoD homodimers described earlier.

1.8 MEF2D and TnIc

1.8.1 Human MEF2D

In acute or chronic injury to the adult hearts, pathological responses such as cardiomyocytes hypertrophy, and ventricular wall thinning have been observed to associate with development of pathological hypertrophy, which may result in heart failure and death [346, 347]. These pathological responses are actually caused by the activation of complex signaling pathways that activate various TFs, including myocyte enhancer factor-2 (MEF2), nuclear factor of activated T cells (NFAT), and GATA-4 [348], leading to additional sarcomere assembly and expression of genes typical of the fetal cardiac gene programme. There are four members (MEF2A ~ D) of the MEF2 family of MADS (MCM1, agamous, deficiency, serum response factor) box transcription factor family, which function as stress-dependent regulators of gene expression involved in pathological cardiac remodeling and multiple aspects of striated muscle development and diseases. Recent studies about MEF2D in knock-out mice lacking the *Mef2d* gene found that the mice with a Mef2d loss-of-function mutation are viable, but display impaired response to stress signals that are normally associated with the development of cardiac hypertrophy, fibrosis, and fetal gene activation. However, the forced over-expression of Mef2d in mice can cause severe cardiomyopathy. These results suggest that Mef2d is possibly a key factor in regulating stress-dependent gene in adult heart, raising expectation that human MEF2D might provide new target in developing "transcriptional" therapies for enhancing heart functions [349]. In the TRR (transcription regulatory region) of human *MEF2D* gene, several G4-forming motifs exist within 500bp upstream to the transcription start site (TSS; +1). In particular, the motif starting at -169 contains four consecutive runs of CCCC and its complementary strand would possibly form a stable 4-tetrad G4. In light of the previous studies about G4-forming motifs in the upstream TRR of *MYC*, *KIT*, *HIF-1α*, *KRAS*, *BCL-2*, *PDGF-A* and *VEGF* [296], this 4-tetrad G4 might play certain roles in regulating the expression of MEF2D in respect to stress-dependent signaling, and therefore could be a target for the development of drugs to counter heart failure.

1.8.2 Human TnIc

Troponin is a complex of three regulatory proteins (troponin I/C/T) in the thin filament that regulates muscle contraction in response to changes of intracellular calcium concentration in muscle. Troponin I is the inhibitory (I) subunit and binds with actin in the myofilaments, preventing the actomyosin ATPase activity before the calcium level reaches a threshold [350]. In mammals, three troponin I isoforms are encoded by three different genes, the fast skeletal muscle isoform (TNNI2, or TnIf), the slow skeletal muscle isoform (TNNI1, or TnIs), and cardiac isoform (TNNI3, or TnIc). TnIs is mainly expressed in the fast twitch skeletal muscle fibers [351]. TnIs is expressed in both cardiac and slow twitch skeletal muscle fibers during development, but then restricted to slow twitch skeletal muscle fibers in adult [352]. During the embryonic and fetal life of human, TnIs is the main troponin I isoform expressed in cardiac myocytes; while after 9 months postnatal development, only TnIc can be identified in human cardiac myocytes. TnIc has long been an ideal biomarker for myocardial infarction (MI) [353-355], and in 2000, TnIc was established as the preferred biomarker in MI diagnosis [356, 357]. It also has been proved that early detection of abnormal changes of TnIc in patients can effectively reduce the risk of death from heart attacks [355, 358, 359]. Furthermore, due to the direct correlation between mutations in TnIc in some populations and onset of hypertrophic cardiomyopathy (HCM), *TnIc* has been proved as one of several genes in which mutations can give rise to HCM [360]. Although TnIc is known as the only sarcomeric protein found to be exclusively expressed in cardiac myocytes, the molecular mechanisms that control this cardiac-specific expression are still poorly understood. Evidence shows that the expression of TnIc protein may be regulated at the transcription level [361]. By analyzing the proximal 1.1 kb of 5'-flanking region of human *TnIc*, Bhavsar *et al.* identified a group of cardiac-related TF binding sites. These include an MCAT element at -1100/-1094, a potential MEF-3 binding sites at -988/-978, four putative E-boxes (CANNTG) at -822/-817 (hE4), -766/-761 (hE3), -310/-305 (hE2), and -108/-103 (hE1), a CACC-box at -94/-87 (CCCACCCC), two GATA binding sties in reverse orientation at -68/-63, and -7/-2 (AGATAA), a potential TATA-box at -33/-24 (GTATATTTAG), and several Sp1 binding sites (Figure 5.1) [362]. Subsequently, the region between -531/+67 was determined to be essential for optimal transcription activity of *TnIc* in neonatal rat cardiomyocytes, in

which GATA, MEF-2, Sp1 and CACC box-binding factors are involved [363]. By investigating this promoter region, two types of intriguing G4-forming motifs were identified. The first type is a G4-forming minisatellite (MNSG4s). 11 repeats of a chromosome 19-specific potential MNSG4 were found between -729/-319. These are relatively conserved between -528/-319, and some of them contain an Sp1 binding site. These conserved MNSG4s were found to be involved in up-regulating the expression level of TnIc in a previous promoter deletion analysis [363]. Interestingly, on both sides of these MNSG4s, there are two E-boxes (hE3 and hE2). In the fifth intron of human TnIc, more than 30 repeats of the same conserved MNSG4 exist on the template strand. By comparing the sequences of the *TnIc* gene among different primates, the same MNSG4s were identified at similar positions, i.e. within the promoter and the fifth intron. Moreover, about 2.6 kb downstream to the human *TnIc* gene, there is a gene encoding another troponin isoform, the slow skeletal muscle troponin T (TNNT1, or TnT), whose expression is highly restricted to the skeletal muscle [364]. In the first intron of *TnT*, more than 20 repeats of the same G4-forming minisatellite sequence were found on the coding strand. The second type of G4-forming motifs in the promoter of human *TnIc* is the one identified at the position around -80 to the TSS of human *TnIc*. Between -97 and -73, this potential G4-forming sequence centred at around -80 ("-80G4) is found on the template strand, which overlaps with a CACC-box and an Sp1 binding site. It is also adjacent to an E-box and a GATA binding site (Figure 5.1).

1.9 Hypotheses and aims

My general hypotheses are that promoter G4s are also important in regulating cardiac function-related genes in both physiological and pathological processes, and these G4s might be used as potential targets via the multi-targeting strategy in the treatment of certain cardiomyopathies. My aims are: (1) to evaluate, by bioinformatics approaches, the importance of G4s in TRRs of key cardiac genes as compared to other genes, especially those involved in cancer, (2) to choose two model genes containing multiple G4s in their promoters according to bioinformatics results, and to characterise their biophysical properties in solution, and (3) to select one of them as a model system to investigate the regulatory roles of promoter G4s in selected cardiac

genes, and investigate the potential for multi-targeting strategy at the single gene level.

1.10 References

1. Watson, J. D., and Crick, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature 171*, 737-738.
2. Wang, A. H. J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., Vanboom, J. H., Vandermarel, G., and Rich, A. (1979) Molecular-Structure of a Left-Handed Double Helical DNA Fragment at Atomic Resolution, *Nature 282*, 680-686.
3. Drew, H., Takano, T., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980) High-salt d(CpGpCpG), a left-handed Z' DNA double helix, *Nature 286*, 567-573.
4. Franklin, R. E., and Gosling, R. G. (1953) Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate, *Nature 172*, 156-157.
5. Wang, A. H., Fujii, S., van Boom, J. H., van der Marel, G. A., van Boeckel, S. A., and Rich, A. (1982) Molecular structure of r(GCG)d(TATACGC): a DNA--RNA hybrid helix joined to double helical DNA, *Nature 299*, 601-604.
6. Zhou, W., Brand, N. J., and Ying, L. (2011) G-quadruplexes-novel mediators of gene function, *J Cardiovasc Transl Res 4*, 256-270.
7. Riley, M., and Maling, B. (1966) Physical and chemical characterization of two- and three-stranded adenine-thymine and adenine-uracil homopolymer complexes, *J Mol Biol 20*, 359-389.
8. Morgan, A. R., and Wells, R. D. (1968) Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences, *J Mol Biol 37*, 63-80.
9. Panayotatos, N., and Wells, R. D. (1981) Cruciform structures in supercoiled DNA, *Nature 289*, 466-470.
10. Lilley, D. M. J. (1980) The Inverted Repeat as a Recognizable Structural Feature in Supercoiled DNA-Molecules, *P Natl Acad Sci-Biol 77*, 6468-6472.
11. Gao, X., Huang, X., Smith, G. K., and Zheng, M. (1998) *Genetic Instabilities and Hereditary Neurological Diseases*, Wells, R. D., and Warren, S. T. ed., Academic Press, San Diego, CA.
12. Chou, S. H., Chin, K. H., and Wang, A. H. (2003) Unusual DNA duplex and hairpin motifs, *Nucleic Acids Res 31*, 2461-2474.
13. Phan, A. T., and Mergny, J. L. (2002) Human telomeric DNA: G-quadruplex, i-motif and Watson-Crick double helix, *Nucleic Acids Res 30*, 4618-4625.
14. Bacolla, A., and Wells, R. D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease, *J Biol Chem 279*, 47411-47414.
15. Wells, R. D. (2007) Non-B DNA conformations, mutagenesis and disease, *Trends Biochem Sci 32*, 271-278.
16. Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E., and Wells, R. D. (2006) The involvement of non-B DNA structures in gross chromosomal rearrangements, *DNA Repair (Amst) 5*, 1161-1170.
17. Huppert, J. L., and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome, *Nucleic Acids Res 35*, 406-413.
18. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr., and Blackburn, E. H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs, *Cell 51*, 899-908.
19. Bang, I. (1910) Untersuchungen über die Guanylsäure, *Biochemische Zeitschrift 26*, 293-311.
20. Gellert, M., Lipsett, M. N., and Davies, D. R. (1962) Helix formation by guanylic acid, *Proc Natl Acad Sci U S A 48*, 2013-2018.
21. Chantot, J. F., and Guschlbauer, W. (1972) Mechanism of gel formation by guanine nucleosides, *Jerusalem Symposium on Quantum Chemistry and Biochemistry 4*, 205-214.

22. Ralph, R. K., Connors, W. J., and Khorana, H. G. (1962) Secondary structure and aggregation in deoxyguanosine oligonucleotides, *Journal of the American Chemical Society 84*, 2265-2266.
23. Arnott, S., Chandrasekaran, R., and Marttila, C. M. (1974) Structures for polyinosinic acid and polyguanylic acid, *Biochem J 141*, 537-543.
24. Zimmerman, S. B., Cohen, G. H., and Davies, D. R. (1975) X-ray fiber diffraction and model-building study of polyguanylic acid and polyinosinic acid, *J Mol Biol 92*, 181-192.
25. Dai, J., Dexheimer, T. S., Chen, D., Carver, M., Ambrus, A., Jones, R. A., and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution, *J Am Chem Soc 128*, 1096-1098.
26. Dai, J., Chen, D., Jones, R. A., Hurley, L. H., and Yang, D. (2006) NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region, *Nucleic Acids Res 34*, 5133-5144.
27. Guo, K., Gokhale, V., Hurley, L. H., and Sun, D. (2008) Intramolecularly folded G-quadruplex and i-motif structures in the proximal promoter of the vascular endothelial growth factor gene, *Nucleic Acids Res 36*, 4598-4608.
28. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, *Proc Natl Acad Sci U S A 99*, 11593-11598.
29. Ambrus, A., Chen, D., Dai, J., Jones, R. A., and Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization, *Biochemistry 44*, 2048-2058.
30. Phan, A. T., Kuryavyi, V., Burge, S., Neidle, S., and Patel, D. J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter, *J Am Chem Soc 129*, 4386-4392.
31. Hazel, P., Huppert, J., Balasubramanian, S., and Neidle, S. (2004) Loop-length-dependent folding of G-quadruplexes, *J Am Chem Soc 126*, 16405-16415.
32. Bugaut, A., and Balasubramanian, S. (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes, *Biochemistry 47*, 689-697.
33. Risitano, A., and Fox, K. R. (2004) Influence of loop size on the stability of intramolecular DNA quadruplexes, *Nucleic Acids Res 32*, 2598-2606.
34. Rachwal, P. A., Brown, T., and Fox, K. R. (2007) Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes, *Biochemistry 46*, 3036-3044.
35. Kumar, N., Sahoo, B., Varun, K. A., and Maiti, S. (2008) Effect of loop length variation on quadruplex-Watson Crick duplex competition, *Nucleic Acids Res 36*, 4433-4442.
36. Cogoi, S., Paramasivam, M., Spolaore, B., and Xodo, L. E. (2008) Structural polymorphism within a regulatory element of the human KRAS promoter: formation of G4-DNA recognized by nuclear proteins, *Nucleic Acids Research 36*, 3765-3780.
37. Arnott, S., Chandras.R, and Marttila, C. M. (1974) Structures for Polyinosinic Acid and Polyguanylic Acid, *Biochemical Journal 141*, 537-&.
38. Smargiasso, N., Rosu, F., Hsia, W., Colson, P., Baker, E. S., Bowers, M. T., De Pauw, E., and Gabelica, V. (2008) G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation, *J Am Chem Soc 130*, 10208-10216.
39. Parkinson, G. N., Lee, M. P., and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA, *Nature 417*, 876-880.
40. Dailey, M. M., Miller, M. C., Bates, P. J., Lane, A. N., and Trent, J. O. (2010) Resolution and characterization of the structural polymorphism of a single quadruplex-forming sequence, *Nucleic Acids Res 38*, 4877-4888.

41. Li, J., Correia, J. J., Wang, L., Trent, J. O., and Chaires, J. B. (2005) Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal, *Nucleic Acids Res 33*, 4649-4659.
42. Phan, A. T., Kuryavyi, V., Luu, K. N., and Patel, D. J. (2007) Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K+ solution, *Nucleic Acids Res 35*, 6517-6525.
43. Lim, K. W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D. J., Luu, K. N., and Phan, A. T. (2009) Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers, *J Am Chem Soc 131*, 4301-4309.
44. Bell, G. I., Karam, J. H., and Rutter, W. J. (1981) Polymorphic DNA region adjacent to the 5' end of the human insulin gene, *Proc Natl Acad Sci U S A 78*, 5759-5763.
45. Bell, G. I., Selby, M. J., and Rutter, W. J. (1982) The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences, *Nature 295*, 31-35.
46. Bell, G. I., Horita, S., and Karam, J. H. (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus, *Diabetes 33*, 176-183.
47. Lucassen, A. M., Julier, C., Beressi, J. P., Boitard, C., Froguel, P., Lathrop, M., and Bell, J. I. (1993) Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR, *Nat Genet 4*, 305-310.
48. Hammond-Kosack, M. C., and Docherty, K. (1992) A consensus repeat sequence from the human insulin gene linked polymorphic region adopts multiple quadriplex DNA structures in vitro, *FEBS Lett 301*, 79-82.
49. Hammond-Kosack, M. C., Dobrinski, B., Lurz, R., Docherty, K., and Kilpatrick, M. W. (1992) The human insulin gene linked polymorphic region exhibits an altered DNA structure, *Nucleic Acids Res 20*, 231-236.
50. Hammond-Kosack, M. C., Kilpatrick, M. W., and Docherty, K. (1993) The human insulin gene-linked polymorphic region adopts a G-quartet structure in chromatin assembled in vitro, *J Mol Endocrinol 10*, 121-126.
51. Kennedy, G. C., German, M. S., and Rutter, W. J. (1995) The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription, *Nat Genet 9*, 293-298.
52. Lew, A., Rutter, W. J., and Kennedy, G. C. (2000) Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ, *Proc Natl Acad Sci U S A 97*, 12508-12512.
53. Catasti, P., Chen, X., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1996) Structure-function correlations of the insulin-linked polymorphic region, *J Mol Biol 264*, 534-545.
54. Schonhoft, J. D., Das, A., Achamyeleh, F., Samdani, S., Sewell, A., Mao, H. B., and Basu, S. (2010) ILPR Repeats Adopt Diverse G-Quadruplex Conformations that Determine Insulin Binding, *Biopolymers 93*, 21-31.
55. Xiao, J., and McGown, L. B. (2009) Mass spectrometric determination of ILPR G-quadruplex binding sites in insulin and IGF-2, *J Am Soc Mass Spectrom 20*, 1974-1982.
56. Connor, A. C., Frederick, K. A., Morgan, E. J., and McGown, L. B. (2006) Insulin capture by an insulin-linked polymorphic region G-quadruplex DNA oligonucleotide, *J Am Chem Soc 128*, 4986-4991.
57. Yu, Z., Schonhoft, J. D., Dhakal, S., Bajracharya, R., Hegde, R., Basu, S., and Mao, H. (2009) ILPR G-quadruplexes formed in seconds demonstrate high mechanical stabilities, *J Am Chem Soc 131*, 1876-1882.
58. Schonhoft, J. D., Bajracharya, R., Dhakal, S., Yu, Z., Mao, H., and Basu, S. (2009) Direct experimental evidence for quadruplex-quadruplex interaction within the human ILPR, *Nucleic Acids Res 37*, 3310-3320.

59. Santoro, I. M., Yi, T. M., and Walsh, K. (1991) Identification of Single-Stranded-DNA-Binding Proteins That Interact with Muscle Gene Elements, *Mol Cell Biol 11*, 1944-1953.
60. Walsh, K., and Gualberto, A. (1992) MyoD binds to the guanine tetrad nucleic acid structure, *J Biol Chem 267*, 13714-13718.
61. Yafe, A., Etzioni, S., Weisman-Shomer, P., and Fry, M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes, *Nucleic Acids Res 33*, 2887-2900.
62. Shklover, J., Weisman-Shomer, P., Yafe, A., and Fry, M. (2010) Quadruplex structures of muscle gene promoter sequences enhance in vivo MyoD-dependent gene expression, *Nucleic Acids Res 38*, 2369-2377.
63. O'Sullivan, R. J., and Karlseder, J. (2010) Telomeres: protecting chromosomes against genome instability, *Nat Rev Mol Cell Biol 11*, 171-181.
64. Murti, K. G., and Prescott, D. M. (1999) Telomeres of polytene chromosomes in a ciliated protozoan terminate in duplex DNA loops, *Proc Natl Acad Sci U S A 96*, 14436-14439.
65. Verdun, R. E., and Karlseder, J. (2007) Replication and protection of telomeres, *Nature 447*, 924-931.
66. de Lange, T. (2004) T-loops and the origin of telomeres, *Nat Rev Mol Cell Biol 5*, 323-329.
67. Olovnikov, A. M. (1973) A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon, *J Theor Biol 41*, 181-190.
68. Harley, C. B., Futcher, A. B., and Greider, C. W. (1990) Telomeres shorten during ageing of human fibroblasts, *Nature 345*, 458-460.
69. Greider, C. W., and Blackburn, E. H. (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity, *Cell 51*, 887-898.
70. Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L., Coviello, G. M., Wright, W. E., Weinrich, S. L., and Shay, J. W. (1994) Specific association of human telomerase activity with immortal cells and cancer, *Science 266*, 2011-2015.
71. Neidle, S. (2010) Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer, *FEBS J 277*, 1118-1125.
72. Lipps, H. J., and Rhodes, D. (2009) G-quadruplex structures: in vivo evidence and function, *Trends Cell Biol 19*, 414-422.
73. Zhang, N., Phan, A. T., and Patel, D. J. (2005) (3 + 1) Assembly of three human telomeric repeats into an asymmetric dimeric G-quadruplex, *J Am Chem Soc 127*, 17277-17285.
74. Griffith, J. D., Comeau, L., Rosenfield, S., Stansel, R. M., Bianchi, A., Moss, H., and de Lange, T. (1999) Mammalian telomeres end in a large duplex loop, *Cell 97*, 503-514.
75. Xu, Y., Ishizuka, T., Kurabayashi, K., and Komiyama, M. (2009) Consecutive formation of G-quadruplexes in human telomeric-overhang DNA: a protective capping structure for telomere ends, *Angew Chem Int Ed Engl 48*, 7833-7836.
76. Han, H., and Hurley, L. H. (2000) G-quadruplex DNA: a potential target for anti-cancer drug design, *Trends Pharmacol Sci 21*, 136-142.
77. Arthanari, H., and Bolton, P. H. (2001) Functional and dysfunctional roles of quadruplex DNA in cells, *Chem Biol 8*, 221-230.
78. Price, C. (2006) *Ciliate Telomere, In Telomeres*, 2nd ed., Cold Spring Harbor Laboratory Press.
79. Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D., and Lipps, H. J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo, *Nat Struct Mol Biol 12*, 847-854.

80. Paeschke, K., Juranek, S., Simonsson, T., Hempel, A., Rhodes, D., and Lipps, H. J. (2008) Telomerase recruitment by the telomere end binding protein-beta facilitates G-quadruplex DNA unfolding in ciliates, *Nat Struct Mol Biol 15*, 598-604.
81. Sen, D., and Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis, *Nature 334*, 364-366.
82. Wang, Y., and Patel, D. J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex, *Structure 1*, 263-282.
83. He, Y., Neumann, R. D., and Panyutin, I. G. (2004) Intramolecular quadruplex conformation of human telomeric DNA assessed with 125I-radioprobing, *Nucleic Acids Res 32*, 5359-5367.
84. Phan, A. T., and Patel, D. J. (2003) Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: distinct topologies, thermodynamic properties, and folding/unfolding kinetics, *J Am Chem Soc 125*, 15021-15027.
85. Vorlickova, M., Chladkova, J., Kejnovska, I., Fialova, M., and Kypr, J. (2005) Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats, *Nucleic Acids Res 33*, 5851-5860.
86. Lee, J. Y., Okumus, B., Kim, D. S., and Ha, T. (2005) Extreme conformational diversity in human telomeric DNA, *Proc Natl Acad Sci U S A 102*, 18938-18943.
87. Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R. A., and Yang, D. (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution, *Nucleic Acids Res 34*, 2723-2735.
88. Dai, J., Carver, M., Punchihewa, C., Jones, R. A., and Yang, D. (2007) Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K+ solution: insights into structure polymorphism of the human telomeric sequence, *Nucleic Acids Res 35*, 4927-4940.
89. Xu, Y., Noguchi, Y., and Sugiyama, H. (2006) The new models of the human telomere d[AGGG(TTAGGG)3] in K+ solution, *Bioorg Med Chem 14*, 5584-5591.
90. Luu, K. N., Phan, A. T., Kuryavyi, V., Lacroix, L., and Patel, D. J. (2006) Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold, *J Am Chem Soc 128*, 9963-9970.
91. Zahler, A. M., Williamson, J. R., Cech, T. R., and Prescott, D. M. (1991) Inhibition of telomerase by G-quartet DNA structures, *Nature 350*, 718-720.
92. Prescott, D. M. (1994) The DNA of ciliated protozoa, *Microbiol Rev 58*, 233-267.
93. Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H. J., and Pluckthun, A. (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei, *Proc Natl Acad Sci U S A 98*, 8572-8577.
94. Lipps, H. J. (1980) In vitro aggregation of the gene-sized DNA molecules of the ciliate Stylonychia mytilus, *Proc Natl Acad Sci U S A 77*, 4104-4107.
95. Sundquist, W. I., and Klug, A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops, *Nature 342*, 825-829.
96. Lu, Q., Schierer, T., Kang, S. G., and Henderson, E. (1998) Purification, characterization and molecular cloning of TGP1, a novel G-DNA binding protein from Tetrahymena thermophila, *Nucleic Acids Res 26*, 1613-1620.
97. Lu, Q., and Henderson, E. (2000) Two Tetrahymena G-DNA-binding proteins, TGP1 and TGP3, share novel motifs and may play a role in micronuclear division, *Nucleic Acids Res 28*, 2993-3001.
98. Oganesian, L., Moon, I. K., Bryan, T. M., and Jarstfer, M. B. (2006) Extension of G-quadruplex DNA by ciliate telomerase, *EMBO J 25*, 1148-1159.
99. Salvati, E., Leonetti, C., Rizzo, A., Scarsella, M., Mottolese, M., Galati, R., Sperduti, I., Stevens, M. F., D'Incalci, M., Blasco, M., Chiorino, G., Bauwens, S., Horard, B., Gilson, E., Stoppacciaro, A., Zupi, G., and Biroccio, A. (2007) Telomere damage induced by the G-quadruplex ligand RHPS4 has an antitumor effect, *J Clin Invest 117*, 3236-3247.

100. Rodriguez, R., Muller, S., Yeoman, J. A., Trentesaux, C., Riou, J. F., and Balasubramanian, S. (2008) A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres, *J Am Chem Soc 130*, 15758-15759.
101. Gomez, D., Wenner, T., Brassart, B., Douarre, C., O'Donohue, M. F., El Khoury, V., Shin-Ya, K., Morjani, H., Trentesaux, C., and Riou, J. F. (2006) Telomestatin-induced telomere uncapping is modulated by POT1 through G-overhang extension in HT1080 human tumor cells, *J Biol Chem 281*, 38721-38729.
102. Zaug, A. J., Podell, E. R., and Cech, T. R. (2005) Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro, *Proc Natl Acad Sci U S A 102*, 10864-10869.
103. De Cian, A., Cristofari, G., Reichenbach, P., De Lemos, E., Monchaud, D., Teulade-Fichou, M. P., Shin-Ya, K., Lacroix, L., Lingner, J., and Mergny, J. L. (2007) Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action, *Proc Natl Acad Sci U S A 104*, 17347-17352.
104. Pelengaris, S., Rudolph, B., and Littlewood, T. (2000) Action of Myc in vivo - proliferation and apoptosis, *Curr Opin Genet Dev 10*, 100-105.
105. Vita, M., and Henriksson, M. (2006) The Myc oncoprotein as a therapeutic target for human cancer, *Semin Cancer Biol 16*, 318-330.
106. Berberich, S. J., and Postel, E. H. (1995) PuF/NM23-H2/NDPK-B transactivates a human c-myc promoter-CAT gene via a functional nuclease hypersensitive element, *Oncogene 10*, 2343-2347.
107. Simonsson, T., Pecinka, P., and Kubista, M. (1998) DNA tetraplex formation in the control region of c-myc, *Nucleic Acids Res 26*, 1167-1172.
108. Simonsson, T., and Sjoback, R. (1999) DNA tetraplex formation studied with fluorescence resonance energy transfer, *J Biol Chem 274*, 17379-17383.
109. Phan, A. T., Modi, Y. S., and Patel, D. J. (2004) Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter, *J Am Chem Soc 126*, 8710-8716.
110. Mathad, R. I., Hatzakis, E., Dai, J., and Yang, D. (2011) c-MYC promoter G-quadruplex formed at the 5'-end of NHE III1 element: insights into biological relevance and parallel-stranded G-quadruplex stability, *Nucleic Acids Res 39*, 9023-9033.
111. Grand, C. L., Han, H., Munoz, R. M., Weitman, S., Von Hoff, D. D., Hurley, L. H., and Bearss, D. J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo, *Mol Cancer Ther 1*, 565-573.
112. Sun, D., and Hurley, L. H. (2009) The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression, *J Med Chem 52*, 2863-2874.
113. Gonzalez, V., Guo, K., Hurley, L., and Sun, D. (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein, *J Biol Chem 284*, 23622-23635.
114. Borgognone, M., Armas, P., and Calcaterra, N. B. (2010) Cellular nucleic-acid-binding protein, a transcriptional enhancer of c-Myc, promotes the formation of parallel G-quadruplexes, *Biochem J 428*, 491-498.
115. Gonzalez, V., and Hurley, L. H. (2010) The c-MYC NHE III(1): function and regulation, *Annu Rev Pharmacol Toxicol 50*, 111-129.
116. Miettinen, M., and Lasota, J. (2005) KIT (CD117): a review on expression in normal and neoplastic tissues, and mutations and their clinicopathologic correlation, *Appl Immunohistochem Mol Morphol 13*, 205-220.
117. Edling, C. E., and Hallberg, B. (2007) c-Kit--a hematopoietic cell essential receptor tyrosine kinase, *Int J Biochem Cell Biol 39*, 1995-1998.
118. Siehl, J., and Thiel, E. (2007) C-kit, GIST, and imatinib, *Recent Results Cancer Res 176*, 145-151.

119. Agaimy, A., Wunsch, P. H., Hofstaedter, F., Blaszyk, H., Rummele, P., Gaumann, A., Dietmaier, W., and Hartmann, A. (2007) Minute gastric sclerosing stromal tumors (GIST tumorlets) are common in adults and frequently show c-KIT mutations, *Am J Surg Pathol 31*, 113-120.
120. Thakur, R. K., Kumar, P., Halder, K., Verma, A., Kar, A., Parent, J. L., Basundra, R., Kumar, A., and Chowdhury, S. (2009) Metastases suppressor NM23-H2 interaction with G-quadruplex DNA within c-MYC promoter nuclease hypersensitive element induces c-MYC expression, *Nucleic Acids Res 37*, 172-183.
121. Balasubramanian, S., Hurley, L. H., and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?, *Nat Rev Drug Discov 10*, 261-275.
122. Park, G. H., Plummer, H. K., 3rd, and Krystal, G. W. (1998) Selective Sp1 binding is critical for maximal activity of the human c-kit promoter, *Blood 92*, 4138-4149.
123. Rankin, S., Reszka, A. P., Huppert, J., Zloh, M., Parkinson, G. N., Todd, A. K., Ladame, S., Balasubramanian, S., and Neidle, S. (2005) Putative DNA quadruplex formation within the human c-kit oncogene, *J Am Chem Soc 127*, 10584-10589.
124. Todd, A. K., Haider, S. M., Parkinson, G. N., and Neidle, S. (2007) Sequence occurrence and structural uniqueness of a G-quadruplex in the human c-kit promoter, *Nucleic Acids Res 35*, 5799-5808.
125. Hsu, S. T., Varnai, P., Bugaut, A., Reszka, A. P., Neidle, S., and Balasubramanian, S. (2009) A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics, *J Am Chem Soc 131*, 13399-13409.
126. Fernando, H., Reszka, A. P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A. R., Neidle, S., and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene, *Biochemistry 45*, 7854-7860.
127. Kuryavyi, V., Phan, A. T., and Patel, D. J. (2010) Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter, *Nucleic Acids Res 38*, 6757-6773.
128. Bejugam, M., Sewitz, S., Shirude, P. S., Rodriguez, R., Shahid, R., and Balasubramanian, S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression, *J Am Chem Soc 129*, 12926-12927.
129. McLuckie, K. I., Waller, Z. A., Sanders, D. A., Alves, D., Rodriguez, R., Dash, J., McKenzie, G. J., Venkitaraman, A. R., and Balasubramanian, S. (2011) G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells, *J Am Chem Soc 133*, 2658-2663.
130. Shirude, P. S., Okumus, B., Ying, L., Ha, T., and Balasubramanian, S. (2007) Single-molecule conformational analysis of G-quadruplex formation in the promoter DNA duplex of the proto-oncogene c-kit, *J Am Chem Soc 129*, 7484-7485.
131. Kranenburg, O. (2005) The KRAS oncogene: past, present, and future, *Biochim Biophys Acta 1756*, 81-82.
132. Cogoi, S., and Xodo, L. E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, *Nucleic Acids Res 34*, 2536-2549.
133. Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H., and Xodo, L. E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription, *Nucleic Acids Res 37*, 2841-2853.
134. Xodo, L., Paramasivam, M., Membrino, A., and Cogoi, S. (2008) Protein hnRNPA1 binds to a critical G-rich element of KRAS and unwinds G-quadruplex structures: implications in transcription, *Nucleic Acids Symp Ser (Oxf)*, 159-160.
135. Soldatenkov, V. A., Vetcher, A. A., Duka, T., and Ladame, S. (2008) First evidence of a functional interaction between DNA quadruplexes and poly(ADP-ribose) polymerase-1, *ACS Chem Biol 3*, 214-219.

136. Cogoi, S., Paramasivam, M., Spolaore, B., and Xodo, L. E. (2008) Structural polymorphism within a regulatory element of the human KRAS promoter: formation of G4-DNA recognized by nuclear proteins, *Nucleic Acids Res 36*, 3765-3780.
137. Cogoi, S., Paramasivam, M., Filichev, V., Geci, I., Pedersen, E. B., and Xodo, L. E. (2009) Identification of a new G-quadruplex motif in the KRAS promoter and design of pyrene-modified G4-decoys with antiproliferative activity in pancreatic cancer cells, *J Med Chem 52*, 564-568.
138. Yu, J., Ustach, C., and Kim, H. R. (2003) Platelet-derived growth factor signaling and human cancer, *J Biochem Mol Biol 36*, 49-59.
139. Kaetzel, D. M., Jr., Maul, R. S., Liu, B., Bonthron, D., Fenstermaker, R. A., and Coyne, D. W. (1994) Platelet-derived growth factor A-chain gene transcription is mediated by positive and negative regulatory regions in the promoter, *Biochem J 301 (Pt 2)*, 321-327.
140. Qin, Y., Rezler, E. M., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4, *Nucleic Acids Res 35*, 7698-7713.
141. Arighi, E., Borrello, M. G., and Sariola, H. (2005) RET tyrosine kinase signaling in development and cancer, *Cytokine Growth Factor Rev 16*, 441-467.
142. Munnes, M., Patrone, G., Schmitz, B., Romeo, G., and Doerfler, W. (1998) A 5'-CG-3'-rich region in the promoter of the transcriptionally frequently silenced RET protooncogene lacks methylated cytidine residues, *Oncogene 17*, 2573-2583.
143. Guo, K., Pourpak, A., Beetz-Rogers, K., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene, *J Am Chem Soc 129*, 10220-10228.
144. Neufeld, G., Cohen, T., Gengrinovitch, S., and Poltorak, Z. (1999) Vascular endothelial growth factor (VEGF) and its receptors, *FASEB J 13*, 9-22.
145. Sun, D., Guo, K., Rusche, J. J., and Hurley, L. H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents, *Nucleic Acids Res 33*, 6070-6080.
146. Sun, D., Liu, W. J., Guo, K., Rusche, J. J., Ebbinghaus, S., Gokhale, V., and Hurley, L. H. (2008) The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex-interactive agents, *Mol Cancer Ther 7*, 880-889.
147. Sun, D., Guo, K., and Shin, Y. J. (2011) Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene, *Nucleic Acids Res 39*, 1256-1265.
148. Sun, D. (2010) In vitro footprinting of promoter regions within supercoiled plasmid DNA, *Methods Mol Biol 613*, 223-233.
149. Murphree, A. L., and Benedict, W. F. (1984) Retinoblastoma: clues to human oncogenesis, *Science 223*, 1028-1033.
150. Murchie, A. I., and Lilley, D. M. (1992) Retinoblastoma susceptibility genes contain 5' sequences with a high propensity to form guanine-tetrad structures, *Nucleic Acids Res 20*, 49-53.
151. Xu, Y., and Sugiyama, H. (2004) Highly efficient photochemical 2'-deoxyribonolactone formation at the diagonal loop of a 5-iodouracil-containing antiparallel G-quartet, *J Am Chem Soc 126*, 6274-6279.
152. Xu, Y., and Sugiyama, H. (2006) Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Res 34*, 949-954.
153. Xu, Y., and Sugiyama, H. (2005) Structural and functional characterizations of the G-quartet and i-motif elements in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Symp Ser (Oxf)*, 177-178.
154. Adams, J. M., and Cory, S. (1998) The Bcl-2 protein family: arbiters of cell survival, *Science 281*, 1322-1326.

155. Young, R. L., and Korsmeyer, S. J. (1993) A negative regulatory element in the bcl-2 5'-untranslated region inhibits expression from an upstream promoter, *Mol Cell Biol 13*, 3686-3697.
156. Dexheimer, T. S., Sun, D., and Hurley, L. H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter, *J Am Chem Soc 128*, 5404-5415.
157. Del Toro, M., Bucek, P., Avino, A., Jaumot, J., Gonzalez, C., Eritja, R., and Gargallo, R. (2009) Targeting the G-quadruplex-forming region near the P1 promoter in the human BCL-2 gene with the cationic porphyrin TMPyP4 and with the complementary C-rich strand, *Biochimie 91*, 894-902.
158. Douarre, C., Gomez, D., Morjani, H., Zahm, J. M., O'Donohue M, F., Eddabra, L., Mailliet, P., Riou, J. F., and Trentesaux, C. (2005) Overexpression of Bcl-2 is associated with apoptotic resistance to the G-quadruplex ligand 12459 but is not sufficient to confer resistance to long-term senescence, *Nucleic Acids Res 33*, 2192-2203.
159. Ratcliffe, P. J. (2002) From erythropoietin to oxygen: hypoxia-inducible factor hydroxylases and the hypoxia signal pathway, *Blood Purif 20*, 445-450.
160. Ferrara, N. (2004) Vascular endothelial growth factor: basic science and clinical progress, *Endocr Rev 25*, 581-611.
161. Ebbinghaus, S. W., and Gordon, M. S. (2004) Renal cell carcinoma: rationale and development of therapeutic inhibitors of angiogenesis, *Hematol Oncol Clin North Am 18*, 1143-1159, ix-x.
162. De Armond, R., Wood, S., Sun, D., Hurley, L. H., and Ebbinghaus, S. W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter, *Biochemistry 44*, 16341-16350.
163. Guan, Y., Reddy, K. R., Zhu, Q., Li, Y., Lee, K., Weerasinghe, P., Prchal, J., Semenza, G. L., and Jing, N. (2010) G-rich oligonucleotides inhibit HIF-1alpha and HIF-2alpha and block tumor growth, *Mol Ther 18*, 188-197.
164. Lyonnais, S., Gorelick, R. J., Mergny, J. L., Le Cam, E., and Mirambeau, G. (2003) G-quartets direct assembly of HIV-1 nucleocapsid protein along single-stranded DNA, *Nucleic Acids Res 31*, 5754-5763.
165. Erlitzki, R., and Fry, M. (1997) Sequence-specific binding protein of single-stranded and unimolecular quadruplex telomeric DNA from rat hepatocytes, *J Biol Chem 272*, 15881-15890.
166. Giraldo, R., and Rhodes, D. (1994) The yeast telomere-binding protein RAP1 binds to and promotes the formation of DNA quadruplexes in telomeric DNA, *EMBO J 13*, 2411-2420.
167. Muniyappa, K., Anuradha, S., and Byers, B. (2000) Yeast meiosis-specific protein Hop1 binds to G4 DNA and promotes its formation, *Mol Cell Biol 20*, 1361-1369.
168. Arimondo, P. B., Riou, J. F., Mergny, J. L., Tazi, J., Sun, J. S., Garestier, T., and Helene, C. (2000) Interaction of human DNA topoisomerase I with G-quartet structures, *Nucleic Acids Res 28*, 4832-4838.
169. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H., and Toole, J. J. (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin, *Nature 355*, 564-566.
170. Fry, M., and Loeb, L. A. (1999) Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n, *J Biol Chem 274*, 12797-12802.
171. Mohaghegh, P., Karow, J. K., Brosh, R. M., Jr., Bohr, V. A., and Hickson, I. D. (2001) The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases, *Nucleic Acids Res 29*, 2843-2849.
172. Ribeyre, C., Lopes, J., Boule, J. B., Piazza, A., Guedin, A., Zakian, V. A., Mergny, J. L., and Nicolas, A. (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo, *PLoS Genet 5*, e1000475.

173. Sanders, C. M. (2010) Human Pif1 helicase is a G-quadruplex DNA-binding protein with G-quadruplex DNA-unwinding activity, *Biochem J 430*, 119-128.
174. Sarig, G., Weisman-Shomer, P., Erlitzki, R., and Fry, M. (1997) Purification and characterization of qTBP42, a new single-stranded and quadruplex telomeric DNA-binding protein from rat hepatocytes, *J Biol Chem 272*, 4474-4482.
175. Riva, S., Morandi, C., Tsoulfas, P., Pandolfo, M., Biamonti, G., Merrill, B., Williams, K. R., Multhaup, G., Beyreuther, K., Werr, H., and et al. (1986) Mammalian single-stranded DNA binding protein UP I is derived from the hnRNP core protein A1, *EMBO J 5*, 2267-2273.
176. Merrill, B. M., LoPresti, M. B., Stone, K. L., and Williams, K. R. (1986) High pressure liquid chromatography purification of UP1 and UP2, two related single-stranded nucleic acid-binding proteins from calf thymus, *J Biol Chem 261*, 878-883.
177. Liu, Z., Frantz, J. D., Gilbert, W., and Tye, B. K. (1993) Identification and characterization of a nuclease activity specific for G4 tetrastranded DNA, *Proc Natl Acad Sci U S A 90*, 3157-3161.
178. Liu, Z., Lee, A., and Gilbert, W. (1995) Gene disruption of a G4-DNA-dependent nuclease in yeast leads to cellular senescence and telomere shortening, *Proc Natl Acad Sci U S A 92*, 6002-6006.
179. Todd, A. K., Johnston, M., and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA, *Nucleic Acids Res 33*, 2901-2907.
180. Huppert, J. L., and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome, *Nucleic Acids Res 33*, 2908-2916.
181. Huppert, J. L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res 36*, 6260-6268.
182. Eddy, J., and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, *Nucleic Acids Res 36*, 1321-1333.
183. Rawal, P., Kummarasetti, V. B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S. K., and Chowdhury, S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation, *Genome Res 16*, 644-655.
184. Hershman, S. G., Chen, Q., Lee, J. Y., Kozak, M. L., Yue, P., Wang, L. S., and Johnson, F. B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae, *Nucleic Acids Res 36*, 144-156.
185. Verma, A., Halder, K., Halder, R., Yadav, V. K., Rawal, P., Thakur, R. K., Mohd, F., Sharma, A., and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species, *J Med Chem 51*, 5641-5649.
186. Du, Z., Kong, P., Gao, Y., and Li, N. (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome, *Biochem Biophys Res Commun 354*, 1067-1070.
187. Zhao, Y., Du, Z., and Li, N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals, *FEBS Lett 581*, 1951-1956.
188. Eddy, J., and Maizels, N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome, *Nucleic Acids Res 34*, 3887-3896.
189. Du, Z., Zhao, Y., and Li, N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription, *Genome Res 18*, 233-241.
190. Du, Z., Zhao, Y., and Li, N. (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs, *Nucleic Acids Res 37*, 6784-6798.
191. Kumar, P., Yadav, V. K., Baral, A., Saha, D., and Chowdhury, S. (2011) Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide, *Nucleic Acids Res 39*, 8005-8016.

192. Bartlett, J. M., and Stirling, D. (2003) *PCR protocols*, Humana Press, Totowa.
193. Maxam, A. M., and Gilbert, W. (1977) A new method for sequencing DNA, *Proc Natl Acad Sci U S A 74*, 560-564.
194. Russell, R., Zhuang, X., Babcock, H. P., Millett, I. S., Doniach, S., Chu, S., and Herschlag, D. (2002) Exploring the folding landscape of a structured RNA, *Proc Natl Acad Sci U S A 99*, 155-160.
195. Seenisamy, J., Rezler, E. M., Powell, T. J., Tye, D., Gokhale, V., Joshi, C. S., Siddiqui-Jain, A., and Hurley, L. H. (2004) The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4, *J Am Chem Soc 126*, 8702-8709.
196. Matsugami, A., Okuizumi, T., Uesugi, S., and Katahira, M. (2003) Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA, *J Biol Chem 278*, 28147-28153.
197. Dapic, V., Abdomerovic, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J. O., and Bates, P. J. (2003) Biophysical and biological properties of quadruplex oligodeoxyribonucleotides, *Nucleic Acids Res 31*, 2097-2107.
198. Kelly, J. A., Feigon, J., and Yeates, T. O. (1996) Reconciliation of the X-ray and NMR structures of the thrombin-binding aptamer d(GGTTGGTGTGGTTGG), *J Mol Biol 256*, 417-422.
199. Mergny, J. L., Phan, A. T., and Lacroix, L. (1998) Following G-quartet formation by UV-spectroscopy, *FEBS Lett 435*, 74-78.
200. Balagurumoorthy, P., and Brahmachari, S. K. (1994) Structure and stability of human telomeric sequence, *J Biol Chem 269*, 21858-21869.
201. Li, W., Wu, P., Ohmichi, T., and Sugimoto, N. (2002) Characterization and thermodynamic properties of quadruplex/duplex competition, *FEBS Lett 526*, 77-81.
202. Selvin, P. R. (2000) The renaissance of fluorescence resonance energy transfer, *Nat Struct Biol 7*, 730-734.
203. Clegg, R. M. (1992) Fluorescence resonance energy transfer and nucleic acids, *Methods Enzymol 211*, 353-388.
204. Deniz, A. A., Laurence, T. A., Dahan, M., Chemla, D. S., Schultz, P. G., and Weiss, S. (2001) Ratiometric single-molecule studies of freely diffusing biomolecules, *Annu Rev Phys Chem 52*, 233-253.
205. Ying, L., Green, J. J., Li, H., Klenerman, D., and Balasubramanian, S. (2003) Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer, *Proc Natl Acad Sci U S A 100*, 14629-14634.
206. Mukhopadhyay, S., and Deniz, A. A. (2007) Fluorescence from diffusing single molecules illuminates biomolecular structure and dynamics, *J Fluoresc 17*, 775-783.
207. Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., and Ha, T. (2008) Advances in single-molecule fluorescence methods for molecular biology, *Annu Rev Biochem 77*, 51-76.
208. Ying, L. (2007) Single molecule biology: coming of age, *Mol Biosyst 3*, 377-380.
209. Deniz, A. A., Mukhopadhyay, S., and Lemke, E. A. (2008) Single-molecule biophysics: at the interface of biology, physics and chemistry, *J R Soc Interface 5*, 15-45.
210. Green, J. J., Ying, L., Klenerman, D., and Balasubramanian, S. (2003) Kinetics of unfolding the human telomeric DNA quadruplex using a PNA trap, *J Am Chem Soc 125*, 3763-3767.
211. Green, J. J., Ladame, S., Ying, L., Klenerman, D., and Balasubramanian, S. (2006) Investigating a quadruplex-ligand interaction by unfolding kinetics, *J Am Chem Soc 128*, 9809-9812.
212. Fegan, A., Shirude, P. S., Ying, L., and Balasubramanian, S. (2010) Ensemble and single molecule FRET analysis of the structure and unfolding kinetics of the c-kit promoter quadruplexes, *Chem Commun (Camb) 46*, 946-948.

213. Kumari, S., Bugaut, A., Huppert, J. L., and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation, *Nat Chem Biol 3*, 218-221.
214. Haider, S. M., Parkinson, G. N., and Neidle, S. (2003) Structure of a G-quadruplex-ligand complex, *J Mol Biol 326*, 117-125.
215. Haq, I., Trent, J. O., Chowdhry, B. Z., and Jenkins, T. C. (1999) Intercalative G-tetraplex stabilization of telomeric DNA by a cationic porphyrin, *Journal of the American Chemical Society 121*, 1768-1779.
216. Rodriguez, R., Pantos, G. D., Goncalves, D. P., Sanders, J. K., and Balasubramanian, S. (2007) Ligand-driven G-quadruplex conformational switching by using an unusual mode of interaction, *Angew Chem Int Ed Engl 46*, 5405-5407.
217. Sun, D., Thompson, B., Cathers, B. E., Salazar, M., Kerwin, S. M., Trent, J. O., Jenkins, T. C., Neidle, S., and Hurley, L. H. (1997) Inhibition of human telomerase by a G-quadruplex-interactive compound, *J Med Chem 40*, 2113-2116.
218. Perry, P. J., Gowan, S. M., Reszka, A. P., Polucci, P., Jenkins, T. C., Kelland, L. R., and Neidle, S. (1998) 1,4- and 2,6-disubstituted amidoanthracene-9,10-dione derivatives as inhibitors of human telomerase, *J Med Chem 41*, 3253-3260.
219. Perry, P. J., Reszka, A. P., Wood, A. A., Read, M. A., Gowan, S. M., Dosanjh, H. S., Trent, J. O., Jenkins, T. C., Kelland, L. R., and Neidle, S. (1998) Human telomerase inhibition by regioisomeric disubstituted amidoanthracene-9,10-diones, *J Med Chem 41*, 4873-4884.
220. Read, M. A., Wood, A. A., Harrison, J. R., Gowan, S. M., Kelland, L. R., Dosanjh, H. S., and Neidle, S. (1999) Molecular modeling studies on G-quadruplex complexes of telomerase inhibitors: structure-activity relationships, *J Med Chem 42*, 4538-4546.
221. Read, M., Harrison, R. J., Romagnoli, B., Tanious, F. A., Gowan, S. H., Reszka, A. P., Wilson, W. D., Kelland, L. R., and Neidle, S. (2001) Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors, *Proc Natl Acad Sci U S A 98*, 4844-4849.
222. Schultes, C. M., Guyen, B., Cuesta, J., and Neidle, S. (2004) Synthesis, biophysical and biological evaluation of 3,6-bis-amidoacridines with extended 9-anilino substituents as potent G-quadruplex-binding telomerase inhibitors, *Bioorg Med Chem Lett 14*, 4347-4351.
223. Moore, M. J., Schultes, C. M., Cuesta, J., Cuenca, F., Gunaratnam, M., Tanious, F. A., Wilson, W. D., and Neidle, S. (2006) Trisubstituted acridines as G-quadruplex telomere targeting agents. Effects of extensions of the 3,6- and 9-side chains on quadruplex binding, telomerase activity, and cell proliferation, *J Med Chem 49*, 582-599.
224. White, E. W., Tanious, F., Ismail, M. A., Reszka, A. P., Neidle, S., Boykin, D. W., and Wilson, W. D. (2007) Structure-specific recognition of quadruplex DNA by organic cations: influence of shape, substituents and charge, *Biophys Chem 126*, 140-153.
225. Gowan, S. M., Harrison, J. R., Patterson, L., Valenti, M., Read, M. A., Neidle, S., and Kelland, L. R. (2002) A G-quadruplex-interactive potent small-molecule inhibitor of telomerase exhibiting in vitro and in vivo antitumor activity, *Mol Pharmacol 61*, 1154-1162.
226. Burger, A. M., Dai, F., Schultes, C. M., Reszka, A. P., Moore, M. J., Double, J. A., and Neidle, S. (2005) The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase function, *Cancer Res 65*, 1489-1496.
227. Taetz, S., Baldes, C., Murdter, T. E., Kleideiter, E., Piotrowska, K., Bock, U., Haltner-Ukomadu, E., Mueller, J., Huwer, H., Schaefer, U. F., Klotz, U., and Lehr, C. M. (2006) Biopharmaceutical characterization of the telomerase inhibitor BRACO19, *Pharm Res 23*, 1031-1037.
228. Martins, C., Gunaratnam, M., Stuart, J., Makwana, V., Greciano, O., Reszka, A. P., Kelland, L. R., and Neidle, S. (2007) Structure-based design of benzylamino-acridine

compounds as G-quadruplex DNA telomere targeting agents, *Bioorg Med Chem Lett 17*, 2293-2298.
229. Ladame, S., Schouten, J. A., Stuart, J., Roldan, J., Neidle, S., and Balasubramanian, S. (2004) Tetrapeptides induce selective recognition for G-quadruplexes when conjugated to a DNA-binding platform, *Org Biomol Chem 2*, 2925-2931.
230. Redman, J. E., Granadino-Roldan, J. M., Schouten, J. A., Ladame, S., Reszka, A. P., Neidle, S., and Balasubramanian, S. (2009) Recognition and discrimination of DNA quadruplexes by acridine-peptide conjugates, *Org Biomol Chem 7*, 76-84.
231. Randazzo, A., Galeone, A., and Mayol, L. (2001) H-1-NMR study of the interaction of distamycin A and netropsin with the parallel stranded tetraplex [d(TGGGGT)](4), *Chem Commun*, 1030-1031.
232. Martino, L., Virno, A., Pagano, B., Virgilio, A., Di Micco, S., Galeone, A., Giancola, C., Bifulco, G., Mayol, L., and Randazzo, A. (2007) Structural and thermodynamic studies of the interaction of distamycin A with the parallel quadruplex structure [d(TGGGGT)]4, *J Am Chem Soc 129*, 16048-16056.
233. Zhang, H., Xiang, J. F., Hu, H. Y., Li, L., Jin, X., Liu, Y., Li, P. F., Tang, Y., and Chen, C. F. (2010) Stabilizing G-quadruplex DNA by a scissors-shaped binaphthyl derivative through the entangling mode: cooperation of binaphthylene and the ethoxy chain, *Biochemistry 49*, 10351-10353.
234. Jantos, K., Rodriguez, R., Ladame, S., Shirude, P. S., and Balasubramanian, S. (2006) Oxazole-based peptide macrocycles: a new class of G-quadruplex binding ligands, *J Am Chem Soc 128*, 13662-13663.
235. Monchaud, D., and Teulade-Fichou, M. P. (2008) A hitchhiker's guide to G-quadruplex ligands, *Org Biomol Chem 6*, 627-636.
236. Rentzeperis, D., Marky, L. A., Dwyer, T. J., Geierstanger, B. H., Pelton, J. G., and Wemmer, D. E. (1995) Interaction of minor groove ligands to an AAATT/AATTT site: correlation of thermodynamic characterization and solution structure, *Biochemistry 34*, 2937-2945.
237. Zaffaroni, N., Lualdi, S., Villa, R., Bellarosa, D., Cermele, C., Felicetti, P., Rossi, C., Orlandi, L., and Daidone, M. G. (2002) Inhibition of telomerase activity by a distamycin derivative: effects on cell proliferation and induction of apoptosis in human cancer cells, *Eur J Cancer 38*, 1792-1801.
238. Pagano, B., Fotticchia, I., De Tito, S., Mattia, C. A., Mayol, L., Novellino, E., Randazzo, A., and Giancola, C. (2010) Selective Binding of Distamycin A Derivative to G-Quadruplex Structure [d(TGGGGT)](4), *J Nucleic Acids 2010*.
239. Kaiser, M., De Cian, A., Sainlos, M., Renner, C., Mergny, J. L., and Teulade-Fichou, M. P. (2006) Neomycin-capped aromatic platforms: quadruplex DNA recognition and telomerase inhibition, *Org Biomol Chem 4*, 1049-1057.
240. Mergny, J. L., Lacroix, L., Teulade-Fichou, M. P., Hounsou, C., Guittat, L., Hoarau, M., Arimondo, P. B., Vigneron, J. P., Lehn, J. M., Riou, J. F., Garestier, T., and Helene, C. (2001) Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay, *Proc Natl Acad Sci U S A 98*, 3062-3067.
241. Xue, L., Ranjan, N., and Arya, D. P. (2011) Synthesis and spectroscopic studies of the aminoglycoside (neomycin)--perylene conjugate binding to human telomeric DNA, *Biochemistry 50*, 2838-2849.
242. Chakraborty, T. K., Arora, A., Roy, S., Kumar, N., and Maiti, S. (2007) Furan based cyclic oligopeptides selectively target G-quadruplex, *J Med Chem 50*, 5539-5542.
243. Shirude, P. S., Gillies, E. R., Ladame, S., Godde, F., Shin-Ya, K., Huc, I., and Balasubramanian, S. (2007) Macrocyclic and helical oligoamides as a new class of G-quadruplex ligands, *J Am Chem Soc 129*, 11890-11891.
244. Waller, Z. A., Shirude, P. S., Rodriguez, R., and Balasubramanian, S. (2008) Triarylpyridines: a versatile small molecule scaffold for G-quadruplex recognition, *Chem Commun (Camb)*, 1467-1469.
245. Bugaut, A., Jantos, K., Wietor, J. L., Rodriguez, R., Sanders, J. K., and Balasubramanian, S. (2008) Exploring the differential recognition of DNA G-

quadruplex targets by small molecules using dynamic combinatorial chemistry, *Angew Chem Int Ed Engl 47*, 2677-2680.
246. Phan, A. T., Kuryavyi, V., Gaw, H. Y., and Patel, D. J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter, *Nat Chem Biol 1*, 167-173.
247. Wei, C., Jia, G., Yuan, J., Feng, Z., and Li, C. (2006) A spectroscopic study on the interactions of porphyrin with G-quadruplex DNAs, *Biochemistry 45*, 6681-6691.
248. Parkinson, G. N., Ghosh, R., and Neidle, S. (2007) Structural basis for binding of porphyrin to human telomeres, *Biochemistry 46*, 2390-2397.
249. Goncalves, D. P., Rodriguez, R., Balasubramanian, S., and Sanders, J. K. (2006) Tetramethylpyridiniumporphyrazines--a new class of G-quadruplex inducing and stabilising ligands, *Chem Commun (Camb)*, 4685-4687.
250. Seenisamy, J., Bashyam, S., Gokhale, V., Vankayalapati, H., Sun, D., Siddiqui-Jain, A., Streiner, N., Shin-Ya, K., White, E., Wilson, W. D., and Hurley, L. H. (2005) Design and synthesis of an expanded porphyrin that has selectivity for the c-MYC G-quadruplex structure, *J Am Chem Soc 127*, 2944-2959.
251. Rezler, E. M., Seenisamy, J., Bashyam, S., Kim, M. Y., White, E., Wilson, W. D., and Hurley, L. H. (2005) Telomestatin and diseleno sapphyrin bind selectively to two different forms of the human telomeric G-quadruplex structure, *J Am Chem Soc 127*, 9439-9447.
252. Stanslas, J., Hagan, D. J., Ellis, M. J., Turner, C., Carmichael, J., Ward, W., Hammonds, T. R., and Stevens, M. F. (2000) Antitumor polycyclic acridines. 7. Synthesis and biological properties of DNA affinic tetra- and pentacyclic acridines, *J Med Chem 43*, 1563-1572.
253. Gowan, S. M., Heald, R., Stevens, M. F., and Kelland, L. R. (2001) Potent inhibition of telomerase by small-molecule pentacyclic acridines capable of interacting with G-quadruplexes, *Mol Pharmacol 60*, 981-988.
254. Gavathiotis, E., Heald, R. A., Stevens, M. F., and Searle, M. S. (2003) Drug recognition and stabilisation of the parallel-stranded DNA quadruplex d(TTAGGGT)4 containing the human telomeric repeat, *J Mol Biol 334*, 25-36.
255. Cookson, J. C., Heald, R. A., and Stevens, M. F. (2005) Antitumor polycyclic acridines. 17. Synthesis and pharmaceutical profiles of pentacyclic acridinium salts designed to destabilize telomeric integrity, *J Med Chem 48*, 7198-7207.
256. Cookson, J. C., Dai, F., Smith, V., Heald, R. A., Laughton, C. A., Stevens, M. F., and Burger, A. M. (2005) Pharmacodynamics of the G-quadruplex-stabilizing telomerase inhibitor 3,11-difluoro-6,8,13-trimethyl-8H-quino[4,3,2-kl]acridinium methosulfate (RHPS4) in vitro: activity in human tumor cells correlates with telomere length and can be enhanced, or antagonized, with cytotoxic agents, *Mol Pharmacol 68*, 1551-1558.
257. Keating, L. R., and Szalai, V. A. (2004) Parallel-stranded guanine quadruplex interactions with a copper cationic porphyrin, *Biochemistry 43*, 15891-15900.
258. Evans, S. E., Mendez, M. A., Turner, K. B., Keating, L. R., Grimes, R. T., Melchoir, S., and Szalai, V. A. (2007) End-stacking of copper cationic porphyrins on parallel-stranded guanine quadruplexes, *J Biol Inorg Chem 12*, 1235-1249.
259. Maraval, A., Franco, S., Vialas, C., Pratviel, G., Blasco, M. A., and Meunier, B. (2003) Porphyrin-aminoquinoline conjugates as telomerase inhibitors, *Org Biomol Chem 1*, 921-927.
260. Dixon, I. M., Lopez, F., Esteve, J. P., Tejera, A. M., Blasco, M. A., Pratviel, G., and Meunier, B. (2005) Porphyrin derivatives for telomere binding and telomerase inhibition, *Chembiochem 6*, 123-132.
261. Dixon, I. M., Lopez, F., Tejera, A. M., Esteve, J. P., Blasco, M. A., Pratviel, G., and Meunier, B. (2007) A G-quadruplex ligand with 10000-fold selectivity over duplex DNA, *J Am Chem Soc 129*, 1502-1503.

262. Bertrand, H., Bombard, S., Monchaud, D., and Teulade-Fichou, M. P. (2007) A platinum-quinacridine hybrid as a G-quadruplex ligand, *J Biol Inorg Chem 12*, 1003-1014.
263. Rao, L., and Bierbach, U. (2007) Kinetically favored platination of adenine in the g-rich human telomeric repeat, *J Am Chem Soc 129*, 15764-15765.
264. Bertrand, H., Monchaud, D., De Cian, A., Guillot, R., Mergny, J. L., and Teulade-Fichou, M. P. (2007) The importance of metal geometry in the recognition of G-quadruplex-DNA by metal-terpyridine complexes, *Org Biomol Chem 5*, 2555-2559.
265. Reed, J. E., Arnal, A. A., Neidle, S., and Vilar, R. (2006) Stabilization of G-quadruplex DNA and inhibition of telomerase activity by square-planar nickel(II) complexes, *J Am Chem Soc 128*, 5992-5993.
266. Bertrand, H., Bombard, S., Monchaud, D., and Teulade-Fichou, M. P. (2008) New platinum(II) complexes targeting the loops of the human telomeric G-quadruplex, *Nucleic Acids Symp Ser (Oxf)*, 163-164.
267. Suntharalingam, K., White, A. J., and Vilar, R. (2010) Two metals are better than one: investigations on the interactions between dinuclear metal complexes and quadruplex DNA, *Inorg Chem 49*, 8371-8380.
268. Shin-ya, K., Wierzba, K., Matsuo, K., Ohtani, T., Yamada, Y., Furihata, K., Hayakawa, Y., and Seto, H. (2001) Telomestatin, a novel telomerase inhibitor from Streptomyces anulatus, *J Am Chem Soc 123*, 1262-1263.
269. Kim, M. Y., Gleason-Guzman, M., Izbicka, E., Nishioka, D., and Hurley, L. H. (2003) The different biological effects of telomestatin and TMPyP4 can be attributed to their selectivity for interaction with intramolecular or intermolecular G-quadruplex structures, *Cancer Res 63*, 3247-3256.
270. Kim, M. Y., Vankayalapati, H., Shin-Ya, K., Wierzba, K., and Hurley, L. H. (2002) Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular g-quadruplex, *J Am Chem Soc 124*, 2098-2099.
271. Gomez, D., Paterski, R., Lemarteleur, T., Shin-Ya, K., Mergny, J. L., and Riou, J. F. (2004) Interaction of telomestatin with the telomeric single-strand overhang, *J Biol Chem 279*, 41487-41494.
272. Binz, N., Shalaby, T., Rivera, P., Shin-ya, K., and Grotzer, M. A. (2005) Telomerase inhibition, telomere shortening, cell growth suppression and induction of apoptosis by telomestatin in childhood neuroblastoma cells, *Eur J Cancer 41*, 2873-2881.
273. Tahara, H., Shin-Ya, K., Seimiya, H., Yamada, H., Tsuruo, T., and Ide, T. (2006) G-Quadruplex stabilization by telomestatin induces TRF2 protein dissociation from telomeres and anaphase bridge formation accompanied by loss of the 3' telomeric overhang in cancer cells, *Oncogene 25*, 1955-1966.
274. Gomez, D., O'Donohue, M. F., Wenner, T., Douarre, C., Macadre, J., Koebel, P., Giraud-Panis, M. J., Kaplan, H., Kolkes, A., Shin-ya, K., and Riou, J. F. (2006) The G-quadruplex ligand telomestatin inhibits POT1 binding to telomeric sequences in vitro and induces GFP-POT1 dissociation from telomeres in human cells, *Cancer Res 66*, 6908-6912.
275. Doi, T., Yoshida, M., Shin-ya, K., and Takahashi, T. (2006) Total synthesis of (R)-telomestatin, *Org Lett 8*, 4165-4167.
276. Doi, T., Shibata, K., Yoshida, M., Takagi, M., Tera, M., Nagasawa, K., Shin-ya, K., and Takahashi, T. (2011) (S)-stereoisomer of telomestatin as a potent G-quadruplex binder and telomerase inhibitor, *Org Biomol Chem 9*, 387-393.
277. Linder, J., Garner, T. P., Williams, H. E., Searle, M. S., and Moody, C. J. (2011) Telomestatin: formal total synthesis and cation-mediated interaction of its seco-derivatives with G-quadruplexes, *J Am Chem Soc 133*, 1044-1051.
278. Minhas, G. S., Pilch, D. S., Kerrigan, J. E., LaVoie, E. J., and Rice, J. E. (2006) Synthesis and G-quadruplex stabilizing properties of a series of oxazole-containing macrocycles, *Bioorg Med Chem Lett 16*, 3891-3895.

279. Tera, M., Sohtome, Y., Ishizuka, H., Doi, T., Takagi, M., Kazuo, S. Y., and Nagasawa, K. (2006) Design and synthesis of telomestatin derivatives and their inhibitory activity of telomerase, *Heterocycles 69*, 505-+.
280. Hamon, F., Largy, E., Guedin-Beaurepaire, A., Rouchon-Dagois, M., Sidibe, A., Monchaud, D., Mergny, J. L., Riou, J. F., Nguyen, C. H., and Teulade-Fichou, M. P. (2011) An Acyclic Oligoheteroaryle That Discriminates Strongly between Diverse G-Quadruplex Topologies, *Angew Chem Int Edit 50*, 8745-8749.
281. Kolb, H. C., Finn, M. G., and Sharpless, K. B. (2001) Click Chemistry: Diverse Chemical Function from a Few Good Reactions, *Angew Chem Int Ed Engl 40*, 2004-2021.
282. Evans, R. A. (2007) The rise of azide-alkyne 1,3-dipolar 'click' cycloaddition and its application to polymer science and surface modification, *Aust J Chem 60*, 384-395.
283. Moorhouse, A. D., and Moses, J. E. (2008) Click chemistry and medicinal chemistry: a case of "cyclo-addiction", *ChemMedChem 3*, 715-723.
284. Moses, J. E., and Moorhouse, A. D. (2007) The growing applications of click chemistry, *Chem Soc Rev 36*, 1249-1262.
285. Moorhouse, A. D., Santos, A. M., Gunaratnam, M., Moore, M., Neidle, S., and Moses, J. E. (2006) Stabilization of G-quadruplex DNA by highly selective ligands via click chemistry, *J Am Chem Soc 128*, 15972-15973.
286. Moorhouse, A. D., Haider, S., Gunaratnam, M., Munnur, D., Neidle, S., and Moses, J. E. (2008) Targeting telomerase and telomeres: a click chemistry approach towards highly selective G-quadruplex ligands, *Mol Biosyst 4*, 629-642.
287. Schouten, J. A., Ladame, S., Mason, S. J., Cooper, M. A., and Balasubramanian, S. (2003) G-quadruplex-specific peptide-hemicyanine ligands by partial combinatorial selection, *J Am Chem Soc 125*, 5594-5595.
288. Oganesian, L., and Bryan, T. M. (2007) Physiological relevance of telomeric G-quadruplex formation: a potential drug target, *Bioessays 29*, 155-165.
289. Kendrick, S., and Hurley, L. H. (2010) The role of G-quadruplex/i-motif secondary structures as cis-acting regulatory elements, *Pure Appl Chem 82*, 1609-1621.
290. Gunaratnam, M., Swank, S., Haider, S. M., Galesa, K., Reszka, A. P., Beltran, M., Cuenca, F., Fletcher, J. A., and Neidle, S. (2009) Targeting human gastrointestinal stromal tumor cells with a quadruplex-binding small molecule, *J Med Chem 52*, 3774-3783.
291. Riou, J. F., Guittat, L., Mailliet, P., Laoui, A., Renou, E., Petitgenet, O., Megnin-Chanet, F., Helene, C., and Mergny, J. L. (2002) Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands, *Proc Natl Acad Sci U S A 99*, 2672-2677.
292. Gomez, D., Mergny, J. L., and Riou, J. F. (2002) Detection of telomerase inhibitors based on g-quadruplex ligands by a modified telomeric repeat amplification protocol assay, *Cancer Res 62*, 3365-3368.
293. Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J. L., and Riou, J. F. (2004) Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing, *Nucleic Acids Res 32*, 371-379.
294. Brooks, T. A., and Hurley, L. H. (2010) Targeting MYC Expression through G-Quadruplexes, *Genes Cancer 1*, 641-649.
295. Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C. B., Proffitt, C., Trent, K., Whitten, J. P., Lim, J. K., Von Hoff, D., Anderes, K., and Rice, W. G. (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis, *Cancer Res 69*, 7653-7661.
296. Qin, Y., and Hurley, L. H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, *Biochimie 90*, 1149-1171.
297. Alvarez, R. H., Kantarjian, H. M., and Cortes, J. E. (2006) Biology of platelet-derived growth factor and its involvement in disease, *Mayo Clin Proc 81*, 1241-1257.

298. Dobrian, A., Wade, S. S., and Prewitt, R. L. (1999) PDGF-A expression correlates with blood pressure and remodeling in 1K1C hypertensive rat arteries, *Am J Physiol 276*, H2159-2167.
299. Parker, S. B., Wade, S. S., and Prewitt, R. L. (1998) Pressure mediates angiotensin II-induced arterial hypertrophy and PDGF-A expression, *Hypertension 32*, 452-458.
300. Li, L., Blumenthal, D. K., Terry, C. M., He, Y., Carlson, M. L., and Cheung, A. K. (2011) PDGF-induced proliferation in human arterial and venous smooth muscle cells: molecular basis for differential effects of PDGF isoforms, *J Cell Biochem 112*, 289-298.
301. Liao, C. H., Akazawa, H., Tamagawa, M., Ito, K., Yasuda, N., Kudo, Y., Yamamoto, R., Ozasa, Y., Fujimoto, M., Wang, P., Nakauchi, H., Nakaya, H., and Komuro, I. (2010) Cardiac mast cells cause atrial fibrillation through PDGF-A-mediated fibrosis in pressure-overloaded mouse hearts, *J Clin Invest 120*, 242-253.
302. Tuuminen, R., Nykanen, A. I., Krebs, R., Soronen, J., Pajusola, K., Keranen, M. A., Koskinen, P. K., Alitalo, K., and Lemstrom, K. B. (2009) PDGF-A, -C, and -D but not PDGF-B increase TGF-beta1 and chronic rejection in rat cardiac allografts, *Arterioscler Thromb Vasc Biol 29*, 691-698.
303. Ferrara, N., Gerber, H. P., and LeCouter, J. (2003) The biology of VEGF and its receptors, *Nat Med 9*, 669-676.
304. Tammela, T., Enholm, B., Alitalo, K., and Paavonen, K. (2005) The biology of vascular endothelial growth factors, *Cardiovasc Res 65*, 550-563.
305. Woolard, J., Bevan, H. S., Harper, S. J., and Bates, D. O. (2009) Molecular diversity of VEGF-A as a regulator of its biological activity, *Microcirculation 16*, 572-592.
306. Alon, T., Hemo, I., Itin, A., Pe'er, J., Stone, J., and Keshet, E. (1995) Vascular endothelial growth factor acts as a survival factor for newly formed retinal vessels and has implications for retinopathy of prematurity, *Nat Med 1*, 1024-1028.
307. Safran, M., and Kaelin, W. G., Jr. (2003) HIF hydroxylation and the mammalian oxygen-sensing pathway, *J Clin Invest 111*, 779-783.
308. Miyagi, Y., Chiu, L. L., Cimini, M., Weisel, R. D., Radisic, M., and Li, R. K. (2011) Biodegradable collagen patch with covalently immobilized VEGF for myocardial repair, *Biomaterials 32*, 1280-1290.
309. Adams, J. M., Difazio, L. T., Rolandelli, R. H., Lujan, J. J., Hasko, G., Csoka, B., Selmeczy, Z., and Nemeth, Z. H. (2009) HIF-1: a key mediator in hypoxia, *Acta Physiol Hung 96*, 19-28.
310. Ravi, R., Mookerjee, B., Bhujwalla, Z. M., Sutter, C. H., Artemov, D., Zeng, Q., Dillehay, L. E., Madan, A., Semenza, G. L., and Bedi, A. (2000) Regulation of tumor angiogenesis by p53-induced degradation of hypoxia-inducible factor 1alpha, *Genes Dev 14*, 34-44.
311. Shiojima, I., and Walsh, K. (2006) Regulation of cardiac growth and coronary angiogenesis by the Akt/PKB signaling pathway, *Genes Dev 20*, 3347-3365.
312. Chi, N. C., and Karliner, J. S. (2004) Molecular determinants of responses to myocardial ischemia/reperfusion injury: focus on hypoxia-inducible and heat shock factors, *Cardiovasc Res 61*, 437-447.
313. Hanakahi, L. A., Sun, H., and Maizels, N. (1999) High affinity interactions of nucleolin with G-G-paired rDNA, *J Biol Chem 274*, 15908-15912.
314. Khurts, S., Masutomi, K., Delgermaa, L., Arai, K., Oishi, N., Mizuno, H., Hayashi, N., Hahn, W. C., and Murakami, S. (2004) Nucleolin interacts with telomerase, *J Biol Chem 279*, 51508-51515.
315. Ginisty, H., Sicard, H., Roger, B., and Bouvet, P. (1999) Structure and functions of nucleolin, *J Cell Sci 112 (Pt 6)*, 761-772.
316. van Weerd, J. H., Koshiba-Takeuchi, K., Kwon, C., and Takeuchi, J. K. (2011) Epigenetic factors and cardiac development, *Cardiovasc Res 91*, 203-211.
317. Behe, M. J. (1995) An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes, *Nucleic Acids Res 23*, 689-695.

318. Brahmachari, S. K., Sarkar, P. S., Raghavan, S., Narayan, M., and Maiti, A. K. (1997) Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators, *Gene 190*, 17-26.
319. Bates, D. O., and Curry, F. E. (1996) Vascular endothelial growth factor increases hydraulic conductivity of isolated perfused microvessels, *Am J Physiol 271*, H2520-2528.
320. Ferrara, N., Houck, K. A., Jakeman, L. B., Winer, J., and Leung, D. W. (1991) The vascular endothelial growth factor family of polypeptides, *J Cell Biochem 47*, 211-218.
321. Mitchell, C. A., Rutland, C. S., Walker, M., Nasir, M., Foss, A. J., Stewart, C., Gerhardt, H., Konerding, M. A., Risau, W., and Drexler, H. C. (2006) Unique vascular phenotypes following over-expression of individual VEGFA isoforms from the developing lens, *Angiogenesis 9*, 209-224.
322. Bates, D. O., Cui, T. G., Doughty, J. M., Winkler, M., Sugiono, M., Shields, J. D., Peat, D., Gillatt, D., and Harper, S. J. (2002) VEGF165b, an inhibitory splice variant of vascular endothelial growth factor, is down-regulated in renal cell carcinoma, *Cancer Res 62*, 4123-4131.
323. Bevan, H. S., van den Akker, N. M., Qiu, Y., Polman, J. A., Foster, R. R., Yem, J., Nishikawa, A., Satchell, S. C., Harper, S. J., Gittenberger-de Groot, A. C., and Bates, D. O. (2008) The alternatively spliced anti-angiogenic family of VEGF isoforms VEGFxxxb in human kidney development, *Nephron Physiol 110*, p57-67.
324. Glass, C. A., Harper, S. J., and Bates, D. O. (2006) The anti-angiogenic VEGF isoform VEGF165b transiently increases hydraulic conductivity, probably through VEGF receptor 1 in vivo, *J Physiol 572*, 243-257.
325. Harper, S. J., and Bates, D. O. (2008) VEGF-A splicing: the key to anti-angiogenic therapeutics?, *Nat Rev Cancer 8*, 880-887.
326. van Rooij, E., Sutherland, L. B., Liu, N., Williams, A. H., McAnally, J., Gerard, R. D., Richardson, J. A., and Olson, E. N. (2006) A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure, *Proc Natl Acad Sci U S A 103*, 18255-18260.
327. Divakaran, V., and Mann, D. L. (2008) The emerging role of microRNAs in cardiac remodeling and heart failure, *Circ Res 103*, 1072-1083.
328. Thum, T., Catalucci, D., and Bauersachs, J. (2008) MicroRNAs: novel regulators in cardiac development and disease, *Cardiovasc Res 79*, 562-570.
329. Sacca, B., Lacroix, L., and Mergny, J. L. (2005) The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides, *Nucleic Acids Res 33*, 1182-1192.
330. Joachimi, A., Benz, A., and Hartig, J. S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities, *Bioorg Med Chem 17*, 6811-6815.
331. Kostadinov, R., Malhotra, N., Viotti, M., Shine, R., D'Antonio, L., and Bagga, P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences, *Nucleic Acids Res 34*, D119-124.
332. Mullen, M. A., Olson, K. J., Dallaire, P., Major, F., Assmann, S. M., and Bevilacqua, P. C. (2010) RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles, *Nucleic Acids Res 38*, 8149-8163.
333. Bugaut, A., Rodriguez, R., Kumari, S., Hsu, S. T., and Balasubramanian, S. (2010) Small molecule-mediated inhibition of translation by targeting a native RNA G-quadruplex, *Org Biomol Chem 8*, 2771-2776.
334. Collie, G. W., and Parkinson, G. N. (2011) The application of DNA and RNA G-quadruplexes to therapeutic medicines, *Chem Soc Rev 40*, 5867-5892.
335. Marcel, V., Tran, P. L., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M. P., Hall, J., Mergny, J. L., Hainaut, P., and Van Dyck, E. (2011) G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms, *Carcinogenesis 32*, 271-278.

336. Decorsiere, A., Cayrel, A., Vagner, S., and Millevoi, S. (2011) Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3 '-end processing and function during DNA damage, *Gene Dev 25*, 220-225.
337. Todd, A. K., and Neidle, S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements, *Nucleic Acids Res 36*, 2700-2704.
338. Raiber, E. A., Kranaster, R., Lam, E., Nikan, M., and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro, *Nucleic Acids Res 40*, 1499-1508.
339. Nemer, M., and Horb, M. E. (2007) The KLF family of transcriptional regulators in cardiomyocyte proliferation and differentiation, *Cell Cycle 6*, 117-121.
340. Fisch, S., Gray, S., Heymans, S., Haldar, S. M., Wang, B., Pfister, O., Cui, L., Kumar, A., Lin, Z., Sen-Banerjee, S., Das, H., Petersen, C. A., Mende, U., Burleigh, B. A., Zhu, Y., Pinto, Y. M., Liao, R., and Jain, M. K. (2007) Kruppel-like factor 15 is a regulator of cardiomyocyte hypertrophy, *Proc Natl Acad Sci U S A 104*, 7074-7079.
341. Haldar, S. M., Lu, Y., Jeyaraj, D., Kawanami, D., Cui, Y., Eapen, S. J., Hao, C., Li, Y., Doughman, Y. Q., Watanabe, M., Shimizu, K., Kuivaniemi, H., Sadoshima, J., Margulies, K. B., Cappola, T. P., and Jain, M. K. (2010) Klf15 deficiency is a molecular link between heart failure and aortic aneurysm formation, *Sci Transl Med 2*, 26ra26.
342. Kaczynski, J., Cook, T., and Urrutia, R. (2003) Sp1- and Kruppel-like transcription factors, *Genome Biol 4*, 206.
343. Srivastava, D. (2006) Making or breaking the heart: from lineage determination to morphogenesis, *Cell 126*, 1037-1048.
344. Risebro, C. A., Smart, N., Dupays, L., Breckenridge, R., Mohun, T. J., and Riley, P. R. (2006) Hand1 regulates cardiomyocyte proliferation versus differentiation in the developing heart, *Development 133*, 4595-4606.
345. Scott, I. C., Anson-Cartwright, L., Riley, P., Reda, D., and Cross, J. C. (2000) The HAND1 basic helix-loop-helix transcription factor regulates trophoblast differentiation via multiple mechanisms, *Mol Cell Biol 20*, 530-541.
346. Czubryt, M. P., and Olson, E. N. (2004) Balancing contractility and energy production: the role of myocyte enhancer factor 2 (MEF2) in cardiac hypertrophy, *Recent Prog Horm Res 59*, 105-124.
347. Frey, N., Katus, H. A., Olson, E. N., and Hill, J. A. (2004) Hypertrophy of the heart: a new therapeutic target?, *Circulation 109*, 1580-1589.
348. McKinsey, T. A., and Olson, E. N. (2005) Toward transcriptional therapies for the failing heart: chemical screens to modulate genes, *J Clin Invest 115*, 538-546.
349. Kim, Y., Phan, D., van Rooij, E., Wang, D. Z., McAnally, J., Qi, X., Richardson, J. A., Hill, J. A., Bassel-Duby, R., and Olson, E. N. (2008) The MEF2D transcription factor mediates stress-dependent cardiac remodeling in mice, *J Clin Invest 118*, 124-132.
350. Zot, A. S., and Potter, J. D. (1987) Structural aspects of troponin-tropomyosin regulation of skeletal muscle contraction, *Annu Rev Biophys Biophys Chem 16*, 535-559.
351. Baldwin, A. S., Jr., Kittler, E. L., and Emerson, C. P., Jr. (1985) Structure, evolution, and regulation of a fast skeletal muscle troponin I gene, *Proc Natl Acad Sci U S A 82*, 8080-8084.
352. Corin, S. J., Juhasz, O., Zhu, L., Conley, P., Kedes, L., and Wade, R. (1994) Structure and expression of the human slow twitch skeletal muscle troponin I gene, *J Biol Chem 269*, 10651-10659.
353. Adams, J. E., 3rd, Bodor, G. S., Davila-Roman, V. G., Delmez, J. A., Apple, F. S., Ladenson, J. H., and Jaffe, A. S. (1993) Cardiac troponin I. A marker with high specificity for cardiac injury, *Circulation 88*, 101-106.

354. Adams, J. E., 3rd, Schechtman, K. B., Landt, Y., Ladenson, J. H., and Jaffe, A. S. (1994) Comparable detection of acute myocardial infarction by creatine kinase MB isoenzyme and cardiac troponin I, *Clin Chem 40*, 1291-1295.
355. Antman, E. M., Tanasijevic, M. J., Thompson, B., Schactman, M., McCabe, C. H., Cannon, C. P., Fischer, G. A., Fung, A. Y., Thompson, C., Wybenga, D., and Braunwald, E. (1996) Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes, *N Engl J Med 335*, 1342-1349.
356. Alpert, J. S., Thygesen, K., Antman, E., and Bassand, J. P. (2000) Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction, *J Am Coll Cardiol 36*, 959-969.
357. Thygesen, K., Alpert, J. S., and White, H. D. (2007) Universal definition of myocardial infarction, *J Am Coll Cardiol 50*, 2173-2195.
358. Benamer, H., Steg, P. G., Benessiano, J., Vicaut, E., Gaultier, C. J., Aubry, P., Boudvillain, O., Sarfati, L., Brochet, E., Feldman, L. J., Himbert, D., Juliard, J. M., and Assayag, P. (1999) Elevated cardiac troponin I predicts a high-risk angiographic anatomy of the culprit lesion in unstable angina, *Am Heart J 137*, 815-820.
359. Wong, G. C., Morrow, D. A., Murphy, S., Kraimer, N., Pai, R., James, D., Robertson, D. H., Demopoulos, L. A., DiBattiste, P., Cannon, C. P., and Gibson, C. M. (2002) Elevations in troponin T and I are associated with abnormal tissue level perfusion: a TACTICS-TIMI 18 substudy. Treat Angina with Aggrastat and Determine Cost of Therapy with an Invasive or Conservative Strategy-Thrombolysis in Myocardial Infarction, *Circulation 106*, 202-207.
360. Kimura, A., Harada, H., Park, J. E., Nishi, H., Satoh, M., Takahashi, M., Hiroi, S., Sasaoka, T., Ohbuchi, N., Nakamura, T., Koyanagi, T., Hwang, T. H., Choo, J. A., Chung, K. S., Hasegawa, A., Nagai, R., Okazaki, O., Nakamura, H., Matsuzaki, M., Sakamoto, T., Toshima, H., Koga, Y., Imaizumi, T., and Sasazuki, T. (1997) Mutations in the cardiac troponin I gene associated with hypertrophic cardiomyopathy, *Nat Genet 16*, 379-382.
361. Bhavsar, P. K., Dhoot, G. K., Cumming, D. V., Butler-Browne, G. S., Yacoub, M. H., and Barton, P. J. (1991) Developmental expression of troponin I isoforms in fetal human heart, *FEBS Lett 292*, 5-8.
362. Bhavsar, P. K., Brand, N. J., Yacoub, M. H., and Barton, P. J. (1996) Isolation and characterization of the human cardiac troponin I gene (TNNI3), *Genomics 35*, 11-23.
363. Bhavsar, P. K., Dellow, K. A., Yacoub, M. H., Brand, N. J., and Barton, P. J. (2000) Identification of cis-acting DNA elements required for expression of the human cardiac troponin I gene promoter, *J Mol Cell Cardiol 32*, 95-108.
364. Barton, P. J., Cullen, M. E., Townsend, P. J., Brand, N. J., Mullen, A. J., Norman, D. A., Bhavsar, P. K., and Yacoub, M. H. (1999) Close physical linkage of human troponin genes: organization, sequence, and expression of the locus encoding cardiac troponin I and slow skeletal troponin T, *Genomics 57*, 102-109.

Chapter 2 Materials and Methods

2.1 Materials

2.1.1 General reagents

All stock chemicals and solutions were purchased from VWR International (UK), Invitrogen (UK), or Sigma-Aldrich (UK) unless indicated otherwise. Glassware and plastic ware purchased from VWR International, Anachem (UK), or Lab-Tek (Denmark). For biophysical studies, water was purified by MilliQ Reagent Grade Water Ultrapurification System (Millipore, USA), while for *in vitro* studies, water was further autoclaved for 20 minutes at 15 psi, 121 °C. Glassware, plastic ware, solutions and media used in *in vitro* studies were also autoclaved as described above prior to use. Pipette tips, chambered coverglass, and other disposable plastic ware used in biophysical studies were purchased ready-sterilized.

2.1.2 Oligos used in biophysical studies

All labeled and unlabeled DNA oligos used in biophysical studies were purchased from IBA Biotechnology (Germany) and HPLC purified (high-performance liquid chromatography, double HPLC purification for labeled oligos). The sequences and labeling sites of these oligos are listed in Table 2.1. Fluorophores (Cy3 or Cy5) were coupled to nucleotides via NHS (N-hydroxysuccinimide) esters.

Table 2.1: Details of oligos used in biophysical studies in solution.

Name	Sequence (5'-3')
MEF-I	CAG GCG TGG GGG CCA ATG GGA GGC ACC TGT GGG GGT GGC GGG GCG G^{Cy3}AG CCG GGG CTG GGG
MEF-II	CAG GCG TGG GGG CCA ATG GGA GGC ACC TGT GGG GGT GGC GGG GCG GAG CCG GGG CTG GG^{Cy3}G
MEF-III	CCC CAG CCC CGG CTC CGC CCC GCC ACC CCC ACA GG^{Cy5}T GCC TCC CAT TGG CCC CCA CGC CTG
MEF-IV	ACA GG^{Cy5}T GCC TCC CAT TGG CCC CCA CGC CTG
MEF-V	GGG GGT GGC GGG GCG GAG CCG GGG CTG GGG
MEF-VI	CCC CAG CCC CGG CTC CGC CCC GCC ACC CCC
Tr$_{MNS}$-I	TGG GTC TGA GGG AGG AGG GGC TGG GGG C
Tr$_{MNS}$-II	GCC CCC AGC CCC TCC TCC CTC AGA CCC A
Tr$_{MNS}$-III	^{Cy3}TGG GTC TGA GGG AGG AGG GGC TGG GGG CGA GGT AAA AGG ATA ATG GCT ACG GTG CGG ACG GC
Tr$_{MNS}$-IV	^{Cy3}TGG GTC TGA GGG AGG AGG GGC TGG GGG ^{Cy5}C
Tr$_{2\times MNS}$-I	TGG GTC TGA GGG AGG AGG GGC TGG GGG CCT GGA CTC CTG GGT CTG AGG GAG GAG GGG CTG GGG C
Tr$_{2\times MNS}$-II	GCC CCA GCC CCT CCT CCC TCA GAC CCA GGA GTC CAG GCC CCC AGC CCC TCC TCC CTC AGA CCC A
Tr$_{2\times MNS}$-III	^{Cy5}TGG GTC TGA GGG AGG AGG GGC TGG GGG CCT GGA CTC CTG GGT CTG AGG GAG GAG GGG CTG GGG ^{Cy3}C
Tr$_{-80}$-I	CGG GGG CGC GTG AGG GGC GGG GTG GGC
Tr$_{-80}$-II	GCC CAC CCC GCC CCT CAC GCG CCC CCG
Tr$_{-80}$-III	^{Cy3}CGG GGG CGC GTG AGG GGC GGG GTG GGC GAG GTA AAA GGA TAA TGG CTA CGG TGC GGA CGG C
Tr$_{-80}$-IV	^{Cy3}CGG GGG CGC GTG AGG GGC GGG GTG GG ^{Cy5}C
Comp-Cy5	GCC GTC CGC ACC GTA GCC ATT ATC CTT ^{Cy5}TTA CCT C

Note: $^{Cy3/5}$N indicates a Cy3 or Cy5-labeled nucleotide, which is coupled via NHS ester.

2.1.3 Primers and synthetic oligos used in *in vitro* studies

All primers and synthetic oligos that are used in *in vitro* studies were synthesised to order by Eurofins MWG Operon (Germany) except where otherwise indicated. All primers and synthetic oligos were subjected to HPLC purification, and their sequences are listed in Table 2.2.

Table 2.2: Details of primers and synthetic oligos used in *in vitro* studies.

Name	Sequences (5' - 3')
Primers	
M13 Forward	GTA AAA CGA CGG CCA GT
M13 Reverse	CAG GAA ACA GCT ATG ACC
RV3	CTA GCA AAA TAG GCT GTC CCC AG
GL2	CTT TAT GTT TTT GGC GTC TTC CAT GG
5'-PCR-PstI	GGC TGC AGG AAA AGG GGG CTG GG
3'-PCR-XhoI	GGC TCG AGC CCA GGG ACC GTC AGT C
-80G4-G4M-F	CCC GCC CCT CAC GCG CTT TCG TTA TCT GGC ATA GT
-80G4-G4M-R	ACT ATG CCA GAT AAC GAA AGC GCG TGA GGG GCG GG
-80G4-sp1M1-F	GAG CAG GGC CCA CCC CAC CCC TCA CGC GCC CCC
-80G4-sp1M1-R	GGG GGC GCG TGA GGG GTG GGG TGG GCC CTG CTC
-80G4-sp1M2-F	GAG CAG GGC CCA CCC CTC CCC TCA CGC GCC CCC
-80G4-sp1M2-R	GGG GGC GCG TGA GGG GAG GGG TGG GCC CTG CTC
-80G4-M-F	GCA GGG CCC ACC CCG CTT CTC ACG CGC CCC CGT T
-80G4-M-R	AAC GGG GGC GCG TGA GAA GCG GGG TGG GCC CTG C
Synthetic Oligos	
MNS_G	CCT GGA CTC TTG GGT CTG AGG GAG GAG GGG CTG GGG G
MNS_C	TCC AGG CCC CCA GCC CCT CCT CCC TCA GAC CCA AGA G
MNSM_G	CCT GGA CTC TTG AAT CTG AGA GAG GAG AGA CTG AGA G
MNSM_C	TCC AGG CTC TCA GTC TCT CCT CTC TCA GAT TCA AGA G
5'Linker_G	GGG GAC AAG TTT GTA CAA AAA AGC AGG CTG AAT TC
5'Linker_C	TCC AGG GAA TTC AGC CTG CTT TTT TGT ACA AAC TTG TCC CC
3'Linker_G	CCT GGA CAC TTG AGT CTG CAG ACC CAG CTT TCT TGT ACA AAG TGG TCC CC
3'Linker_C	GGG GAC CAC TTT GTA CAA GAA AGC TGG GTC TGC AGA CTC AAG TG

2.1.4 Plasmids used in *in vitro* studies

The original plasmids pGL3-Basic_-533 and pGL3-Basic-Gateway were kindly provided by Dr. Nigel Brand (Heart Science Centre, Harefield, Imperial College London). All other Gateway® vectors and Gateway® reaction kits used in this study were purchased from Invitrogen. All plasmids that were used in this study are listed in Table 2.3. Modifications of these plasmids will be described in detail in Section 2.3.

Table 2.3: Details of plasmids used in *in vitro* studies.

Plasmid	Brief Description	Size (kb)	Antibiotic Used for Selections	Reference
pRL-TK	*Renilla* luciferase reporter vector driven by herpes simplex virus thymidine kinase promoter P_{HSV-TK}.	4	Ampicillin	Promega, UK
pGL3-Basic-GW	Promoterless pGL3-Basic vector with the Gateway®*attR* Reading Frame Cassette B (RfB).	6.5	Ampicillin	[1]
pDONR™221	Gateway® donor vector to give entry vectors via BP reaction.	4.7	Kanamycin	Invitrogen
hTnIc-6MNS-WT	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by wild type hTnIc -531/+67.	5.4	Ampicillin	[1]
hTnIc-6MNS(E)-WT	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by 6 repeats of consensus MNSG4 and wild type hTnIc -299/+51.	5.4	Ampicillin	This study
hTnIc-3MNS(E)-WT	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by 3 repeats of consensus MNSG4 and wild type hTnIc -299/+51.	5.3	Ampicillin	This study
hTnIc-1MNS(E)-WT	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by 1 repeats of consensus MNSG4 and wild type hTnIc -299/+51.	5.2	Ampicillin	This study
hTnIc-299-WT	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by wild type hTnIc -299/+51.	5.1	Ampicillin	This study
hTnIc-80G4M	Similar as hTnIc-6MNS(E)-WT, but the -80G4-forming sequence was mutated and the Sp1 binding site is intact.	5.4	Ampicillin	This study
hTnIc-80sp1M-1	Similar as hTnIc-6MNS(E)-WT, but the Sp1 binding site was mutated (G to A), and -80G4-forming sequence is intact.	5.4	Ampicillin	This study
hTnIc-80sp1M-2	Similar as hTnIc-6MNS(E)-WT, but the Sp1 binding site was mutated (G to T), and -80G4-forming sequence is intact.	5.4	Ampicillin	This study
hTnIc-80G4/sp1M	Similar as hTnIc-6MNS(E)-WT, but both of the -80G4-forming sequence and Sp1 binding site was mutated (CCCCGCCCC to CCCCGCTTC).	5.4	Ampicillin	This study
hTnIc-6MNS-M	Firefly luciferase reporter vector (pGL3-Basic-GW) driven by 6 repeats of mutated MNSG4 (G4-forming sequence disrupted) and wild type hTnIc -299/+51.	5.4	Ampicillin	This study
hTnIc-6MNS/80G4-M	Similar as hTnIc-6MNS-M, but both of the -80G4-forming sequence and Sp1 binding site was mutated (CCCCGCCCC to CCCCGCTTC).	5.4	Ampicillin	This study

2.1.5 Strains and media

Bacteria (*E. coli.*) were grown at 37 °C in Luria-Bertani (LB) broth or on LB agar (2 %) plates containing ampicillin (100 μg/ml) or kanamycin (50 μg/ml) as required. Bacteria strains are listed in Table 2.4, and the recipes used to prepare bacteria media are listed in Table 2.5.

Table 2.4: Bacterial strains used in *in vitro* studies.

Bacterial Strain	Genotype / Construction	Reference
E. coli DH5α (sub-cloning efficiency competent cells)	$F^-\phi$80*lacZ*Δ*M15*, Δ(*lacZY A-argF*), U169, *deoR*, *recA1*, *endA1*, *hsdR17*(r_k^-, m_k^+), *phoA*, *supE44*, *thi-1*, *gyrA96*, *relA1*, λ^-	Invitrogen
E. coli XL-Blue competent cells	Δ(*mcrA*)*183*, Δ(*mcrCB-hsdSMR-mrr*)*173*, *endA1*, *supE44*, *thi-1*, *recA1*, *gyrA96*, *relA1*, *lac*[F' *proAB lac1^{q}Z* Δ*M15* Tn*10*(Tet^r)]	Stratagene, USA

Table 2.5: Recipes to prepare bacterial media.

Media	Recipe to Prepare 1 Litre of Media
LB Broth	10 g/l Bacto-tryptone, 5 g/l yeast extract, and 10 g/l NaCl
LB Agar	10 g/l Bacto-tryptone, 5 g/l yeast extract, 10 g/l NaCl, and 2 % agar (w/v)

Note: for LB broth, antibiotics were added just prior to use; while for LB agar plates, antibiotics were added when the media had cooled to 45 °C.

2.1.6 Cell line and media

Human Embryonic Kidney 293 (HEK293) cell line was a kind gift from Dr. Nigel Brand. HEK293 cells were inoculated in 75 cm^2 flasks and grown to 70-80 % confluence at 37 °C and under 5 % CO_2 atmosphere in complete medium DMEM (Dulbecco Modified Eagle's Medium) supplemented with 4500 mg/L D-glucose, 5.0 mM L-glutamine, 10 % fetal calf serum, and 83.3 U/ml of penicillin and streptomycin. All media and supplements were purchased from Sigma Aldrich.

2.1.7 G4-binding ligands

G4-binding ligands used in this study were designed and synthesised in Prof Ramon Vilar's group (Department of Chemistry, Imperial College London). The di-copper and di-platinum complexes (compound 2 and 3), as well as the precursor of [Pt(terpy-

O-morph)Cl]Cl for the preparation of compound 1, were synthesised following the protocol that has been recently reported [2]. Structures of ligands are shown in Figure 5.6.

2.2 Biophysical characterisations

2.2.1 Oligo characterisations

The lyophilized oligos (labeled and unlabeled) purchased from IBA Biotechnology were dissolved in MilliQ water to 20 μM concentration as stock solutions according to the reported synthesis scale measured by the manufacturer. The actual concentration of each oligo stock solution was determined by UV-Vis spectroscopy technique at room temperature. It has been reported that the base stacking in G4s can result in hyperchromicity, and hence changes the extinction coefficient of the G4-forming oligo [3]. Due to the lack of UV-Vis thermal denaturation equipment, an alternative approach was used. The stock solution was firstly incubated at 95 °C for 15 minutes to disrupt all secondary structures, then loaded into a pre-heated cuvette with optical path length of 10 mm (Perkin Elmer, USA). The absorbance at 260 nm for DNA (A_{260}), 550 nm for Cy3 (A_{550}) and 650 nm for Cy5 (A_{650}) were measured immediately. DNA or fluorophore concentration was determined by following equation 2.1 according to Beer's law.

$$c = \frac{A}{\varepsilon b} \qquad \text{Eq. 2.1}$$

where A is the measured absorbance, ε is the molar extinction coefficient for DNA oligo or fluorophores, while b is the optical path length. The ε of each oligo was calculated using the nearest-neighbour method [4, 5], and ε for Cy3 and Cy5 were measured to be 150,000 and 250,000 M^{-1} cm^{-1} [6]. For Cy3/Cy5 dual-labeled oligos, crosstalk between fluorophores was determined to be 13 %. Then, all oligo stock solutions were adjusted to 5 μM concentration.

2.2.2 G4 preparation

The G4-forming and non-G4-forming DNA oligos were first dissolved in 10 mM Tris-HCl buffer (pH 7.4) with or without the presence of cation ions (100 mM KCl, NaCl, or LiCl) as required to a strand concentration of 5 μM. Prior to subsequent experiments, the samples containing G4-forming DNA oligos were heated to 95 °C for 15 minutes and then slowly cooled down to room temperature over 4 hours to allow the formation of G4s. For samples with non-G4-forming DNA oligos, samples were left on ice after incubating at 95 °C for 15 minutes to reduce the formation of other DNA secondary structures. Before each measurement, the stock solutions were further diluted to lower concentration as required.

2.2.3 Electrophoresis mobility shift assay (EMSA)

EMSA experiments were carried out by using the Mini-PROTEAN Tetra Electrophoresis System (BioRad, UK). 12 % Native and denaturing PAGE were prepared as described in Table 2.6. Upon the addition of TEMED (Tetramethylethylenediamine) and APS (ammonium persulfate), the gel was poured immediately to the pre-assembled plates carefully in order to prevent bubble formation. Then, the 1 mm thick comb was inserted and the gel was allowed to set for at least 30 minutes, using the polymerisation control to verify that the gel polymerised between the plates. The electrophoresis apparatus was assembled and the plates and gel were moved into the tank compartment of the electrophoresis chamber. The inner (top) chamber was filled with fresh 1 × TBE (tris-borate-EDTA buffer), and the comb was washed with 1 × TBE to get rid of debris. The apparatus was left for 30 minutes to check if there was any leakage. Once the gel had fully set, the comb was gently removed, and the wells were thoroughly flushed to avoid gel particles and bubbles by using a 200 μl pipette filled with 1× TBE. Then the outer (bottom) chamber was filled with 300 ml fresh 1 × TBE. All DNA oligo samples in denaturing EMSA and the control samples in native EMSA were heated to 95 °C for 15 minutes, and then left on ice before being loaded into the gel. DNA oligo samples were diluted with 1 × TBE buffer containing 5 % glycerol. 5 μl of diluted DNA oligo samples were loaded into individual well. For denaturing gel, each well was flushed by using a needle-like

pipette filled with 1 × TBE buffer to push out the high concentrated urea before loading the samples. 5 µl 1 × TBE buffer containing 5 % glycerol and 0.015 % Bromophenol Blue was loaded to an empty well to indicate the position of DNA oligos during electrophoresis. The whole electrophoresis apparatus was assembled and transferred it to an ice box to avoid unfolding of G4s induced by the heat generated during the processes of electrophoresis. For the native EMSA, both PAGE and TBE buffer contained 100 mM KCl. Control experiments were carried out by loading the same samples into 7 M urea denaturing PAGE in 1 × TBE buffer without KCl. All EMSA experiments were conducted at 90 V. Once the Bromophenol Blue band reached ¼ to bottom of the gel, the electrophoresis was terminated, and the buffer chambers of the apparatus were emptied. The apparatus was disassembled and the plates were unclamped, leaving only the gel attached between the two plates. The two glass plates were carefully separated, and the gel was transferred to a staining box. After rinsing the gel twice by MilliQ water, the gel was stained with 50 ml 1 × TBE buffer containing 1 × SYBR® Gold for 30 minutes. Stained gel was scanned by using an Ettan DIGE imager (GE Healthcare, UK) on the Cy2 (for SYBR® Gold), Cy3 or Cy5 channel.

Table 2.6: Recipes to prepare 12 % native and denaturing gels for EMSA experiments.

Component	12 % Native PAGE	12 % Denaturing PAGE
30 % Acrylamide/Bis Solution (29:1)	4.8 ml	4.8 ml
5 × TBE	2.4 ml	2.4 ml
2M KCl	600 µl	--
Urea	--	5.05 g
ddH_2O	4 ml	Up to 12 ml
APS	50 µl	50 µl
TEMED	5 µl	5 µl

Note: This recipe is for 12 ml gel, which is sufficient for 2 mini gels.

2.2.4 CD spectroscopy and CD melting

The test G4-binding compounds were dissolved in DMSO (dimethyl sulfoxide) to yield a 1 mM stock solution. This was then diluted using 10 mM Tris-HCl buffer (10 mM, pH 7.4) supplemented with 100 mM cation ions to the appropriate concentrations. DNA oligo samples in the absence of G4-binding compounds were

diluted to 5 μM in Tris-HCl buffer (10 mM, pH 7.4) with 100 mM KCl, NaCl or LiCl. For the MEF2D G4-forming sequence, 12.5 μM concentration was used. Prior to CD measurements, the DNA oligo samples were annealed by heating to 95 °C for 15 minutes and then by cooling to room temperature over 4 hours. To test the stabilising ability of G4-binding compounds, DNA oligo samples at 5 μM concentration were mixed with 10 μM compounds, and annealed by following the same protocol. CD spectra were recorded on an Applied Photophysics Ltd. Chirascan spectrometer (Applied Photophysics, UK) at 25 °C with following parameters: bandwidth- 1 nm, spectral range- 230 ~ 330 nm, step size- 0.5 nm, and time per point- 1 second. 1 mm optical path length quartz cell was used. The baselines of all CD spectra were corrected by using blank buffer. The CD spectra were collected and processed using the Chirascan and Chirascan Viewer softwares (Applied Photophysics) respectively. CD melting spectra were recorded at 2 °C intervals while the temperature was increased at a rate of 2 °C/min with 2 minutes dwelling time for each temperature. All melting curves were analysed in Origin 7.0 (OriginLab Corporation, USA).

In CD melting experiments, the thermal denaturation (melting) processes of G4s were assumed to be a simple two-state transition from folded (*F*) state to unfolded state (*U*) with a equilibrium constant *K*:

$$F \overset{K}{\longleftrightarrow} U \qquad \text{Eq. 2.2}$$

The population of the folded state (p_f) will be related to the equilibrium constant as:

$$p_f = \frac{[F]}{[F]+[U]} = \frac{1}{1+K} \qquad \text{Eq. 2.3}$$

Thus, the observed ellipticity of G4s in the process of thermal denaturation is:

$$E_{(\text{measured ellipticity})} = (u-l)\frac{1}{1+K} + l \qquad \text{Eq. 2.4}$$

in which *u* is the mean residue ellipticity of 100 % folded G4s, *l* is the mean residue ellipticity of 100 % unfolded G4s.

The basic thermodynamic relationships that are relevant to G4 stability can be summarized as:

$$\Delta G = \Delta H - T\Delta S = -RT\ln(K) \qquad \text{Eq. 2.5}$$

where, ΔG is the Gibbs free energy change, ΔH and ΔS are the enthalpy and entropy changes, respectively, R is the gas constant, and T is the absolute temperature. Since at $T = T_m$ (melting temperature), $K = 1$, and thus $\Delta G = 0$, the relationship between ΔH and ΔS is:

$$\Delta S = \frac{\Delta H}{T_m} \qquad \text{Eq. 2.6}$$

Thus, the Eq. 2.5 can be rewritten as:

$$\ln(K) = \frac{\Delta H}{RT}\left(\frac{T}{T_m} - 1\right) \qquad \text{Eq. 2.7}$$

and

$$K = \exp\left[\frac{\Delta H}{RT}\left(\frac{T}{T_m} - 1\right)\right] \qquad \text{Eq. 2.8}$$

Thus, by fitting the melting curves obtained at 263 or 295 nm with Eq. 2.4 and 2.8, the G4 unfolding van't Hoff enthalpy (ΔH) and melting temperature (T_m) can be derived. Then, by following Eq. 2.6 and 2.5, G4 unfolding van't Hoff entropy (ΔS) and free energy change at 37 °C ($\Delta G_{(310K)}$) can be calculated.

Because the calculated ΔH and T_m are intensively dependent on the choice of pre- and post-transition baselines, the truncated melting curves observed in certain cases (mostly due to the high melting temperature of certain G4s) would then jeopardize the accurate estimation of the stability of G4s. Thus, van't Hoff analysis was also carried out by using a model proposed by John *et al.* [7] as:

$$\frac{d(E_{\text{measured ellipticity}})}{dT} = Ap_f(1 - p_f)(T^2) \qquad \text{Eq. 2.9}$$

where $d(E_{\text{measured ellipticity}}) / dT$ is the algebraic derivative of thermal denaturation curve of G4s with arbitrary units, A is a scaling factor, p_f is population of folded state as described in Eq. 2.3, and T is absolute temperature. By fitting the first derivative curve of G4 melting curve with Eq. 2.3, 2.8, and 2.9, van't Hoff enthalpy (ΔH) and melting temperature (T_m) can be derived without baselines.

In analysing CD melting curves, both direct (Eq. 2.4 and 2.8) and indirect (Eq. 2.3, 2.8, and 2.9) methods were used and corresponding van't Hoff enthalpy (ΔH) and melting temperature (T_m) values were compared to assess the accuracy of fitted results.

2.2.5 Single molecule FRET studies

Experiment setup for smFRET measurement

The fluorescence detection system is a home-built dual-channel confocal fluorescence microscope composed of a laser adjustment module, a confocal far-field microscope, and a dual-channel fluorescence detection module [8, 9]. Individual components in each module could be modified and substituted with other optics to satisfy particular experimental applications. Compared to commercial fluorescence microscopy setups, the advantages of this home-built setup are (a) it is much less expensive and (b) it is more flexible. Only minor modifications are needed for different types of experiments. Optical components can be changed and adjusted readily. The simplified layout of this home-built setup is shown in Figure 2.1 and Table 2.7. To ensure the proper geometry of the observation volume and a good collection of fluorescence signals, the optical path needs to be aligned prior to single molecule measurements. For this alignment, a diluted solution of single-labeled molecules, for example single-labeled DNA, is measured at high laser power simultaneously with two APD (SPCM-ARQ-14, Perkin Elmer) (using 50/50 dichroic mirror).

To adjust the proper excitation and observation volumes, protected aluminum mirrors (B components in Figure 2.1) need to be tuned to achieve a sharp and bright laser dot right in the center of the ocular field of view. To ensure the good collection of fluorescence, the position of the 50/50 splitter (component J in Figure 2.1) and the two APDs need to be finely tuned in order to be sure that the fluorescence signals reached their maximum values. By tuning these components, it is ensured that the fluorescence signals were focused at the centre of the APDs and the two channels received similar amount of signals (50/50 splitter is in correct position and other filters have similar detection efficiencies) when the signals from both APDs were

recorded by two PC-implemented multichannel scalar card (MCS-PCI, EG & G, ORTEC, USA).

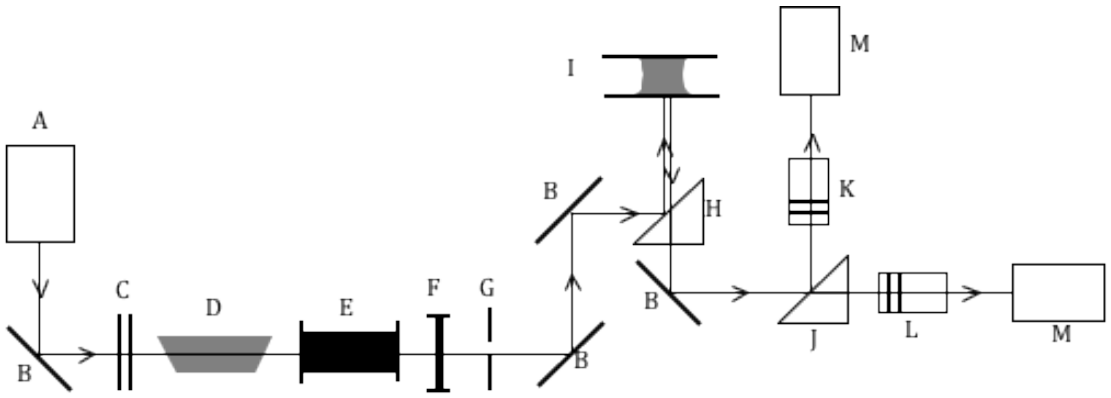


Figure 2.1: Simplified outline of the home-built dual-channel confocal single-molecule fluorescence microscope setup.
Descriptions of components A ~ M are listed in Table 2.7.

Table 2.7: Description of components of the home-built dual-channel confocal single-molecule fluorescence microscope setup.

Component	Description
A	Laser emitter (488 nm, argon ion, model 35LAP321230. Melles Griot, Carlsbad, CA)
B	Protected aluminum mirrors
C	Neutral density filters—adjust laser power
D	Prism—filter plasma lines from laser head
E	Spatial Filter
F	Collimate Lens
G	Collimator, adjust laser beam radius
H	Long pass dichroic mirror
I	Excitation volume in dilute solution
J	50/50 splitter, or long pass dichroic mirror
K	Long pass and band pass filters
L	Long pass and band pass filters
M	Avalanche photodiode detectors (APDs)

smFRET data acquisition and analysis

In each smFRET experiment, background noises of donor and acceptor channels were first determined by measuring the blank buffer under the same conditions. Background noises were normally less than 10 counts per millisecond if the buffer was suitable for smFRET measurements. Fluorophore-labeled DNA oligo samples were diluted from pre-annealed stock solutions to 100 pM in 10 mM Tris-HCl (pH 7.4) buffer containing 200 μM Vitamin C, and 0.05 % (v/v) Tween 20 to reduce photobleaching and surface absorption of DNA molecules. The donor, Cy3, was exited by a 100 μW collimated laser beam at 514 nm (argon ion, model 35LAP321-230, Melles Griot, USA). Fluorescent signals were detected by two APDs and recorded by two multichannel scalar cards. smFRET signals were recorded by MCS-32 software (EG & G, ORTEC) with following paramters: Pass length-8,000, Pass count preset-1, Dwell-internal, Bin width-1 ms, Triggle-channel 1 (donor channel) internal, channel 2 (acceptor channel) external, Acquire mode-replace, and Input control-50 Ohm. Generally, each smFRET measurement for labeled DNA oligos lasted 800 sec, and bin times were set to 1 ms (for smFRET G4 unfolding kinetics measurement, longer measuring time was used). The fluorescence signals from the donor and acceptor were counted as n_D and n_A (number of photons in each burst) respectively. The burst from individual fluorophore-labeled molecule was differentiated as a FRET event when the sum of photons in the donor and acceptor channels exceeded a threshold of 35 photons per millisecond. Cross-talk between Cy3 and Cy5 was determined to be 13 %. FRET efficiencies of each FRET burst was calculated with Eq. 2.10:

$$E_{FRET} = \frac{n_A}{n_A + \gamma n_D} \qquad \text{Eq. 2.10}$$

In this equation, n_D and n_A are the number of photons in the burst of donor and acceptor, respectively, while γ is a constant used to compensate the differences that are due to quantum yields and optical detection efficiencies. γ proved to be close to 1 between Cy3/Cy5 fluorophore pair in previous experiments using the same setup [8]. Histograms of FRET efficiencies of all molecules in each measurement were plotted using Origin 7.0 (E_{FRET} between -0.2 and 1.2, and interval 0.04). Histograms were

analysed by a multiple-peak Gaussian distribution model to reveal individual subpopulations as described in Eq. 2.11:

$$y = y_0 + \sum_{i=1}^{n} \frac{A_i}{w_i\sqrt{\pi/2}} e^{-2\frac{(x-xc_i)^2}{w_i^2}} \qquad \text{Eq. 2.11}$$

where y is the number of event in the histogram, y_0 is the baseline, which is close to 0 when background signals were properly subtracted, x is the FRET efficiency, A_i, w_i, and xc_i are the fraction, width, and the central FRET efficiency of the i^{th} subpopulation. Histograms obtained from six independent measurements were subjected to global fitting to obtain the optimal sets of parameters for the each subpopulation observed.

Conformational analysis by smFRET

The home-built dual-channel confocal fluorescence microscope was used to detect the conformational heterogeneity of G4s in buffer solutions. For double-stranded (ds) dual-labeled G4 systems (Figure 4.3 a and b, and 4.5 a), G4 samples were prepared by mixing 100 nM Cy3-labeled G4-forming DNA oligos with 150 nM Cy5-labeled 5' (MEF2DG4 system, oligo **MEF-IV**) or 3' (TnIc G4 systems, oligo **Comp-Cy5**) linker in 10 mM Tris-HCl buffer (pH 7.4) with or without 100 mM salts (LiCl, NaCl, or KCl) as required. Excess of Cy5-labeled oligo will improve the efficiency of annealing between Cy3-labeled G4-forming oligos and Cy5-labeled linkers, without changing the FRET results. The samples were annealed to allow the formation of G4s by following protocols as previously described (Section 2.2.2). For single-stranded (ss) dual-labeled G4 systems, 100 nM dual-labeled G4-forming oligos were prepared by following similar protocol. Prior to smFRET experiments, annealed samples were further diluted to 100 pM in 10 mM This-HCl buffer (pH 7.4) in the presence of 200 μM Vitamin C, 0.05 % (v/v) Tween 20, and 100 mM salts. Tris-HCl buffer with Vitamin C and Tween 20 was freshly prepared and subjected to background signal check. 1 ml of diluted samples were loaded to the 4-well chambered coverglass. To minimize the variation caused by temperature jumping, coverglass with diluted samples were left on the microscope stage for 30 minutes before each measurement, whose temperature was maintained at 25 °C by a water recirculation system (PE100, Linkam Scientific Instruments, UK). Then, the Cy3 was exited by a 100 μW

collimated laser beam at 514 nm. Fluorescence signals were detected by two APDs and recorded by two multichannel scalar cards and MSC-32 software. Each measurement lasted 800 seconds with 1 ms bin width. Other parameters were set as previously described. For each conformational analysis, 6 independent measurements were carried out and globally fitted in Origin 7.0 with multiple-peak Gaussian models.

G4/duplex competition assay by smFRET

Stability of G4s in the presence of C-rich complimentary strand was assessed by the G4/duplex competition assays. For the MEF2D ds dual-labeled G4 systems, 100 nM of the Cy3-labeled G-rich strand (oligo **MEF-I** or **II**) was mixed with either 150 nM Cy5-labeled 5' linker (oligo **MEF-IV**) and 150 nM C-rich strand complimentary (oligo **MEF-VI**, Figure 4.3 d) to the G4-forming region or 150 nM Cy5-labeled C-rich complementary strand with full length (oligo **MEF-III**, Figure 4.3 c). For the ds dual-labeled TnIc G4 systems (Figure 4.5 a), 100 nM Cy3-labeled G4-forming oligos (oligo **Tr_{MNS}-III** for MNSG4, and oligo **Tr_{-80}-III** for -80G4) were mixed with 150 nM of both Cy5-labeled 3' linker (oligo **Comp-Cy5**) and C-rich complimentary strand (oligo **Tr_{MNS}-II** for MNSG4, and oligo **Tr_{-80}-II** for -80G4). For the ss dual-labeled TnIc G4 systems (Figure 4.5 b and c), 100 nM Cy3/Cy5-labeled G4-forming oligos (oligo **Tr_{MNS}-IV**, **Tr_{-80}-IV**, or **$Tr_{2\times MNS}$-III**) were mixed with 150 nM C-rich complimentary strands (oligo **Tr_{MNS}-II**, **Tr_{-80}-II**, or **$Tr_{2\times MNS}$-II**, respectively). Then, the samples were annealed to allow the formation of G4s by following protocols as previously described (Section 2.2.2). smFRET data were acquired and processed by following the same protocol as described in the conformational analysis.

Unfolding kinetics of two-repeat MNSG4 by smFRET

G4 unfolding kinetics of the dual-labeled two-repeat MNSG4 system (Figure 4.5 c) were measured at single-molecule level using the same instrument as stated above. Basically, the dual-labeled two-repeat MNSG4 oligo (oligo **$Tr_{2\times MNS}$-III**) at 100 nM concentration was annealed in 100 mM K^+ solution (10 mM Tris-HCl, pH 7.4) as stated above. Then, the annealed sample was further diluted to 100 pM by Tris-HCl buffer (10 mM, pH 7.4) containing 200 μM Vitamin C, 0.05 % (v/v) Tween 20, and 100 mM K^+. A 4-well chambered coverglass with 1 ml diluted MNSG4 sample was left on the confocal microscope stage for 30 minutes, and temperature was maintained

at 25 °C. The C-rich complimentary strand (oligo **$Tr_{2\times MNS}$-II**) at 2 μM concentration in Tris-HCl buffer (10 mM, pH 7.4) was denatured at 95 °C for 15 minutes to disrupt other DNA secondary structures, and then immediately annealed on ice. smFRET histogram of the G4 sample was first generated in the absence of C-rich complimentary strand as a reference. Then, 1 μl of annealed C-rich strand was quickly mixed with 1 ml G4 samples in the chambered coverglass to a final concentration of 2 nM (20 folds more than the G-rich strand to mimic the pseudo-first order reaction). The time point when the mixing finished was set to be $t = 0$ for the unfolding reaction, and the decay of smFRET signals were measured for 7,900 seconds with 1 ms bin width. The excitation laser power was 100 μW. Solution background, cross talk between Cy3/Cy5 detection channels, and direct excitation of Cy5 were determined by blank buffer solution, and Cy3/Cy5-labeled samples.

In analyzing smFRET data, the background signals, cross talk between Cy3/Cy5, and direct excitation of Cy5 were subtracted, and a threshold of 35 photon per millisecond bin for the sum of the donor and acceptor fluorescence signals was used to differentiate single molecule bursts from the background. smFRET histograms of fluorescence signals obtained in every 160 seconds were generated as mentioned previously. Each histogram was normalized to the total number of its FRET events observed. All normalized histograms were globally fitted by a 4-peak Gaussian model with shared peak width, and centre (w_i, and xc_i in Eq. 2.11). The fraction of individual subpopulation (A_i in Eq. 2.11) was plotted against reaction time as unfolding kinetics traces, and fitted by either single (Eq. 2.12) or double exponential decay Eq. 2.13 according to the best fit.

$$F = Ae^{-t/\tau} + C \qquad \text{Eq. 2.12}$$

$$F = A_1e^{-t/\tau_1} + A_2e^{-t/\tau_2} + C \qquad \text{Eq. 2.13}$$

where F is the peak fraction of individual subpopulation in the smFRET histogram, t is the time after initiation of unfolding reaction, A, A_1, and A_2 represent the respective amplitude of decaying components, τ, τ_1, and τ_2 represent the unfolding time constant of individual decaying components, and C is the fluorescence at $t = \infty$.

2.2.6 Unfolding kinetics of G4s measured by FRET experiments at bulk level

MEF2D G4

G4 unfolding kinetics of the end dual-labeled MEF2DG4 system were measured at bulk level using the same instrument for smFRET as stated above. 100 nM G4 samples (oligos **MEF-II** 100 nM) in 100 mM K^+ solution (10 mM Tris-HCl, pH 7.4) were prepared as stated previously. Since a much lower laser power and much higher DNA concentration were used in bulk measurements, photobleaching and surface absorption effects are not as prominent as those observed at single molecule level. Thus, Vitamin C and Tween 20 were no longer used in kinetics experiments at bulk level. After annealing, the G4 sample was further diluted to 30 nM concentration in Tris-HCl buffer containing 100 mM K^+ before unfolding kinetics measurements. The C-rich complimentary strand (oligo **MEF-III**, 5 μM) was denatured at 95 °C for 15 minutes and then annealed on ice. Kinetics measurement was carried out by mixing Cy3-labeled G4-forming sequence (oligo **MEF-II**) with its Cy5-labeled complementary strand (oligo **MEF-III**) to a final concentration of 30 nM for both at time 0 of unfolding reaction. The mixing was carried out in a 4-well chambered coverglass. The excitation laser power was reduced to 35 μW to avoid fluorescence signal saturation. All kinetics measurements were carried out at 25 °C, and lasted 12,000 seconds with time bin width 1 second. Fluorescent signals from the acceptor channel (Cy5) was taken as unfolding kinetic trace, and fitted by a second order model as described in Chapter 4.

TnIc Two-repeat MNSG4

Unfolding kinetics of dual-labeled two-repeat MNSG4 was also investigated at bulk level by using the same setup. As compared to the second order unfolding reaction investigated in MEF2DG4 system, a pseudo-first order approach was used in studying the dual-labeled two-repeat MNSG4 (100 nM G4-forming oligo and 2 μM C-rich complimentary oligo). 100 nM dual-labeled two-repeat MNSG4 (oligo **$Tr_{2\times MNS}$-III**) was annealed as previously stated to allow the formation of G4. When the temperature was stabilised (25 °C), 20 μl DNA solution containing 100 μM C-rich complimentary strand (oligo **$Tr_{2\times MNS}$-II**) was mixed with 1 ml G4 samples (100 nM oligo, 10 mM Tris-HCl, and 100 mM K^+) at the time 0 of unfolding reaction. The excitation laser power was further reduced to 10 μW to avoid fluorescence signal saturation. Unfolding reaction was measured for 4,000 seconds with time bin width 1

second. The decrease of fluorescence signal in acceptor channel (Cy5) was taken as unfolding kinetic trace, and fitted by a double exponential decay model (Eq. 2.13).

TnIc MNS and -80 G4s

For both ds dual-labeled MNSG4 and -80G4 systems, their unfolding kinetics were investigated via a pseudo-first order approach. Generally, 100 nM Cy3-labeled MNSG4 (oligo **Tr_{MNS}-III**) and -80G4 (oligo **Tr_{-80}-III**) were mixed with 150 nM Cy5-labeled 3' linker (oligo **Comp-Cy5**) in 10 mM Tris-HCl buffer containing 100 mM K^+. Both G4 samples were annealed to allow the formation G4 as stated previously. Fluorescence signals of 200 μl annealed G4 samples at 100 nM concentration were measured by Fluorometer (Fluor Max-3, JOBIN YVON, HORIBA, USA). Samples were excited at 550 nm, and Cy5 fluorescence signals were collected at 664 nm with 5 nm slit (peak of Cy5 emission). Cy5 signals were measured at 1 second time interval with 0.5 second time bin. Intensity of Cy5 fluorescence signals in the absence of C-rich oligos was monitored for 30 minutes prior to each kinetics experiment. If the signal was found to be systematically changing, it was allowed to further stabilise. 4 μl DNA solution containing 100 μM C-rich complimentary strand (oligo **Tr_{MNS}-II** for MNSG4, and oligo **Tr_{-80}-II** for -80G4) was mixed with 200 μl G4 samples (100 nM oligo, 10 mM Tris-HCl, and 100 mM K^+) to give a final concentration of 2 μM at the time 0 of the unfolding reaction. The progress of G4 unfolding was monitored by following the decrease in Cy5 fluorescence as the duplex formed. Unfolding kinetics runs were repeated at different temperatures (MNSG4: 25, 35, 45, and 55 °C, and -80G4: 35, 45, 55 and 65 °C). In each measurement, the mixing time (adding C-rich complimentary oligo to the G4 samples) was controlled within 10 seconds, and hence has only little influence to determine unfolding kinetic parameter accurately. The cuvette was sealed by a cap in the measurement to prevent evaporation at higher temperature. The resulting traces of Cy5 fluorescence were fitted by a double exponential decay model (Eq. 2.13) in Origin 7.0.

2.2.7 Dimethyl sulfate (DMS) footprinting of TnIc MNSG4 and -80G4

As shown in Figure 2.2, the N7 of guanine is protected from methylation by DMS if it is involved in Hoogsteen hydrogen bonding of a classic G-tetrad, while guanines

involved in GC, and less commonly, GG base paring can still be fully or partially methylated by DMS. Under alkali conditions, the imidazole ring will be opened, and subsequent piperidine treatment can substitute the guanine base group and induce a breakage of the DNA backbone. DMS footprinting is a commonly used technique to resolve guanines involved in G-tetrad of G4s, which are fully or partially protected from oligo breakage induced by DMS and piperidine.

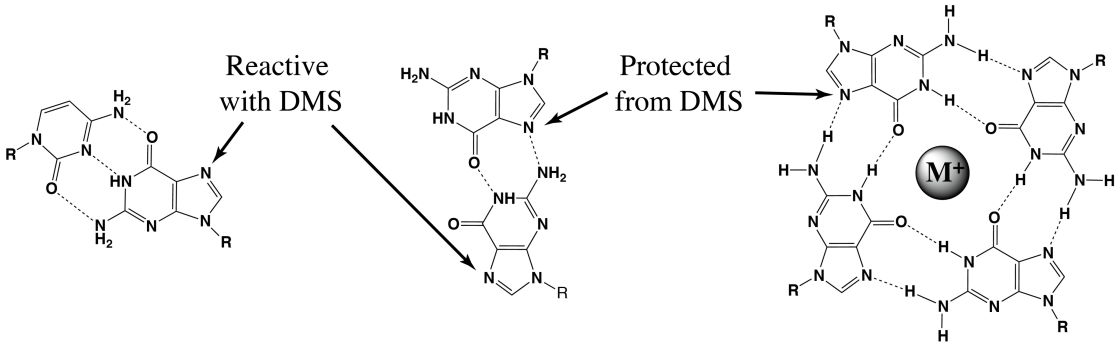


Figure 2.2: Molecular mechanism of DMS footprinting to reveal guanines involved in formation of G4 structures.

Labeling 5'-termini of oligos with [^{32}P]

The first step is to label test oligos with [^{32}P] at their 5'. The following components were added in order stated into a sterile 1.5 ml Eppendorf tube.

Test oligos (10 μM)	50 μl
gamma-^{32}P ATP (6,000 Ci/mmole, 10 mCi/ml)	15 μl
T4 polynucleotide kinase (10 U/μl)	5 μl
10 × kinase buffer	10 μl
Sterile MilliQ H_2O	20 μl

Excess of oligos were used in order to maximize labeling efficiency and decrease the radioactive waste. The reaction mixture was incubated at 37 °C for 1.5 hours in a heating block to label the 5'-termini of oligos. Then, the labeled oligos were separated from free gamma-^{32}P ATP by a Sephadex G-25 spin column (5 ml column centrifuged for 5 minutes at 600 × *g*), and collected in a 15 ml Falcon tube. The column was further eluted with 100 μl 1 × TE buffer (pH 7.4), and then disposed of properly. Gamma-^{32}P ATP was purchased from GE Healthcare.

First round electrophoresis (denaturing) to purify full-length oligos

Since the commercially synthesised oligos may be contaminated with incomplete oligos (i.e. shorter) or other impurities, a first round denaturing gel purification was used to separate labeled full-length oligos from other contaminants. A 30 cm × 16 cm × 1 mm gel was made by following a similar protocol as stated previously. Plates were sealed by adhesive tapes, and a comb with large wells (2 cm width) was used. 45 ml of 16 % polyacrylamide gel solution was prepared by mixing components in the order stated in Table 2.8. Once polymerised, the gel was pre-run at 300 V (constant voltage, DC power supply) to warm the gel for 30 minutes. In the process of pre-run, the labeled oligos were mixed with 20 µl alkaline gel loading buffer (10 ×), denatured at 95 °C for 15 minutes, and then held on ice before loading. Then, the sample was loaded and the gel was run at 200 V until desired resolution was obtained. After the first round electrophoresis was finished, the gel was transferred carefully to a Whatman 3MM paper (GE Healthcare), and then wrapped with plastic Saran wrap. Oligos of the desired length (usually the major product) were visualized by exposing the gel to a BioMax MR film (Kodak, USA) for 30 minutes. DNA bands were cut out with a razor blade, and crushed into small pieces with a metal spatula in a 1.5 ml Eppendorf tube. Gel fragments were soaked in 400 µl 1 × TE buffer (pH 7.4), and incubated on a rotation incubator at room temperature for 2 hours in order to elute the labeled oligos. The supernatant (~ 300 µl) was pipetted off into another 1.5 ml Eppendorf tube and subjected to ethanol precipitation. Generally, the supernatant was mixed with 1/10 volume of 3M NaOAC solution, 3 volumes ice-cold absolute ethanol, and incubated at -80 °C for 3 hours. The DNA sample was then pelleted by centrifuging at 13,000 rpm for 30 minutes at 4 °C. The supernatant was discarded and the DNA pellet was washed twice with 1 ml room-temperature 95 % ethanol. After discarding the ethanol, DNA pellet was air-dried on benchtop, and finally resuspended in 50 µl 1 × TE buffer (pH 7.4). The radioactivity of the labeled oligos were determined by Cerenkov counting, and further adjusted to 5,000 cpm/µl.

Second round electrophoresis (native) to obtain intramolecular G4s

The purified ^{32}P-labeled oligos were annealed to allow the formation of G4 in the presence of 100 mM K^+ as previously stated. A 16 % native polyacrylamide gel was also prepared as described in Table 2.8. The native gel was then pre-run at 150 V for

30 minutes in 1 × TBE buffer containing 100 mM K^+. 10 μl of annealed oligos (~ 50,000 cpm per reaction) were diluted with 0.1x TE buffer to 90 μl per reaction. After adding 1 μl of calf thymus DNA (0.1 μg/μl), DMS methylation reaction was carried out for 5 minutes by adding 1 μl of 25 % DMS solution (1:4 DMS/ethanol v/v). Each reaction was stopped by adding 18 μl of stop buffer (1:6:7 β-mercaptoethanol/water/NaOAc, v/l (3M NaOAc)), 10 μl of gel loading buffer containing 1 μg calf thymus DNA, and immediately loaded on the 16 % native gel for electrophoresis (Bromophenol Blue indicator should remain blue in colour once the methylation reaction is terminated: yellow or purple colour indicates dramatic pH changes in the reaction, which may disrupt formation of G4s). The gel was run at 150 V in 1 × TBE buffer containing 100 mM K^+ until desired resolution was achieved. Then, the gel was recovered and visualized by autoradiography as stated previously. In the process of electrophoresis, a negative control was also prepared (denatured oligos without annealing were DMS methylated directly). After electrophoresis, DNA bands correspond to methylated intramolecular G4s were cut off from the gel, and methylated oligos (including the negative control) were recovered by ethanol precipitation in the presence of 1 μg carrier DNA (calf thymus DNA) to improve the recovery efficiency. Air-dried DNA pellets were resuspended in 30 μl 1M piperidine solution, and heated at 95 °C for 30 minutes. Piperidine-treated DNA samples were air-dried in a speed vac under vacuum, re-washed by 30 μl sterile MilliQ water, and air-dried again in the speed vac. DNA pellets were resuspended in 10 μl alkaline sequencing dye, and stored at -20 °C prior to the third round sequencing electrophoresis.

Chemical sequencing reactions of the labeled oligos

Two aliquots of 10 μl purified DNA samples were diluted to 20 μl in the presence of 0.5 μg calf thymus DNA to prevent excessive chemical modifications. The aliquots were mixed with 20 μl formic acid and hydrazine for purin- and pyrimidine-specific sequencing reactions, respectively. Reactions were performed at room temperature for 30 minutes, and stopped by adding 60 μl sequencing stop solution. DNA samples were recovered by ethanol precipitation, and piperidine-treated as stated above. The air-dried DNA pellets were dissolved in 10 μl alkaline sequencing dye, and stored at -20 °C prior to the third round sequencing electrophoresis.

Third round electrophoresis (sequencing) to resolve guanines involved in G4 formation

A denaturing gel of 38.5 cm × 31 cm × 0.4 mm was prepared from 90 ml gel solution (Table 2.8). The gel was pre-run and warmed for 30 minutes at 1,500 V. During the pre-run, reactive DNA samples (chemical sequencing products, DMS treated G4s, and negative controls) were heated at 95 °C for 5 minutes and chilled on ice. Denatured DNA samples were then loaded on the sequencing gel, and the electrophoresis were performed at 1500 V until desired resolution has been obtained. After the electrophoresis, one plate was removed carefully and the gel was covered with a piece of a Whatman 3MM paper. The 3MM paper was slowly pulled back to transfer the gel to the paper. Then, another piece of paper was placed underneath, and the gel/paper was wrapped with a plastic wrap. The gel sandwich was then dried at 70 °C for 1 hour with a vacuum. The dried gel was exposed to a phosphor screen overnight (GE Healthcare). Then the exposed phosphor screen was scanned with a Typhoon^{TM} 9410 storage phosphor-imager (GE Healthcare) to digitize the results.

Table 2.8: Recipes for preparing 16 % denaturing, 16 % native, and 20 % denaturing polyacrylamide gels for the three rounds of electrophoresis in DMS footprinting experiments.

Component	16 % Denaturing PAGE (45 ml)	16 % Native PAGE (45 ml)	20 % Sequencing PAGE (90 ml)
30 % Acrylamide/Bis Solution (29:1)	24 ml	24 ml	60 ml
10 × TBE	4.5 ml	4.5 ml	18 ml
2M KCl	--	2.25 ml	--
Urea	22.5 g	--	45 g
ddH_2O	Up to 45 ml	14.25 ml	Up to 90 ml
APS (10 %)	150 µl	150 µl	200 µl
TEMED	30 µl	30 µl	50 µl

2.2.8 Ligand-G4 binding affinity assay in solution

Due to the strong fluorescence quenching ability of the di-copper complex (compound 3), the ligand-G4 binding affinity was determined by measuring the decrease of Cy5 fluorescence signals of G4s when titrated by the increase of ligand. Generally, 100

nM ss dual-labeled MNSG4 (oligo **Tr_{MNS}-IV**) and -80G4 (oligo **Tr_{-80}-IV**) were prepared as described in Section 2.2.2. Then, 1 ml of G4 samples (100 nM) were loaded to the 4-well chambered coverglass, and the Cy5 fluorescent signals without ligand were measured at room temperature (25 °C) for 30 seconds as initial fluorescence signals. After adding 1 μl of the ligand stock solution (1,000-fold stock solution), the sample was mixed thoroughly by pipetting and left for 10 minutes. Then, the quenched Cy5 fluorescence signals were measured for 30 seconds. The ligand quenching curves were fitted by following equations.

The formation of ligand-G4 complex (G4•L) can be described by a simple two-state process:

$$G4 + L \leftrightarrow G4 \bullet L \qquad \text{Eq. 2.14}$$

The corresponding dissociation constant (K_d) is defined as:

$$K_d = \frac{[G4][L]}{[G4 \bullet L]} \qquad \text{Eq. 2.15}$$

in which [G4], [L], and [G4•L] represent the concentration of free G4s, free ligands, and ligand-G4 complexes, respectively.

Because the total concentration of G4 ($[G4]_T$) and ligand ($[L]_T$) can be described as:

$$[G4]_T = [G4] + [G4 \bullet L] \qquad \text{Eq. 2.16}$$

$$[L]_T = [L] + [G4 \bullet L] \qquad \text{Eq. 2.17}$$

Eq. 2.15 can be rewritten as:

$$0 = -K_d[G4]_T + ([L]_T - [G4]_T + K_d)[G4] + [G4]^2 \qquad \text{Eq. 2.18}$$

By solving Eq. 2.18, the positive resolution of [G4] is:

$$[G4] = \frac{-([L]_T - [G4]_T + K_d) + \sqrt{([L]_T - [G4]_T + K_d)^2 + 4K_d[G4]_T}}{2} \qquad \text{Eq. 2.19}$$

Due to the strong fluorescence quenching ability of the ligand, it is reasonable to assume the ligand-G4 complex (G4•L) has no fluorescent signal. Thus, the fluorescent signal is solely determined by the concentration of free G4s ([G4]):

$$F_{[L]} = \frac{(F_{initial} - F_{final})}{[G4]_T}[G4] + F_{final} \qquad \text{Eq. 2.20}$$

in which $F_{[L]}$ is the fluorescent signal of the G4 sample titrated by a certain concentration of the ligand, $F_{initial}$ is the fluorescent signal of the G4 sample without ligand, and F_{final} represents the fluorescent signal of the G4 sample fully quenched by the ligand. By taking Eq. 2.19 into Eq. 2.20,

$$F_{[L]} = \frac{(F_{initial} - F_{final})\left(([G4]_T - [L]_T - K_d) + \sqrt{([L]_T - [G4]_T + K_d)^2 + 4K_d[G4]_T}\right)}{2[G4]_T} + F_{final} \qquad \text{Eq. 2.21}$$

By fitting the ligand quenching curves with Eq. 2.21, the dissociation constant K_d will be revealed, and correspondingly, its inverse is taken as the affinity constant (K_a).

2.3 Molecular and cell biology studies

2.3.1 Agarose gel electrophoresis, purification, and quantification of DNA

In order to check PCR or digested DNA products, or for the removal of enzymes and buffer salts, DNA samples were run routinely on a 1.5 % agarose gel in 1 × TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH 7.0). Generally, the agarose gel was prepared by melting 0.75 g agarose in 50 ml of 1 × TAE buffer in the microwave, adding/mixing 1 μl ethidium bromide solution (10 mg/ml) when the solution cooled to 60 °C, gently pouring the solution into a mini gel apparatus without introducing bubbles, and inserting a comb with proper size according to the sample volume. The gel was then left at room temperature to set for 30 minutes. After immersing the gel in an adequate volume of 1 × TAE buffer, the comb was carefully removed. One sixth volume of 6 × gel loading buffer (0.25 % Bromophenol Blue (w/v), 0.25 % xylene cyanol (w/v), and 30 % glycerol (v/v)) was mixed with DNA samples, and then loaded into the gel and run at 90 V (constant voltage) until desired resolution was obtained. DNA bands were visualized by the ChemiGenius documentation system under UV light (Syngene, UK). Bands of interest were carefully excised with a sterile scalpel on a UV transilluminator, and stored in sterile 1.5 ml Eppendorf tubes.

DNA fragments were then extracted from the gel slice using QIAquick extraction kit (QIAGEN, Japan) by following the manufacturer's protocol. This involved melting

the gel slice in 3 volumes of buffer QG at 50 °C for 10 minutes, mixing with 1 volume of isopropanol, binding DNA fragments by QIAquick column, and washing the column by 0.5 ml buffer QG and 0.75 ml buffer PE, respectively. Purified DNA fragments were eluted by adding 50 µl EB buffer (10 mM Tris-HCl, pH 8.5) to the columns placed on Eppendorf tubes, and centrifuging at 13,000 rpm for 1 minute at room temperature.

The DNA concentration in solution was quantified by its optical density (O.D.) at 260 nm using a Shimadzu UV-1601 spectrophotometer (Shimadzu corporation, Japan). The purity of purified DNA samples was determined by the ratio between absorbance at 260 nm and 280 nm ($O.D._{260}$/ $O.D._{280}$), which should be around 1.8. Ratios significantly deviated from this value indicate contamination of protein or phenol. An O.D. of 1 at 260 nm corresponds to a concentration of 50 µg/ml for DNA [10]. Thus, the DNA concentration was determined by following Eq. 2.22:

$$\text{DNA Concentration } (\mu g/\mu l) = O.D._{260} \times \frac{\text{Dilution Factor} \times 50}{1000} \qquad \text{Eq. 2.22}$$

2.3.2 Polymerase chain reaction (PCR)

GC-rich PCR

PCR has been used in this thesis to amplify specific DNA regions of interest. Because of the high GC content in the proximal promoter of human *TnIc*, GC-rich PCR (Invitrogen) strategy was used. Generally, a 50 µl reaction system was set up in a 0.2 ml thin-walled PCR tube as described below:

10 × PCRx Amplification buffer	5 µl
50 mM $MgSO_4$	2 µl
dNTP (10 mM each)	2.5 µl
Primers (100 µM, sense and anti sense)	1.25 µl each
Template DNA (10 µg/µl)	5 µl
Platinum Taq Polymerase (5 U/µl)	0.2 µl
Sterile MilliQ H_2O	up to 50 µl

The optimum concentrations of PCRx Amplification buffer (3 ×) and MgSO4 (2 mM) were determined by prior experiments. 0.5 × PCRx Enhancer solution was also used in certain cases. The PCR system was mixed thoroughly by vortexing, and then amplified by a PCR thermocycler (Stratagene). The PCR amplification was carried out by following the protocol as stated below.

Denaturation at 95 °C for 3 minutes

35 cycles of:

Denaturation at 95 °C for 1 minute

Annealing at 55 to 65 °C for 2 minutes (annealing temperature was optimised according to the melting temperatures of different primer pairs and results of temperature gradient PCR)

Elongation at 68 °C for 1 to 1.5 minutes (1 minute for 1 kb DNA)

Final elongation at 68 °C for 10 minutes

Hold at 4 °C

PCR products were then examined by 1.5 % agarose gel and purified according to the protocol stated above.

Colony PCR

Colony PCR was used to directly amplify the DNA fragments of interest from bacterial colonies, and to identify the sizes of inserted DNA fragments. Generally, 20 µl reaction system was set up proportionally as described above. A 200 µl sterile pipette tip was used to pick a bacterial colony, twirled in the PCR mix solution for 2 or 3 times to introduce the template DNA (bacterial plasmids), and then placed into a 50 ml Falcon tube containing 5 ml LB broth with corresponding antibiotics in order to regenerate the colony. PCR reactions were carried out as described previously, and PCR products were analyzed by agarose gel electrophoresis. Falcon tubes containing colonies with positive inserts were then incubated overnight at 37 °C with shaking at 225 rpm. DNA plasmids were then extracted and sent for sequencing to further examine the sequences of inserted DNA fragments.

2.3.3 Restriction enzyme (RE) digestion

In this thesis, double digests were used in the plasmid sub-cloning step. In the design of a suitable double digest system, enzyme pairs with 100 % activity in the same buffer were chosen. Thus, for all enzyme pairs in the double digest system mentioned in this thesis, the optimum reaction activity could be achieved in the same buffer. Standard double RE digests were performed in a final volume of 20 µl in the presence of 2 µg substrate DNA. The reaction mix was prepared by adding following components in order as stated into a 1.5 ml sterile Eppendorf tube.

Sterile MilliQ H_2O	13.8 µl
Restriction enzyme buffer X (10 ×)	2 µl
BSA (bovine serum albumin, 100 ×)	0.2 µl
Restriction enzyme 1 (20 U/µl)	1 µl
Restriction enzyme 2 (20 U/µl)	1 µl
Substrate DNA (1 µg/µl)	2 µl

The components were mixed gently by finger flicking and then incubated at 37 °C for 2 hours. Digested DNA products were checked and separated by agarose gel electrophoresis, and then extracted from the gel slice using QIAquick extraction kit by following the protocol as described previously. Larger or smaller restriction enzyme digestion systems were prepared by scaling up/down this basic reaction system proportionally.

2.3.4 Plasmid construction via Gateway® recombination cloning

Most of the plasmid constructs used in this thesis were generated via Gateway® recombination cloning technology (Invitrogen), by following the manufacturer's instruction. This technology derives from the bacteriophage lambda site-specific recombination system, which integrates bacteriophage lambda into the chromosome of *E. coli*, and switches between the lytic and lysogenic pathways. The recombination reaction of this bacteriophage lambda-derived system is mainly determined by two components: (i) the site-specific *att* (attachment) recombination sites that provide recognition sites for orientation-specific recombination and (ii) a group of the phage lambda and *E. coli* proteins involved in regulating the recombination reaction.

By following this site/orientation-specific recombination reaction in bacteriophage lambda, Gateway® Technology facilitates the transfer of heterologous DNA sequences of interest flanked by specific *att* sites between different vectors via recombination reactions. Gateway® Technology is composed of two recombination reactions, the BP reaction and LR reaction (Figure 2.3). Firstly, the DNA fragment (e.g. PCR products or *attB* expression clone) flanked with *attB* sites is recombined into a donor vector containing *attP* sites to give an "entry" clone containing *attL* sites flanked around the DNA fragment. This BP reaction is catalysed by BP Clonase™ enzyme mix that is involved in inducing lambda lysogenic pathway and generates an entry clone. The principle behind Gateway cloning is that the entry clone then allows the researcher to introduce the DNA into multiple specific "destination" vectors (for example, allowing expression in mammalian, yeast or insect cells) via a single further reaction, the "LR" step. Next, the DNA fragment between *attL* sites in the entry clone is further recombined into the destination vector, which contains *attR* sites, giving rise to an expression clone containing the DNA fragment of interest flanked with regenerated *attB* sites. This LR reaction is catalysed by LR Clonase™ enzyme mix that facilitates lambda lytic pathway.

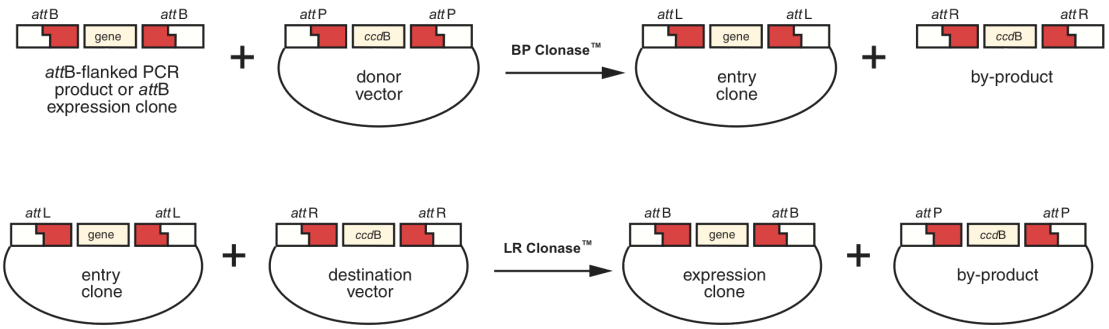


Figure 2.3: Schematic illustration of Gateway® technology.
The Gateway® technology is composed of two reactions, the BP reaction (upper) and LR reaction (lower). In the BP reaction, the DNA fragment of interest flanked with *attB* sites is recombined into a donor vector containing *attP* sites, and gives an entry clone containing the DNA fragment flanked with *attL* sites. LR reaction recombines the DNA fragment into a destination vector containing *attR* sites, and gives rise to an expression clone containing the DNA fragment flanked with *attB* sites. The BP and LR reactions are catalysed by BP Clonase™ enzyme mix and LR Clonase™ enzyme mixes, respectively. This figure was taken from the Invitrogen user manual.

Because only very small amount of initial *attB* flanked DNA fragment is available (will be described later), the high delivery efficiency of Gateway® Technology at room temperature (> 99 % within 1 hour) makes it a very attractive tool. Another advantage of Gateway® Technology is that the shuttle of insert DNA fragment from one expression vector to another is relatively simple. Furthermore, as compared to cloning with more conventional plasmids, the Gateway® vector is very efficient in selecting positive entry and expression clones. Between the two *att* sites in the Gateway® donor and destination vectors, there is a *ccdB* gene ("death gene"), whose product interferes with *E. coli* gyrase, and which is excised upon successful recombination with the DNA product of interest. Thus, only recombinant *E. coli* carrying an entry or expression vector from a successful BP or LR reaction can survive on plates containing the appropriate antibiotic, in which the *ccdB* gene is replaced by the DNA fragment of interest.

Generation of *attB*-flanked MNS concatemers via three rounds of ligation

Wild type (WT) and mutated (M) minisatellite concatemers were constructed by using a three-round ligation process (Figure 2.4). To improve ligation efficiency, single-stranded oligos **MNS_G/C** and **MNSM_G/C** (Table 2.2) were firstly 5' phosphorylated in 10 µl phosphorylation reaction systems as described below.

Sterile MilliQ H_2O	6 µl
DNA oligo (100 µM)	2.5 µl
T4 DNA Ligase buffer (10 ×)	1 µl
PNK (T4 Polynucleotide Kinase 10 U/µl)	0.5 µl

Since the PNK reaction was followed by a ligation reaction, to simplify this process, PNK is optimised for use in T4 Ligase reaction buffer (which contains the appropriate amount of ATP, 1 mM as working concentration). PNK phosphorylation was performed at 37 °C for 1 hour. Then, 5' phosphorylated oligos **MNS_G/C** and **MNSM_G/C** (10 µl G-rich oligo was mixed with 10 µl C-rich oligo giving a 20 µl system) were annealed to each other by heating at 95 °C for 5 minutes, then gradually cooling to 25 °C. 20 µl of annealed **MNS_G/C** and **MNSM_G/C** were mixed with 2 µl T4 DNA Ligase (10 U/µl), and incubated at 16 °C for 2 hours to allow the generation of self-ligated concatemers with random number of MNS repeat. The

different sticky ends at 5'- (5'CCTGGA) and 3'-end (5'TCCAGG) of individual MNS repeat ensures that the self-ligated products are in the same direction. These first round ligation products (MNS concatemers) were purified by QIAGEN Gel extraction kit. DNA concentration was quantified by the absorbance at 260 nm.

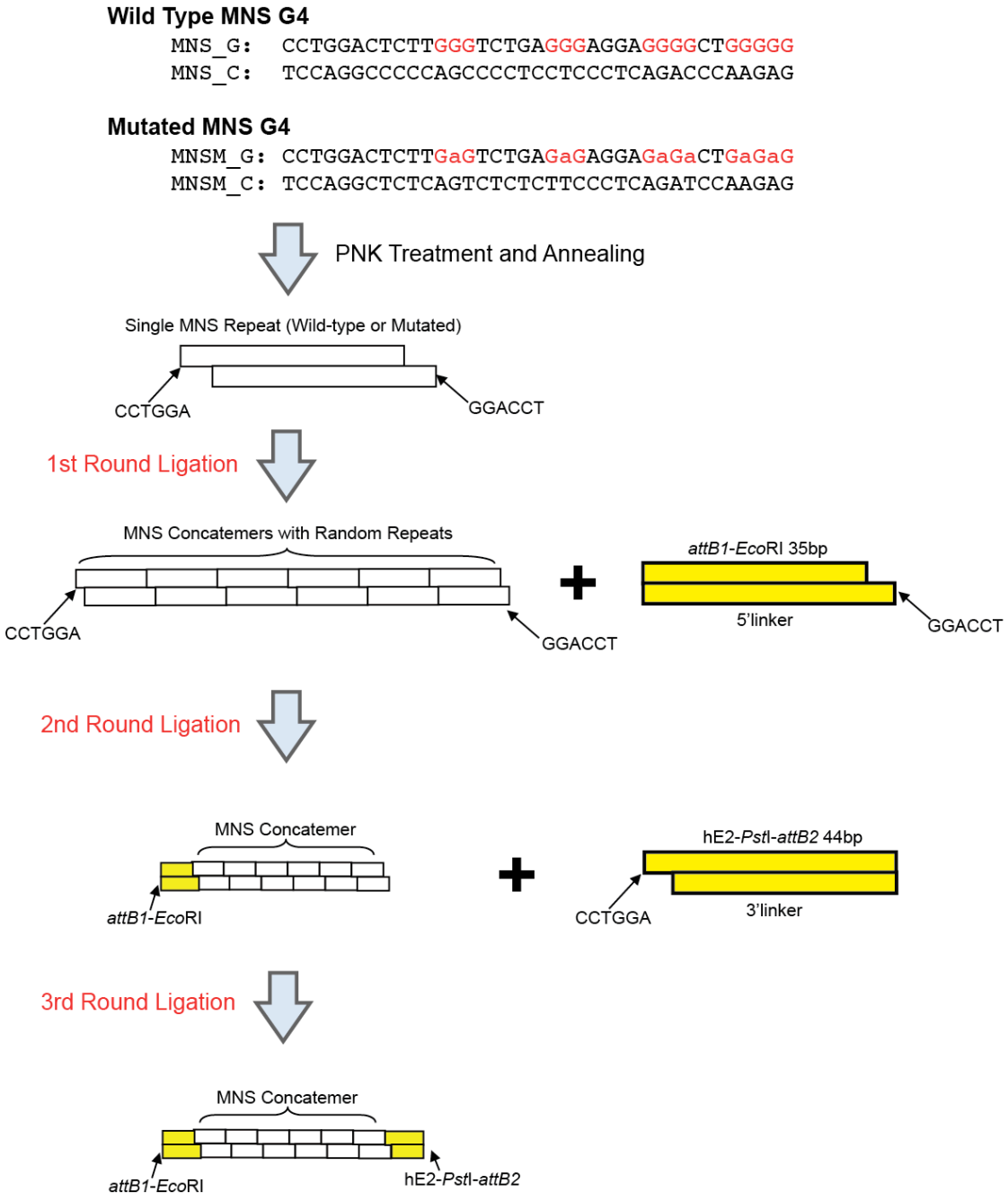


Figure 2.4: Schematic illustration of the protocol to generate *attB*-flanked human TnIc MNS concatemers via three rounds of ligation.

In the process of the first round ligation, oligos **5'Linker_G** and **5'Linker_C** (Table 2.2) were phosphorylated and annealed by following the same protocol as described above. Then, annealed 5'linker (0.75 pmol) was ligated to the purified MNS concatemers (0.25 pmol) in the second round ligation by following a similar protocol as mentioned previously. The 5'linker contains an *attB1* recombination recognition site and an *Eco*RI RE site. The second round ligation products, *attB1-Eco*R1-nMNS fragments, were also purified by QIAGEN Gel extraction kit. In the third round ligation, oligos **3'Linker_G** and **3'Linker_C** (Table 2.2) were phosphorylated, annealed, and ligated to the 3'-end of products purified from the second round ligation by following the same protocol.

Three rounds of ligation give a group of DNA fragments that are composed of an *attB1* site and an *Eco*RI RE site in their 5'-ends, a MNS concatemer with random repeats in the middle, and an E-box (hE2), a *Pst*I RE site, and an *attB2* site in their 3'-end. These *attB*-flanked MNS concatemers were examined by a 12 % native polyacrylamide gel prior to be recombined into a donor vector via BP reaction (Figure 2.5).

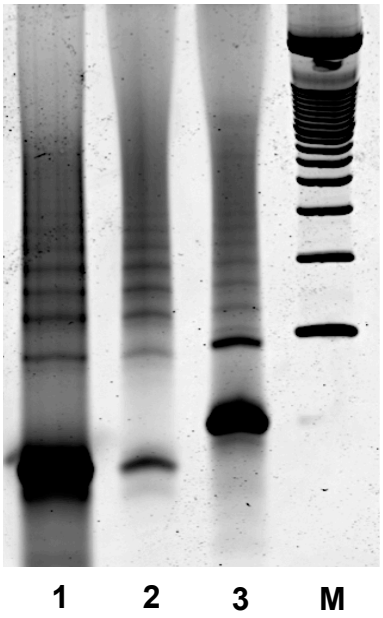


Figure 2.5: Ligation products of wild type MNS in a 12 % native polyacrylamide gel.
1 The first round ligation products of wild type MNS concatemers with random number of repeat. **2** The second round ligation products of wild type MNS concatemers with a linker containing *attB1* and an *Eco*RI site in the 5'-end. **3** The third round ligation products of wild-type MNS concatemers with a linker containing an E-box (hE2), a *Pst*I site, and *attB2* in the 3'-end. **M,** 100 bp ds DNA marker (Invitrogen).

BP recombination reaction

The purified MNS concatemers (both wild type and mutated) flanked with *attB* sites were recombined into the donor vector pDONR^{TM} 221 containing *attP* sites to give an entry clone with MNS concatemers. The BP reaction was set up by adding the following components into a 1.5 ml sterile Eppendorf tube in order as stated.

TE buffer (pH 8.0)	3.5 µl
BP clonase buffer (5 ×)	2 µl
Purified MNS concatemers (100 ng/µl)	1.5 µl
pDONR^{TM} 221 (100 nM)	1 µl

After mixing thoroughly, 2 µl of BP^{TM} Clonase enzyme mix were added, and mixed by vortexing twice for 2 seconds. The reaction was performed by incubating at 25 °C for 3 hours. BP reaction was terminated by adding 1 µl of proteinase K solution, and incubated at 37 °C. A positive and a negative control were also set up by replacing *attB*-flanked MNS concatemers with pEXP7-tet DNA and MilliQ H_2O respectively.

Next, 1 µl of BP reaction product was used to transform 50 µl *E. coli* competent cells (library efficiency® DH5α^{TM} cells, Invitrogen) as described in Section 2.3.6. Transformed cells were selected on an agar plate containing 50 µg/ml kanamycin. After incubating at 37 °C overnight, 10 isolated colonies were screened for the presence of MNS concatemers by colony PCR as described in Section 2.3.2. The M13 forward/reverse primer pair was used in colony PCR (Table 2.2). Colonies containing inserts with required length were then regenerated in LB broth overnight. Plasmid DNA was prepared from these colonies and sent for sequencing as described in Section 2.3.5 and 2.3.7. Based on the sequencing results, entry clones with specific number of MNS repeat proceeded to further LR recombination reaction in the following steps. The BP reaction is illustrated in Figure 2.6.

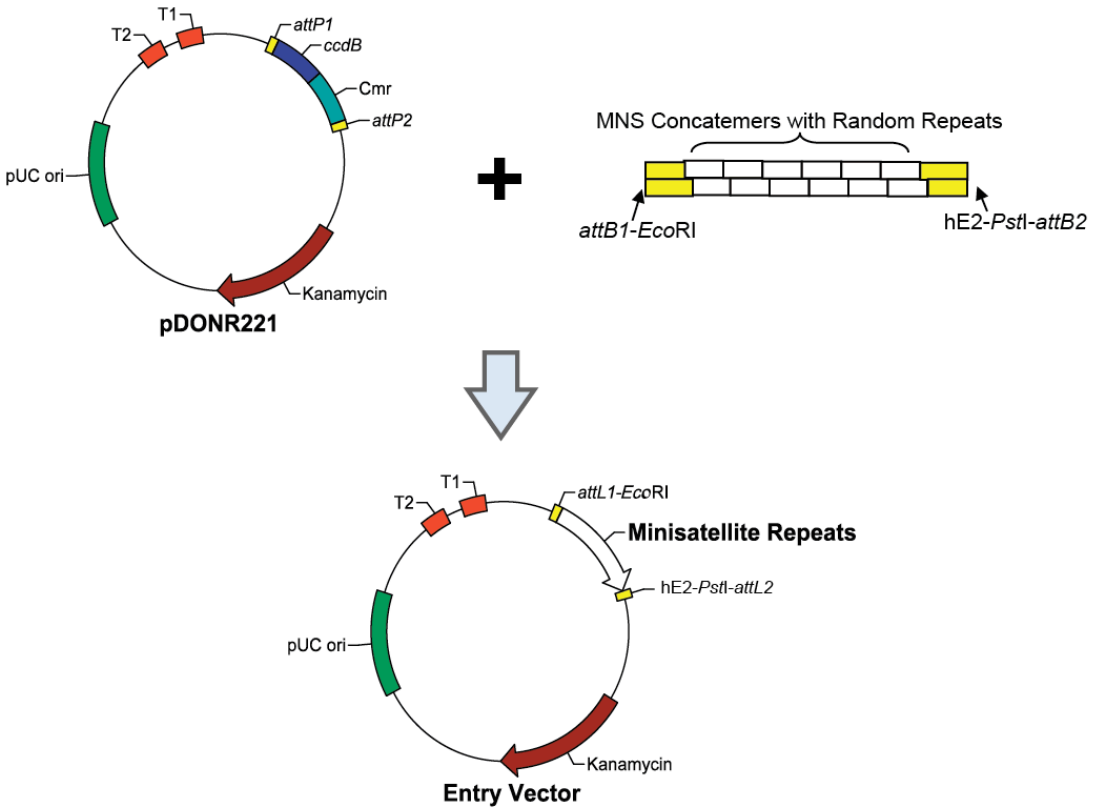


Figure 2.6: Schematic illustration of BP recombination reaction to recombine the MNS concatemers into pDONR221 donor vector in order to create entry vectors.

LR recombination reaction

MNS concatemers were then recombined from the *attL*-containing entry clone into an *attR*-containing destination vector (pGL3-Basic-GW) to create an *attB* containing expression clone (pGL3-Basic-GW-MNS). Destination vector pGL3-Basic-GW, was constructed by Dr. Malathi Raman [1], and kindly provided by Dr. Nigel Brand. Generally, the *attR* Reading Frame Cassette B (RfB) containing *ccd*B and chloramphenicol resistance genes was inserted to the multiple cloning site (MCS) of a promoterless firefly luciferase reporter vector pGL3-Basic (Promega) via *Sma*I blunt end ligation. The LR recombination reaction was set up by adding the following components into a 1.5 ml sterile Eppendorf tube in order as described.

TE buffer (pH 8.0)	3.5 µl
LR clonase buffer (5 ×)	2 µl
pGL3-Basic-GW (200 ng/µl)	1.5 µl

Entry clone with specific MNS concatemer (200 ng/µl) 1 µl

After mixing thoroughly, 2 µl of LR^{TM} Clonase enzyme mix were added, and mixed by vortexing twice for 2 seconds. The reaction was performed by incubating at 25 °C for 3 hours. LR reaction was terminated by adding 1 µl of proteinase K solution, and incubated at 37 °C. A positive and a negative control were also set up by replacing entry clones with pENTR^{TM}-gus DNA and MilliQ H_2O respectively. Transformants were screened by colony PCR and the sequence of MNS inserted concatemers were determined by DNA sequencing as described above. For both colony PCR and sequencing, the RV3 and GL2 primer pair was used (Table 2.2). The LR reaction is illustrated in Figure 2.7.

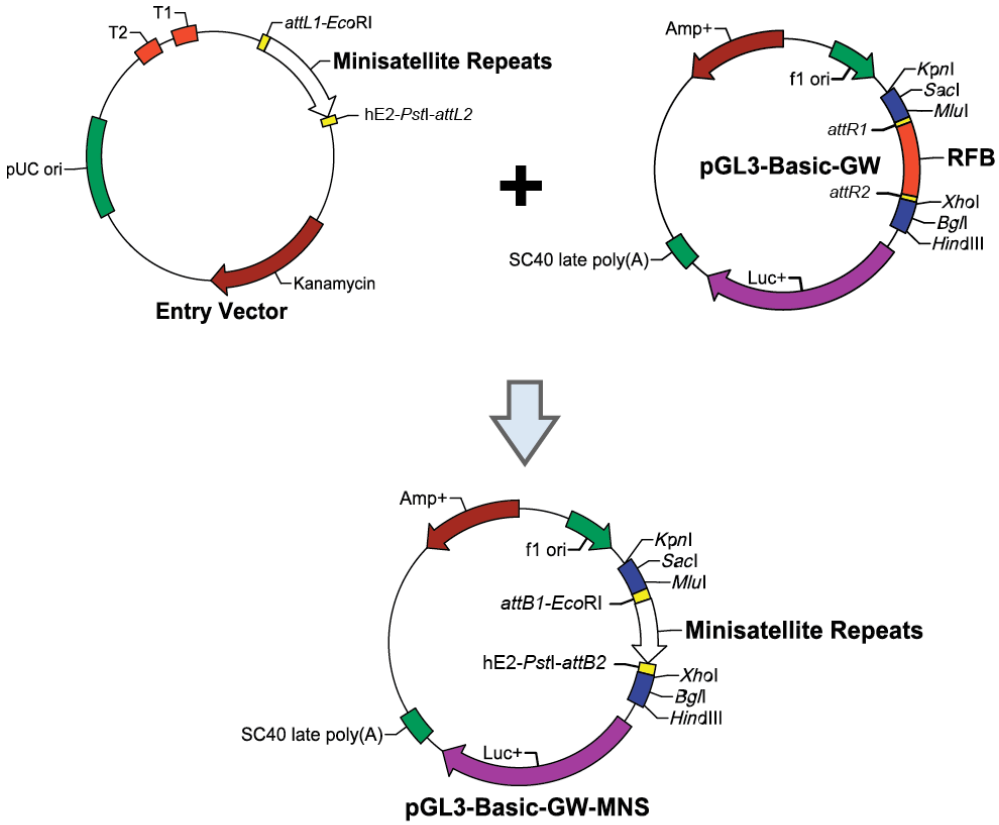


Figure 2.7: Schematic illustration of LR recombination reaction to recombine the MNS concatemers into pGL3-Basic-GW destination vector to give expression vector pGL3-Basic-GW-MNS.

Insert of the -299 ~ +51 bp of *TnIc* into the pGL3-Basic-GW-MNS vector

To construct a luciferase expression vector containing the complete human *TnIc* promoter, the fragment between -299 and +51 bp of the human *TnIc* gene was amplified by PCR and inserted into the pGL3-Basic-GW-MNS (Figure 2.8 a). By using the 5'-PCR-*Pst*I and 3'-PCR-*Xho*I primer pair (Table 2.2), the DNA fragment -299/+51 bp of human *TnIc* was amplified from the template plasmid hTnIc-531WT-*luc* (constructed by Dr. Malathi Raman [1]) containing -531/+67 bp of wild type human *TnIc* promoter sequence. The PCR protocol has been described in Section 2.3.2. The PCR products and pGL3-Basic-GW-MNS vector were separately double-digested by *Xho*I and *Pst*I by following the protocol as described in Section 2.3.3 (1 × NEBuffer 3 was used**)**. DNA fragments were separated by agarose gel electrophoresis and purified by QIAquick extraction kit. Then the digested PCR product and pGL3-Basic-GW-MNS were ligated to each other by adding the following components into a 1.5 ml sterile Eppendorf tube in order as described.

Vector DNA (double digested pGL3-Basic-GW-MNS)	0.03 pmol
Insert DNA (double digested PCR product)	0.3 pmol
T4 DNA Ligase buffer (10 ×)	2 µl
T4 DNA Ligase (10 U/µl)	1 µl
Sterile MilliQ H_2O	up to 20 µl

The mass of DNA added to the ligation reaction system was calculated by following equation.

$$\text{DNA mass (ng)} = \frac{\text{X pmol} \times \text{Y bp} \times 660}{1 \times 10^3} \qquad \text{Eq. 2.23}$$

where X is the amount of DNA in pmol unit, and Y is the length of DNA fragment in bp. The ligation was carried out at room temperature for 1.5 hours. 3 µl of the ligation product was used to transform 50 µl *E. coli* competent cells (library efficiency® DH5α™ cells) as described in Section 2.3.6. Transformed cells were selected on an agar plate containing 50 µg/ml ampicillin. Positive clones were examined by colony PCR and sent for sequencing. By following this protocol, luciferase expression vectors driven by the human *TnIc* promoter containing -299/+51 bp and variable repeats of MNS (engineered wild type and mutated) were generated (Figure 2.8 b).

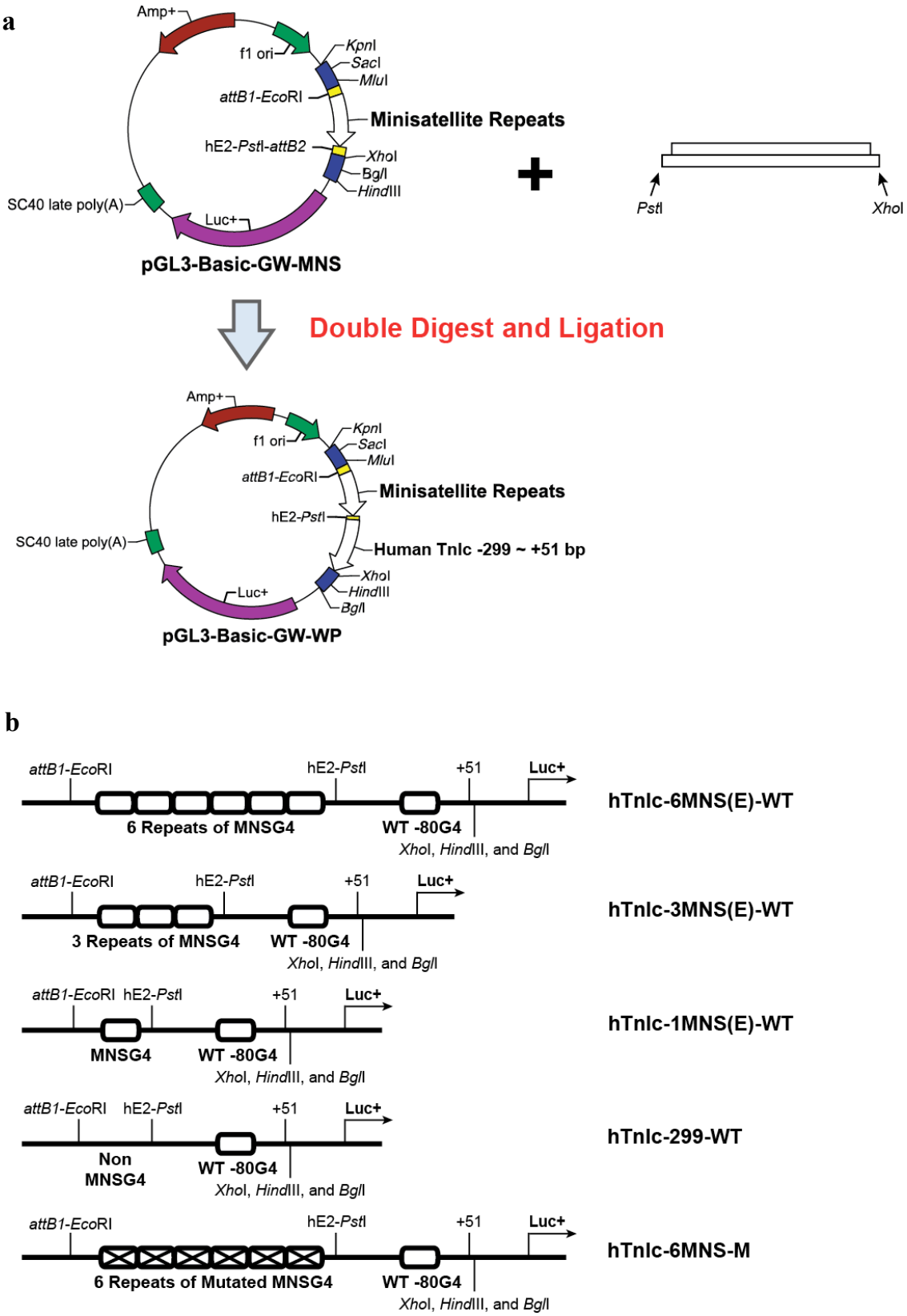


Figure 2.8: Construction of luciferase expression vectors driven by human TnIc MNSG4 and -80G4.

(a) The DNA fragment of -299/+51 of the human *TnIc* promoter region was inserted into the pGL3-Basic-GW-MNS vector via *Pst*I and *Xho*I RE sites. **(b)** Structure of different luciferase expression vectors containing variable repeats of MNSG4 and mutated MNSG4.

Mutagenesis of the -80G4 in luciferase expression vectors

In this thesis, *Dpn*I-mediated site-directed mutagenesis was used to introduce site-specific mutations into the -80G4-forming motif in the luciferase expression vectors. As illustrated in Figure 2.9, the mutagenic primers containing the desired mutations were firstly annealed to the initial plasmid, and extended around the entire plasmid via PCR cycles. Since the product of the PCR reaction is never used as a template in subsequent reactions, the mutagenesis PCR reaction follows a linear amplification rather than an exponential one. Following the reaction, the PCR product is digested with *Dpn*I. This endonuclease only cleaves at methylated sites, which exist only on the template plasmids generated by *E. coli*, but not on the PCR product. Thus, all parental plasmids without mutations are degraded. The PCR reaction of site-directed mutagenesis was set up by adding following components in order as stated.

10 × PCRx Amplification buffer	5 µl
50 mM $MgSO_4$	1.2 µl
dNTP (10 mM each)	4 µl
Primers (100 µM, sense and anti sense)	2.5 µl each
Template DNA (10 ng/µl)	5 µl
Platinum *Pfx*® DNA Polymerase (2.5 U/µl)	1 µl
Sterile MilliQ H_2O	up to 50 µl

The concentration of primer versus template determines whether primer-primer or template-primer interaction dominates. When a reaction fails, modifications such as reducing the primer concentration or increasing the template concentration were necessary. In this thesis, primer pairs, -80G4-sp1M-F/R, -80G4-G4M-F/R, and -80G4-M-F/R (Table 2.2) were used to introduce mutations in Sp1 binding site, G4-forming sequences and both of them, respectively. The PCR amplification was carried out by following the protocol as stated below.

Denaturation at 95 °C for 3 minutes

15 cycles of:

Denaturation at 95 °C for 1 minute

Annealing at 55 to 65 °C for 2 minutes (annealing temperature was optimised according to the melting temperatures of different primer pairs and results of temperature gradient PCR)

Elongation at 72 °C for 6 minutes (1 minute for 1 kb DNA)

Final elongation at 72 °C for 15 minutes

Hold at 4 °C

Parental template plasmid PCR products were then digested by adding 1 µl *Dpn*I (10 U/µl) and incubated at 37 °C for 1.5 hours. Then, 2 µl of the 1:10 diluted digested products were used to transform 50 µl XL-1 Blue competent cells (Stratagene). Desired mutations were subsequently confirmed by sequencing.

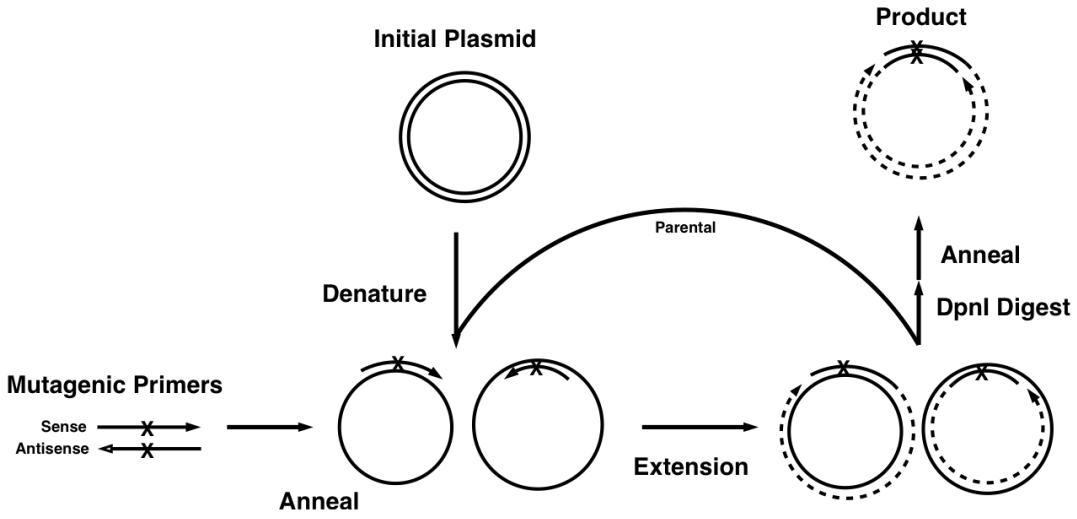


Figure 2.9: Schematic illustration of the *Dpn*I-mediated site-directed mutagenesis.
Primer pair containing desired mutations was first annealed to the parental plasmid, and extended through the whole vector via PCR reaction. Parental plasmids were subsequently degraded by *Dpn*I digestion.

2.3.5 DNA sequencing

Generally, 1.5 µg of mini-prep plasmid DNA (QIAGEN purified) was mixed with 15 pmol of each primer in a volume of 15 µl sterile MilliQ. Plasmid samples and primers were sent to MWG-Eurofin for sequencing in a 1.5 ml sterile Eppendorf tube. The length of sequence read (up to 1100 bases) was with an accuracy > 99 %. The sequencing data of plasmid samples were analysed using Geneious 4.8.5 software (Biomatters Ltd., New Zealand) to confirm the sequences of desired DNA fragments.

2.3.6 Bacterial transformation

In this thesis, DNA plasmids of interest were transformed into sub-cloning efficiency® DH5α (Invitrogen) or XL-1 blue competent (Stratagene) cells according to the manufacturer's instruction. Briefly, 50 µl competent cells were quickly thawed from -80 °C, and maintained on ice. 1 ~ 5 µl of plasmid solution was gently mixed with the thawed competent cells in a chilled 15 ml Falcon 2059 tube, and incubated on ice for 30 minutes. A positive control transformation reaction was carried out by mixing 50 µl DH5α competent cell with 50 pg uncut pUC19 DNA or by mixing XL-1 blue competent with 50 pg uncut pUC18 DNA. Cells were subsequently heat-shocked at 42 °C for 30 seconds in a water bath, and immediately incubated on ice for 2 minutes. Then, 950 µl pre-warmed (37 °C) LB broth without antibiotics was added to the transformation reaction, and the tube was incubated at 37 °C with shaking (~ 200 rpm) for 1 hour. The reaction was transferred to a 1.5 ml sterile Eppendorf tube and cells were harvested by centrifuging at 13,000 rpm for 1 minute. The cell pellet was resuspended in 100 µl LB and immediately spread on a pre-warmed LB agar plate containing appropriate antibiotics. The plate was incubated at 37 °C upside down overnight in order to allow colonies to grow. Colonies were subjected to colony PCR to examine the presence of inserts on the following day.

2.3.7 Small-scale plasmid purification

Small-scale plasmid DNA used in sub-cloning, and transformation was purified using QIAGEN QIAprep Spin Miniprep Kit according to the manufacturer's instruction. The QIAprep miniprep is based on alkaline lysis of bacterial cells followed by absorption of DNA onto silica in the presence of high salt.

Briefly, a single bacterial clone was inoculated into 10 ml of LB broth containing appropriate antibiotics and cultured at 37 °C overnight with shaking at 300 rpm. Cells were then harvested by centrifuging the 50 ml Falcon tube at 13,000 rpm for 5 minutes, and cell pellets were resuspended thoroughly in 250 µl Buffer P1 containing RNase A to degrade RNA. The suspension was then transferred to a 1.5 ml sterile Eppendorf tube, and mixed with 250 µl Buffer P2 (alkaline SDS lysis buffer) by inverting the tube 4 ~ 6 times. Buffer P2 is a NaOH/SDS mixture, in which the SDS is

used to solubilise phospholipids and protein components of the cell membrane leading to cell lysis, and NaOH is used to denature the released DNA and protein components. Then, 350 µl of Buffer N3 was mixed with the lysate immediately by inverting the tube 4 ~ 6 times. The addition of Buffer N3 is to neutralise and adjust salt concentration of the lysate, which precipitates the denatured chromosomal DNA, proteins and other cell components, while plasmid DNA stays in solution. Then, the lysate was centrifuged for 10 minutes at 13,000 rpm to separate plasmid DNA and other cell components. The resulting supernatant containing plasmid DNA was applied to a QIAprep spin column by decanting. By centrifuging the column for 1 minute at 13,000 rpm and discarding the flow-through, plasmid DNA was absorbed onto the silica-gel membrane in the column. 0.75 ml Buffer PE (wash buffer) was added to the column and centrifuged again for 1 minute at 13,000 rpm to remove the salts. The column was centrifuged again to remove residual wash buffer. Then, the column was transferred into a sterile 1.5 ml Eppendorf tube and 50 µl Buffer EB (elution buffer: 10 mM Tris-HCl, pH 8.5) was added to the column. After letting stand for 1 minute, plasmid DNA was eluted by centrifuging the column for 1 minute at 13,000 rpm. Concentration of the plasmid DNA was measured as described in Section 2.3.1. Samples were stored at -20 °C until use.

2.3.8 Large-scale plasmid purification

Due to the requirement of high quality plasmid DNA in mammalian cell transfection experiments, plasmid DNA of interest was purified by using QIAGEN Plasmid Maxi kit according to the manufacturer's instruction. The basic mechanism is similar as that used in small-scale plasmid purification described above. Briefly, a single colony containing the plasmid DNA of interest was inoculated into a starter culture (5 ml LB broth medium) containing appropriate antibiotics, and incubated at 37 °C with vigorous shaking (~ 300 rpm) until the O.D. reaches 0.6 ~ 0.8 at 600 nm (~ 8 hours). Then 0.5 ml of starter culture was added into 250 ml of selective LB broth medium, and cultured overnight at 37 °C with vigorous shaking (~ 300 rpm). Cells were harvested by centrifuging at 6,000 × *g* for 15 minutes at 4 °C, and cell pellets were resuspended in 10 ml of Buffer P1 with RNase A and transferred into a sterile 50 ml Falcon tube. Cells were lysed by mixing thoroughly with 10 ml of Buffer P2 by

gently inverting the tube 4 ~ 6 times, and incubated at room temperature for 5 minutes. Then, cell components other than plasmid DNA were precipitated by adding 10 ml of chilled Buffer P3, and incubating on ice for 20 minutes. After centrifuging twice at 13,000 rpm at 4 °C, the clarified supernatant containing plasmid DNA was applied to a QIAGEN-tip 500 pre-equilibrated with 10 ml Buffer QBT. When the supernatant flow through the anion-exchange resin of the column, the negatively charged DNA was bounded by the positively charged DEAE (diethylaminoethanol) group in the resin. After allowing the supernatant to enter the column by gravity flow, the column was further washed twice by 30 ml of Buffer QC to remove residual contaminants. Then, the plasmid DNA was eluted by 15 ml of Buffer QF, and precipitated by adding 10.5 ml of isopropanol. The sample was mixed by vortexing and centrifuged at ≥ 15,000 ×*g* for 30 minutes at 4 °C. Supernatant was carefully removed without disturbing the precipitated DNA pellet. Next, the DNA pellet was washed with 5 ml of room-temperature 70 % ethanol to remove residual salts, and centrifuged at ≥ 15,000 × *g* for 10 minutes at 4 °C. The supernatant was carefully decanted and the DNA pellet was air-dried for 15 minutes. The DNA pellet was dissolved in 500 µl Buffer EB (10 mM Tris-HCl, pH 8.5), and DNA concentration was measured as described in Section 2.3.1. Samples were stored at -20 °C until use.

2.3.9 Initiation of the HEK293 cell line from frozen stock

HEK293 cells (Human Embryonic Kidney 293 cells) were kind gifts from Dr. Nigel Brand, and stored in liquid nitrogen as described in Section 2.3.11. Prior to use, a frozen aliquot of cells were thawed quickly at 37 °C. When fully thawed, cells were immediately transferred into a 75 cm^2 flask (Greiner Bio-one, UK) containing 15 ml of pre-warmed complete growth medium (Table 2.9). The cells were then incubated at 37 °C to grow as a monolayer and were maintained as described below (Section 2.3.10).

Table 2.9: Recipe to prepare complete growth medium for HEK293 cells.

Component	Volume
Dulbecco's Modified Eagle Medium (Low Glucose 1.0 g/L)	500 ml
Fetal Calf Serum (FCS)	41.5 ml
L-Glutamine (200 mM)	5.5 ml
10,000 Unit/ml Penicillin/streptomycin	5.5 ml

2.3.10 Maintenance and passage of the HEK293 cell line

HEK293 cells were maintained in 75 cm^2 flasks containing complete growth medium and grown to 70-80 % confluence at 37 °C and under 5 % CO_2 atmosphere in a humid LEEC incubator (LEEC, UK). To passage the cell line, the flask was gently shaken to detach loosely attached cells, and the medium was removed by aspiration. Cells were fully detached from the plastic by the addition of 3 ml pre-warmed trypsin-EDTA (0.05 mg/ml, Sigma, UK), and incubated at 37 °C with 5 % CO_2 for 3 minutes. Then, 7 ml of pre-warmed complete growth medium was added to terminate trypsin digestion. In addition, media was pipette along growth surface to lift cells, and to break up cell clumps. Because HEK293 cells are only semi-adherent, and can easily lift from the growth surface during assays that require multiple washes, no PBS wash was needed prior to trypsinizing. Then, cells were transferred to a sterile 50 ml Falcon tube and maintained in a 37 °C water bath. 1 ml of this cell suspension was then diluted ("1:20 split") in 20 ml pre-warmed complete growth medium.

2.3.11 Storing the HEK293 cell line in liquid nitrogen

The cell suspension was centrifuged for 5 minutes at 800 × *g* at room temperature to harvest cells. The supernatant was removed by aspiration, and the cell pellet was resuspended in an appropriate volume of ice-cold freezing medium (complete growth medium supplemented with 10 % DMSO, v/v) to give a cell concentration of 1×10^6 cell/ml. Then, 1 ml aliquots were dispensed into cryovial tubes (Nalge Nunc International, USA) and these were wrapped in Kleenex tissues and placed at -80 °C for 24 hours. Next day, vials were transferred into liquid nitrogen for long-term

storage. This two-step freezing method together with quick thawing method would efficiently reduce ice nucleation in the freeze-thaw cycle that damages the cells.

2.3.12 Calcium phosphate-mediated transient transfection of HEK293 cell line

Calcium phosphate-mediated transfection has been in use since the mid 1970s [11]. Since then, this simple technique has changed little, and continued to be one of the most frequently used techniques to transfer plasmid DNA into mammalian cells *in vitro*. Although the exact mechanism of calcium phosphate-mediated transfection is still not fully understood, it is believed that the calcium, which is poorly soluble in the presence of phosphate ions, forms fine microprecipitates with recombinant plasmid DNA (co-precipitation). These microprecipitates are believed to have a positive effect on transfection efficiency [12]. When exposed on the surface of host cell membranes, the microprecipitates are taken up by the cells by phagocytosis, allowing transient or stable transfection. As compared to other commonly used transfection methods, such as Lipofectamine^{TM} 2000-mediated (Invitrogen) transfection and FuGENE$^{®}$ 6-mediated (Promega) transfection, the calcium phosphate method is simple, inexpensive, and efficient in HEK293 cell line transfection with minimal cytotoxicity [13]. Thus, calcium phosphate-mediated transfection was used in this study.

Generally, on the day before transfection, cells were first split, counted, and diluted in complete growth medium to a concentration of 2×10^4 cells/ml. Next, cells were reseeded in 24 well plates with a volume of 300 μl/well complete growth medium, and cultured overnight. Cell density was ~ 50 % confluent on the following day and ready for transfection. On the day of transfection, the plate was swirled gently to detach dead cells, old medium was aspirated off, and cells were re-fed with 300 μl pre-warmed complete growth medium at least 2 hours prior to transfection. Trasnfection was carried out as follows. Briefly, for each 24-well plate, 3.6 μg testing pGL3 luciferase plasmid and 480 ng pRL-TK control plasmid (*Renilla* luciferase reporter vector driven by herpes simplex virus thymidine kinase promoter $P_{HSV\text{-}TK}$) were diluted by TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) in a 15 ml round-bottomed tube to make up a volume of 315 μl. With slow vortexing, 45 μl of 2M $CaCl_2$ was slowly and dropwisely added to give a final volume of 360 μl. Then, 360 μl of 2 × HBS (pH 7.12, recipe described in Table 2.10) was slowly added with

vortexing. Then, solution was left to stand for 5 minutes to allow the formation of microprecipitates (longer time are unnecessary). Then, 30 µl of precipitate was added to each of triplicate wells, and mixed with original medium by moving the plate up and down, then left and right for 3 times, respectively. Swirling would result in accumulation of precipitates in the middle of the well, and give low transfection efficiency, hence was avoided. Plates were incubated in the presence of precipitates for 24 hours. One day (24 hours) after adding precipitates, plates were swirled briefly, and the medium was aspirated and replaced by pre-warmed fresh complete growth medium. Then, cells were re-incubated for 24 hours before harvest. 48 hours after the addition of precipitates, cells were harvested. Briefly, cells in each well were washed twice with 1 ml ice cold PBS, gently swirled, and aspirated. Plates were left on ice and moved to general lab for subsequent experiments.

Table 2.10: Recipe to prepare 2 × HBS buffer.

Component	Amount Added	Final Concentration
HEPES	11.92 g	50 mM
KCl	745.60 mg	10 mM
Dextrose	2.16 g	12 mM
NaCl	16.36 g	280 mM
Na_2PO_4	211.4 mg	1.5 mM
ddH_2O	Up to 1 L	

Note: pH value was adjusted to 7.12 by addition of 1 M HCl. Solution was filtered using a 0.22 µm filter.

2.3.13 G4-binding ligand treatment of HEK293 cell line

To investigate the regulatory roles of G4-forming motifs *in vitro*, dose response curves of G4-binding ligands were measured by dual luciferase reporter assays. Briefly, G4-binding ligands were stepwisely diluted in DMSO to different concentrations as 1,000 × stock solutions. In the process of calcium phosphate-mediated transfection, 0.3 µl G4-binding ligand stock solution was added to 300 µl pre-warmed fresh medium to replace the transfection medium containing precipitates in each well. Negative controls in the absence of DMSO and in presence of 0.1 % DMSO (v/v) were also measured. Cells were then incubated with G4-binding ligands for 24 hours before cell lysis for luciferase assays.

2.3.14 Dual luciferase reporter assays

HEK293 cell lysates were analysed for luciferase reporter gene expression with and without G4-binding ligands using the Dual Luciferase® Reporter Assay system (Promega) according to the manufacturer's instruction. In these assays, the luminescence of both firefly (*Photinus pyralis*) and *Renilla* (*Renilla reniformi* or sea pansy) luciferase activities were determined sequentially from a single sample. Bioluminescence of luciferase is proportional to the amount of enzymes in the cell lysates, and thus indirectly represents the transcriptional rate of the reporter gene. Due to their different evolutionary origins, the firefly and *Renilla* luciferases have different structures and substrate requirements, which makes it possible to selectively discriminate between their respective bioluminescent reactions (Figure 2.10). In this study, the transcription of firefly luciferase was driven by the engineered human *TnIc* promoter containing different G4-forming motifs, while the *Renilla* luciferase was driven by herpes simplex virus thymidine kinase promoter $P_{HSV\text{-}TK}$, which does not contain G4-forming motif. Thus, bioluminescence of *Renilla* firefly was used as an internal control to investigate the transcription activity differences related to the G4-forming motifs with or without G4-binding ligands.

Transfected cells were first washed twice by ice-cold PBS to reduce background signals from the medium, and cells were lysed by adding 100 μl of passive lysis buffer (Promega) per well. Cells were left on a tilting platform at room temperature to lyse, then scraped in buffer and the cell suspension was added to an Eppendorf tube. Tubes were spun at 13,000 × *g* to pellet debris and the clarified suspension was transferred to a fresh tube. 10 μl of cell lysate was then added to 50 μl Luciferase Assay Reagent II (LAR-II) to initiate the firefly luciferase enzymatic reaction. After 3 seconds delay, luciferase luminescent signals were collected for 10 seconds in a Glomax 20/20 luminometer (Promega). After quantifying the firefly luciferase luminescence, the reaction was quenched and the *Renilla* luciferase reaction was simultaneously initiated by adding 50 μl of Stop and Glo® reagent to the same sample. This allowed the *Renilla* luciferase luminescence signals to be collected over 10 seconds as an internal control. Firefly luciferase (driven by different engineered *TnIc* promoter) luminescence was normalized by dividing by the *Renilla* luciferase

(driven by $P_{HSV\text{-}TK}$ as an internal control) luminescence to minimize experimental variability caused by differences in cell viability and transfection efficiency. Results were presented as mean ± standard deviation (SD) from three independent triplicate experiments (from different batches of HEK293 cells). Statistical significance was determined by a two-tailed unpaired Student's *t*-test, in which $P \leq 0.05$ represent a significant difference.

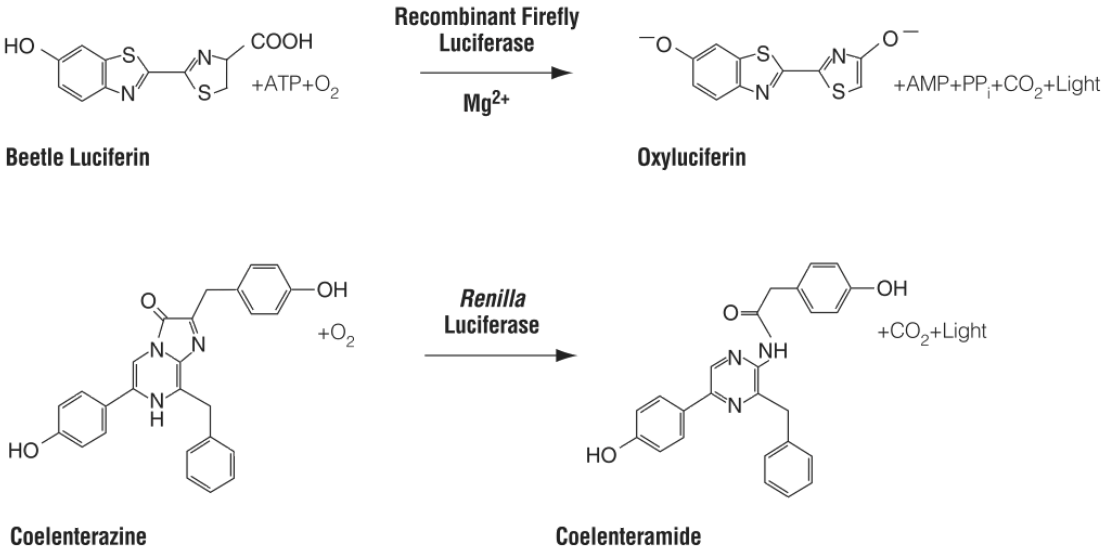


Figure 2.10: Schematic illustration of the mechanism of bioluminescent reactions catalyzed by firefly and *Renilla* luciferases.
Photon emission of firefly luciferase is achieved through oxidation of beetle luciferin in a reaction that requires ATP, Mg^{2+}, and O_2. The bioluminescence reaction catalyzed by *Renilla* luciferase utilizes O_2 and coelenterate-luciferin (coelenterazine), and also gives photon emission. Within the 1st second after the addition of Stop and Glo® reagent, the luminescence signal from firefly luciferase was quenched by a factor of 105, and the enzymatic reaction driven by *Renilla* luciferase is simultaneously activated. This figure was taken from the Dual Luciferase® Reporter Assay System manual (Promega).

2.4 Bioinformatics studies

2.4.1 Obtaining gene/transcript sequences and features

The human genome sequences

Gzip compressed sequences of the human genome (version 37.64) were downloaded from ENSEMBL FTP (ftp://ftp.ensembl.org/pub/release-65/fasta/homo_sapiens/dna/)

with file names as Homo_sapiens.GRCh37.64.dna.chromosome.*.fa.gz (* indicates chromosome number). Gzip compressed sequence files of the human genome were then decompressed to fasta format.

Overall gene/transcript features

Gene/transcript features including ENSEMBL gene ID, ENSEMBL transcript ID, chromosome number, strand polarity, gene start/end position, transcript start/end position, gene description, EntrezGene ID, transcript biotype and gene biotype were exported to a csv format file (comma-separated values) from BioMart (http://www.biomart.org/biomart/martview/). When exporting gene features, "Gene type" and "Status" of genes and transcripts were selected as "protein_coding" and "known", respectively.

Tissue-specific gene/transcript features

For tissue-specific G4 enrichment analysis, tissue-specific gene/transcript lists were exported in csv format by setting "GNF/Atlas organism part" to different tissues in the BioMart. Gene/transcript lists active in 60 different tissues (tissue-specific genes) and one list including all gene/transcript (at whole genome level) were exported, while 11 of them with transcripts number ≤ 800 (i.e. low abundance transcripts) were excluded in subsequent analysis due to limited gene/transcript available (salivary gland, pancreas, tonsil, beta cell islets, lymph, fetal lung, pituitary, thymus, uterus, ovary, and adrenal cortex).

Pathway-specific gene/transcript features

A csv format file containing component genes for all human pathways (234 pathways) was downloaded from KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY database (http://www.genome.jp/kegg/kegg2.html). In each pathway, KEGG ID of each gene was mapped to its ENSEMBL gene ID, and gene features were obtained according to the previously exported csv file from BioMart. All pathways including more than 50 genes were subjected to subsequent pathway analysis.

Transcriptional regulatory regions (TRRs) of individual transcripts were defined as -2,000 bp to +1,000 bp from TSS of the transcript. Coding strand sequences of TRRs

were extracted from the human genome sequence files, according to the transcript features (i.e. chromosome number, strand polarity, and transcript start site). G4-forming motifs on both strands were defined as $G_{\geq 3}(N_{1\sim 7}G_{\geq 3})_{\geq 3}$.

2.4.2 Redundancy issue

In the csv file showing gene/transcript features, most genes are composed of multiple transcripts, which would induce bias in subsequent analysis. Thus, in this study, to take into account this redundancy issue, only one transcript was randomly chosen to represent one gene, and the whole analysis was repeated for 10 times. Briefly, in tissue-specific and pathway-specific analysis, transcripts representing the same gene (with the same ENSEMBL gene ID) were grouped together, and one transcript was chosen according to a uniformly distributed pseudorandom number generated by MATLAB 2009 (MathWorks, Inc., USA) using a built-in function:

$$i = \text{unidrnd}(n,1) \qquad \text{Eq. 2.24}$$

in which i is the number of transcript picked, and n is the total number of transcripts representing one gene. In each iteration, a single transcript representing a single gene without redundancy was randomly chosen by following Eq. 2.24, and the result was presented as mean ± standard deviation (SD) from 10 iterations. Differences between iterations were evaluated by one-way ANOVA test, in which $P > 0.5$ represents no significant difference was found.

2.4.3 Enrichment of G4-forming motifs in different tissues

Tissue-specific enrichment of G4-forming motifs was defined as the percentage of genes active in certain tissue (GNF/Atlas organism part) containing at least one G4-forming motif in a certain region of the TRR (Eq. 2.25).

$$\text{Tissue-specific Enrichment of G4} = \frac{\text{No. of gene containing at least one G4}}{\text{Total No. of gene active in the tissue}} \times 100\% \qquad \text{Eq. 2.25}$$

Besides the full TRRs, enrichment of G4s was also analysed and compared in the distal promoter region (-2,000 ~ -501 bp), proximal promoter region (-500 ~ -1 bp), and downstream region (TSS ~ +1,000 bp). The analysis was carried out in 10 iterations to consider transcript redundancy. Hence the tissue-specific enrichment of G4-forming motifs was presented as mean ± standard deviation (SD). In analysing specific tissue, differences among iterations were evaluated by one-way ANOVA test, in which $P > 0.5$ represents no significant difference was found.

2.4.4 Importance of G4s in different pathways

To evaluate the importance of G4s in individual transcript, two scores, *F* and *Q*, were calculated respectively to reflect the abundance and location significance of G4s identified in each transcript. In the TRR of each transcript, the percentage of nucleotides involved in G4 formation is defined as *P* values according to Eq. 2.26.

$$F = \frac{\sum_{i=1}^{n}\left(ge_i - gs_i + 1\right)}{T} \times 100\% \qquad \text{Eq. 2.26}$$

where n is the number of G4s identified in the TRR ($F = 0$ when $n = 0$), gs_i and ge_i are the start and end positions of i^{th} G4 identified in the TRR relative to the TSS in 5' to 3' direction, and T, the length of TRR being searched, is fixed at 3,000 in this study.

Abundance of G4s on the coding strand, and template strand were separately calculated and labelled as F_{cd}, and F_{ncd}, (subscripts cd: coding strand; ncd: noncoding strand). Overall abundance of G4s, the P_o (o: overall), were calculated as the sum of F_{cd}, and F_{ncd}. To genome-widely compare the abundance of G4s in TRRs of different transcripts, *P* scores of all transcripts without redundancy were calculated (one transcript was randomly picked when multiple transcripts were reported from the same gene), and the cumulative frequencies of *F* scores of all transcript were calculated and denoted as CF_{cd}, CF_{ncd}, and CF_o. To reach a more stringent criterion, transcripts with *P* scores equal to 0 were excluded when calculating *CF*. Genes with transcript *CF* values higher than 50 % were regarded as G4-rich in their TRR, while those with transcript *CF* values lower than 50 % were regarded as G4-scarce.

Except for the G4 abundance, the location of individual G4 in a particular TRR may correlate with its biological significance. As reported in previous genome-wide analyses, G4 was found to be enriched in the promoter region between -200 bp and TSS, and peaked at around -50 bp to TSS in the human genome [14]. This positional bias of G4 distribution in human gene promoters is believed to be a result of evolutionary pressure [15]. Most promoter G4s with confirmed biological functions localize in this region, such as *KIT* and *MYC* promoter G4s. Thus, another indicator, the *Q* score, is introduced to evaluate the potential location significance of G4s in particular TRR.

Following on the previously reported method [14], the probability distribution of each TRR position relative to TSS involved in G4 formation along coding and template (noncoding) strands of all TRRs was calculated according to Eq. 2.27.

$$W_k = \frac{\sum_{j=1}^{m} \alpha_{j,k}}{m} \times 100\% \qquad \text{Eq. 2.27}$$

where W_k is the normalized probability of nucleotide in position k of all TRRs involved in G4 formation, m is the total number of transcripts analyzed. $\alpha_{j,k}$ is the logical value of nucleotide in position k of j^{th} TRR; $a_{j,k} = 1$ when the nucleotide is involved in G4 formation, otherwise $a_{j,k} = 0$. G4 location significance score Q (Q_{cd}, Q_{ncd} and Q_{o} for coding and template strand, and overall, respectively; $Q_{\text{o}} = Q_{\text{cd}} + Q_{\text{ncd}}$) for the TRR of a particular transcript was calculated according to Eq. 2.28,

$$Q = \sum_{i=1}^{n} \frac{\sum_{k=gs}^{ge} W_k}{ge_i - gs_i + 1} \qquad \text{Eq. 2.28}$$

The cumulative frequencies of Q scores, *CQ,* were also calculated. Similarly, transcripts with the zero Q scores were excluded. Genes with both *CF* and *CQ* values higher than 50 % were regarded as G4-important genes, while those with both values lower than 50 % were regarded as G4-less-important.

Correlation between the *CF* and *CQ* scores on the coding, template strand and both strands were also investigated. The correlation coefficient was calculated according to Eq. 2.29.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{Eq. 2.29}$$

in which x_i and y_i are the *CF* and *CQ* scores of the coding, template strand or both strands at genome level, $\bar{x}$ and $\bar{y}$ are the corresponding mean values *CF* and *CQ* scores at genome level.

CF and *CQ* were calculated for all transcripts without redundancy. To evaluate bias introduced in random-picking when multiple transcripts exist in one gene, calculation was carried out for 10 iterations. Gene list for all pathways downloaded from KEGG database. All pathways including more than 50 genes were subject to pathway analysis. When more than 50 % component genes in a pathway were calculated as G4-important (excluding those with both *CF* and *CQ* equal to zero), the pathway was regarded as G4-important. To evaluate the G4-importance of individual pathway, distribution of *CF* and *CQ* values of pathways were compared to the distribution of all transcripts in humane genome by Wilcoxon rank sum test. Differences between iterations of individual pathway were evaluated by one-way ANOVA test.

2.5 References

1. Malathi, R. (2008) Identification and characterisation of transcription factors binding the human cardiac troponin I gene In *National Heart and Lung Institute*, Imperial College London, London.
2. Suntharalingam, K., White, A. J., and Vilar, R. (2010) Two metals are better than one: investigations on the interactions between dinuclear metal complexes and quadruplex DNA, *Inorg Chem 49*, 8371-8380.
3. Sun, X. G., Cao, E. H., He, Y. J., and Qin, J. F. (1999) Spectroscopic comparison of different DNA structures formed by oligonucleotides, *J Biomol Struct Dyn 16*, 863-872.
4. Fasman, G. D. (1975) *Handbook of Biochemistry and Molecular Biology*, Vol. 1: Nucleic Acids, 3rd ed., CRC Press.
5. Cantor, C. R., Warshaw, M. M., and Shapiro, H. (1970) Oligonucleotide interactions. 3. Circular dichroism studies of the conformation of deoxyoligonucleotides, *Biopolymers 9*, 1059-1077.
6. Gill, R. T., Katsoulakis, E., Schmitt, W., Taroncher-Oldenburg, G., Misra, J., and Stephanopoulos, G. (2002) Genome-wide dynamic transcriptional profiling of the light-to-dark transition in Synechocystis sp. strain PCC 6803, *J Bacteriol 184*, 3671-3681.
7. John, D. M., and Weeks, K. M. (2000) van't Hoff enthalpies without baselines, *Protein Sci 9*, 1416-1419.
8. Ying, L. M., Wallace, M. I., Balasubramanian, S., and Klenerman, D. (2000) Ratiometric analysis of single-molecule fluorescence resonance energy transfer using logical combinations of threshold criteria: A study of 12-mer DNA, *J Phys Chem B 104*, 5171-5178.
9. Wallace, M. I., Ying, L., Balasubramanian, S., and Klenerman, D. (2001) Non-Arrhenius kinetics for the loop closure of a DNA hairpin, *Proc Natl Acad Sci U S A 98*, 5584-5589.
10. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, New York.
11. Graham, F. L., and van der Eb, A. J. (1973) A new technique for the assay of infectivity of human adenovirus 5 DNA, *Virology 52*, 456-467.
12. Jordan, M., and Wurm, F. (2004) Transfection of adherent and suspended cells by calcium phosphate, *Methods 33*, 136-143.
13. Jordan, M., Kohne, C., and Wurm, F. M. (1998) Calcium-phosphate mediated DNA transfer into HEK-293 cells in suspension: control of physicochemical parameters allows transfection in stirred media. Transfection and protein expression in mammalian cells, *Cytotechnology 26*, 39-47.
14. Huppert, J. L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res 36*, 6260-6268.
15. Zhao, Y., Du, Z., and Li, N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals, *FEBS Lett 581*, 1951-1956.

Chapter 3 Bioinformatics Studies

3.1 Introduction

As described in the Introduction, a growing body of evidence indicates that G4s are involved in diverse biological processes, especially acting as *cis*-regulatory elements in the promoter regions [1, 2]. Early studies about the regulatory roles of promoter G4s have focused on a few specific gene loci. For example, the formation of G4 structures (mostly stabilised by certain G4-binding ligands) in promoters of the human insulin [3, 4], *MYC* [5, 6], *KRAS* [7, 8], and *PDGF-A* [9] genes were shown to have significant influences on downstream gene transcription activity. Subsequently, promoter G4s have been identified in important functional regions (promoters, NHE regions, enhancers, and so on) of many other cancer-related genes, and their biophysical properties in solution have been thoroughly characterised. Examples include the G4s found in promoters of *KIT* [10-14], *RET* [15], *VEGF* [16-19], *RB1* [20-22], *BCL-2* [23-25], and *HIF-1α* [26] genes. The biological relevance of these G4s has been reviewed in detail in Section 1.2. These isolated examples have led to the hypothesis that G4s may act as a group of common *cis*-regulatory elements via various mechanisms *in vitro* and *in vivo*. Indeed, in the past decade, accumulating bioinformatics studies have supported this hypothesis in several aspects.

Firstly, the prevalence and biased distribution of G4s are conserved in genomes of different organisms, which indicated that these non-canonical DNA structures are favoured by natural selection in the process of evolution. For example, by using different algorithms Todd *et al.* and Huppert *et al.* independently revealed the prevalence of G4s in human genomic DNA, and identified that the length and sequence composition of putative loops involved in G4s are not random [27, 28]. Subsequent studies in the human genome have established that G4s are particularly enriched in promoter regions [29, 30], in the 5'-ends of 5'-UTRs, within the first intron and within regions immediately downstream of the 3'-ends of genes [30, 31]. Such a biased distribution of G4s within promoter regions (with a major peak centred around -50 bp and a minor peak between +150 bp and +300 bp) has been also found in the genomes of chimpanzee, chicken, and other warm-blooded animals, but not in

cold-blooded animals [32-34]. Furthermore, G4s appear to be more enriched in evolutionarily-older genes as compared to younger genes in warm-blooded animals, which indicates that G4s are preferred by the evolutionary selection [34].

Secondly, potential G4-forming motifs closely correlate with functional regions in gene promoters, which further supports the hypothesis that G4s are involved in gene transcription regulation. Du *et al.* have found that clusters of potential G4-forming motifs tend to be co-localized with several types of gene regulatory elements in the human genome [35]. Two independently conducted analyses also revealed that G4s are highly co-localized with several G-rich TF binding sites, such as those of TFs Sp-1, AP-2 (Activator protein-2), MAZ (Myc-associated zinc finger protein), and VDR (vitamin D receptor), and KLFs [36, 37]. Interestingly, recent *in vitro* studies demonstrated that certain zinc-finger and bHLH TFs, which were originally thought to interact exclusively with canonical duplex DNA, are able to interact with G4s as well. For example, the zinc-finger TF Sp1 is able to bind both canonical duplex DNA and G4s in solution with comparable affinity [38] and the homodimeric bHLH TF MyoD interacts preferentially with several muscle-specific bimolecular G4s [39-41]. These data further supported the original hypothesis.

Furthermore, the existence of potential G4-forming motifs is also related with downstream gene expression levels, which indirectly proved the regulatory roles of G4s in gene transcription regulatory regions (TRRs). For example, genes with high expression levels according to the Gene Atlas data set tend to have high G4 abundance on coding strands downstream to the TSS [42]. In another study that combined *in silico* and experimental approaches, stabilisation of conserved G4s in the TRRs by high concentration of the G4-binding ligand TMPyP4 down-regulated corresponding gene expression levels in HeLa S3 and A549 cells [33].

Finally, correlation between prevalence of G4s and gene functions also indicates the potential roles of G4s in the human genome. For example, in an analysis of 16,654 genes in the human RefSeq database, Eddy et *al.* found that potential G4-forming motifs are over-represented in proto-oncogenes, but under-represented in tumor suppressor genes [43].

In summary, these findings suggest that G4-mediated gene transcription regulation might be utilized as a common mechanism in living cells. Thus, it would be

interesting to also analyse if promoter G4s are also involved in regulating gene expression in a tissue- and pathways-specific manner. This could provide useful insight to evaluate the pharmaceutical potential of G4-binding ligands in different tissues, as well as perhaps revealing particular G4-forming motifs as potential targets via the multi-targeting strategy in the treatment of certain diseases.

In this chapter, enrichment of G4s in the TRRs of genes active in different tissues (GNF/Atlas organism) was first analysed. In the analysis, G4 enrichment in different regions of the TRRs was also examined. Then, by taking into account the abundance and location significance of G4 in each TRR, importance of G4 was evaluated and compared in different pathways. Finally, two candidate genes with high G4-importance were chosen as models for subsequent biophysical characterisations, which will be described in the following chapter.

3.2 Enrichment of G4-forming sequences in different tissues

As mentioned previously, enrichment of G4-forming sequences has been thoroughly investigated at the genome level in different species [32-34, 44, 45], and compared between proto-oncogenes and tumor suppressor genes in human [43]. Although it has been reported that the enrichment of G4s is correlated with gene functions, to date there has been no report to investigate potential tissue-specific correlations. Thus, by using the gene atlas of human protein-encoding transcriptomes data [46], tissue-specific enrichment of G4-forming sequences was analysed.

The enrichment of G4s in the transcriptional regulatory regions (TRRs: -2,000 bp to +1,000 bp around Ensembl Transcript Start) of all known protein-coding genes was first analysed as a reference point. For genes with multiple transcripts, only one was chosen randomly. G4s were searched for based on the algorithm $G_{\geq 3}(N_{1\sim 7}G_{\geq 3})_{\geq 3}$. The percentage of genes containing at least one G4-forming sequence in the TRR was calculated to indicate the enrichment of G4. The calculation was repeated for five times (multiple transcripts of one gene were randomly chosen for five times by using Eq. 2.24) to take into account the redundancy effects of multiple transcripts. In total, of 19,633 genes that were analysed, 67.79 ± 0.01 % of them were found to contain at least one G4 (on both coding strand, G4, and template strand, C4), with 52.42 ± 0.02 % containing at least one G4 on the coding strand (G4), and 46.91 ± 0.02 % containing at least one G4 on the template strand (C4). Compared with the widely accepted results that around 40 % human gene promoters contain G4s [29], the much higher percentage (67.79 ± 0.01 %) identified here can be attributed to the much longer region around the TSS that was analysed in this study. Similar results have been previously reported in the chicken genome when regions of similar length were analysed [32]. Following similar methods, genes active in different tissues (different GNF/Atlas organism) were also analysed. GNF/Atlas organism data sets (tissue-specific transcription) with less then 800 transcripts were excluded due to the large errors among different iterations. Tissues excluded in calculation include salivary gland, pancreas, tonsil, beta cell islets, lymph, fetal lung, pituitary, thymus, uterus, ovary, and adrenal cortex. As shown in Figure 3.1 and Table 3.1, G4s are highly enriched in the whole TRR of genes active in the lung, whole brain, caudate nucleus, heart, and cerebellum peduncles. Since the regulatory roles of G4s are found to be

location dependent as mentioned in Chapter 1, enrichment of G4s in different regions

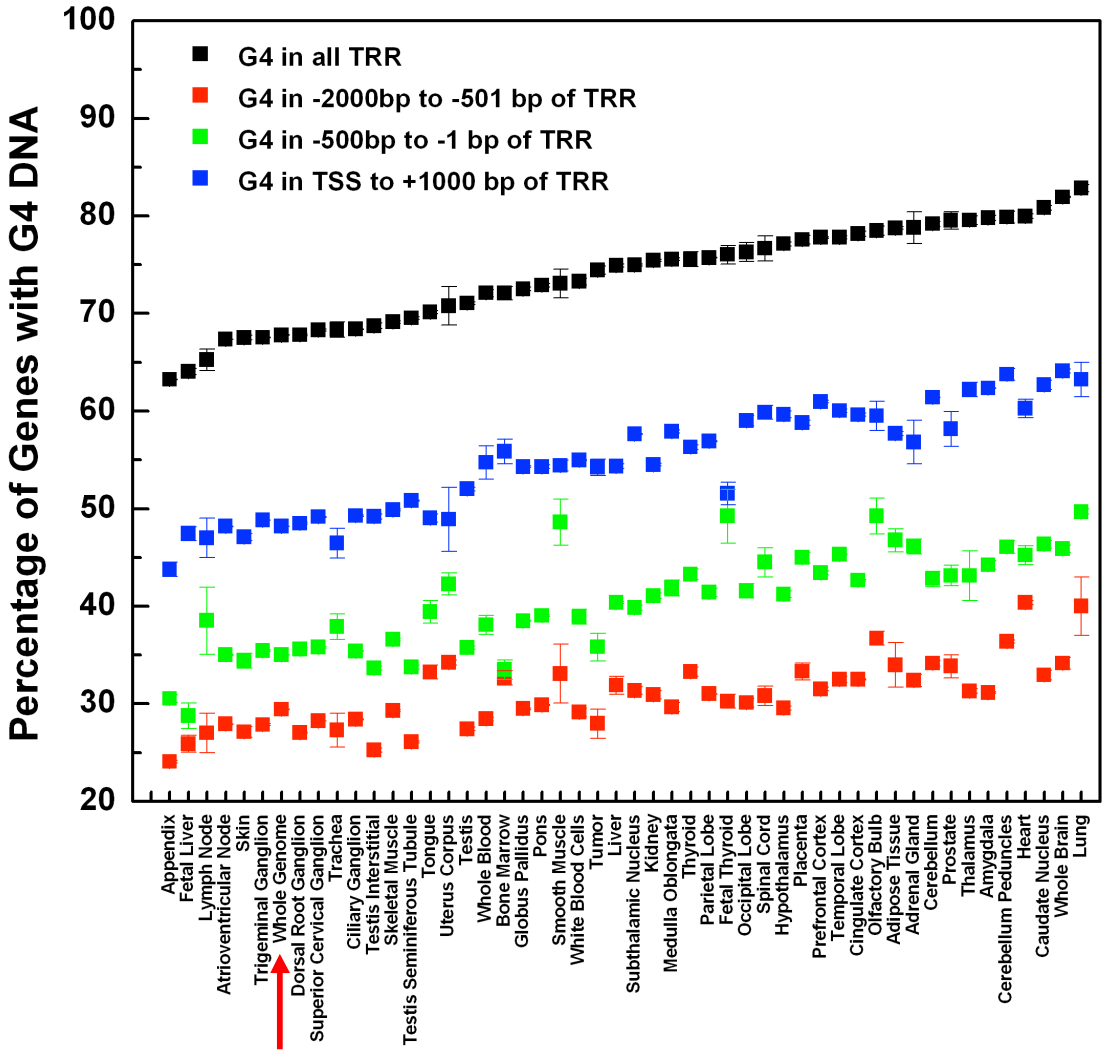


Figure 3.1: Enrichment of G4s in TRRs of genes active in different tissues.
The percentage of genes active in different tissues containing at least one G4-forming sequence in the whole TRR (black), -2,000 ~ -501 bp of TRR (red), -500 ~ -1 bp of TRR (green), and TSS ~ +1,000 bp of TRR. Enrichment of G4s in all protein-coding genes is indicated by the red arrow. The calculation was repeated randomly for five times (multiple transcripts of one gene were randomly chosen for five times) to take into account the redundancy effects of multiple transcripts.

of TRR was also analysed. For all genes found in the human genome (at whole genome level), the G4 enrichments in the full-length TRR, distant promoter (-2,000 ~ -501 bp), proximal promoter (-500 ~ -1 bp), and downstream region (TSS ~ +1,000 bp) were calculated to be 67.79 ± 0.01 %, 29.38 ± 0.03 %, 35.031 ± 0.004 %, and 48.23 ± 0.01 %, respectively.

Table 3.1: The rankings and scores of G4 enrichment in TRRs of genes active in different tissues.

Calculation was carried out by searching G4s in the full-length TRR (Whole TRR), distal promoter region (-2,000 ~ -501 bp), proximal promoter region (-500 ~ -1 bp), and downstream region (TSS ~ +1,000 bp). The calculation was repeated randomly for five times.

Tissues	The Ranking* (descending) and Score (%) of G4 Enrichment in Different Tissues			
	Whole TRR	-2,000 ~ -501 bp	-500 ~ -1 bp	TSS ~ +1,000 bp
Lung	1 82.8 ± 0.4	2 40.0 ± 3.0	1 49.7 ± 0.3	3 63.2 ± 1.8
Whole Brain	2 81.9 ± 0.1	6 34.1 ± 0.6	9 45.7 ± 0.4	1 64.1 ± 0.2
Caudate Nucleus	3 80.8 ± 0.2	14 32.9 ± 0.5	6 46.3 ± 0.4	4 62.7 ± 0.4
Heart	4 80.0 ± 0.2	1 40.4 ± 0.2	11 45.2 ± 1.0	9 60.2 ± 1.0
Cerebellum Peduncles	5 79.9 ± 0.6	4 36.4 ± 0.2	8 46.1 ± 0.6	2 63.8 ± 0.6
Amygdala	6 79.8 ± 0.3	23 31.1 ± 0.1	14 44.2 ± 0.4	5 62.31 ± 0.04
Thalamus	7 79.5 ± 0.4	22 31.3 ± 0.2	18 43.1 ± 2.5	6 62.2 ± 0.7
Prostate	8 79.4 ± 0.9	9 33.9 ± 1.2	17 43.1 ± 1.1	17 58.2 ± 1.8
Cerebellum	9 79.2 ± 0.3	7 34.1 ± 0.3	19 42.8 ± 0.8	7 61.42 ± 0.02
Adrenal Gland	10 78.8 ± 1.6	18 32.4 ± 0.6	7 46.1 ± 0.8	22 56.8 ± 2.2
Adipose Tissue	11 78.7 ± 0.1	8 34.0 ± 2.3	5 46.8 ± 1.2	19 57.7 ± 0.2
Olfactory Bulb	12 78.5 ± 0.5	3 36.7 ± 0.7	3 49.2 ± 1.8	14 59.5 ± 1.5
Cingulate Cortex	13 78.2 ± 0.2	16 32.52 ± 0.03	20 42.6 ± 0.1	13 59.6 ± 0.1
Temporal Lobe	14 77.8 ± 0.2	17 32.5 ± 0.3	10 45.31 ± 0.04	10 60.0 ± 0.2
Prefrontal Cortex	15 77.7 ± 0.1	20 31.5 ± 0.2	15 43.4 ± 0.2	8 61.0 ± 0.1
Placenta	16 77.6 ± 0.4	10 33.3 ± 0.9	12 45.0 ± 0.6	16 58.8 ± 0.3
Hypothalamus	17 77.1 ± 0.2	31 29.6 ± 0.2	25 41.2 ± 0.3	12 59.6 ± 0.3
Spinal Cord	18 76.7 ± 1.3	26 30.8 ± 1.0	13 44.5 ± 1.5	11 59.9 ± 0.7
Occipital Lobe	19 76.3 ± 1.0	28 30.1 ± 0.4	23 41.6 ± 0.7	15 59.0 ± 0.5
Fetal Thyroid	20 76.0 ± 0.9	27 30.3 ± 0.7	2 49.2 ± 2.8	34 51.5 ± 1.2
Parietal Lobe	21 75.72 ± 0.03	24 31.03 ± 0.03	24 41.4 ± 0.4	21 56.9 ± 0.1
Thyroid	22 75.6 ± 0.8	11 33.3 ± 0.5	16 43.3 ± 0.5	23 56.3 ± 0.2
Medulla Oblongata	23 75.6 ± 0.1	30 29.7 ± 0.5	22 41.9 ± 0.8	18 57.9 ± 0.2
Kidney	24 75.4 ± 0.1	25 30.9 ± 0.4	26 41.1 ± 0.3	27 54.5 ± 0.2
Subthalamic Nucleus	25 75.0 ± 0.1	21 31.3 ± 0.4	28 39.8 ± 0.3	20 57.6 ± 0.1
Liver	26 74.9 ± 0.2	19 31.9 ± 0.9	27 40.4 ± 0.1	29 54.3 ± 0.2
Tumor	27 74.5 ± 0.4	39 28.0 ± 1.5	37 35.8 ± 1.4	32 54.2 ± 0.9

(Table 3.1 to be continued on the next page)

(continued Table 3.1)

Tissues	The Ranking* (descending) and Score (%) of G4 Enrichment in Different Tissues			
	Whole TRR	**-2,000 ~ -501 bp**	**-500 ~ -1 bp**	**TSS ~ +1,000 bp**
White Blood Cells	28 73.3 ± 0.3	35 29.1 ± 0.4	31 38.9 ± 0.8	25 55.0 ± 0.1
Smooth Muscle	29 73.1 ± 1.5	13 33.1 ± 3.0	4 48.6 ± 2.3	28 54.4 ± 0.6
Pons	30 72.9 ± 0.2	29 29.9 ± 0.1	30 39.0 ± 0.1	31 54.3 ± 0.2
Globus Pallidus	31 72.5 ± 0.2	32 29.5 ± 0.1	33 38.48 ± 0.04	30 54.3 ± 0.1
Bone Marrow	32 72.1 ± 0.7	15 32.6 ± 0.8	48 33.5 ± 1.0	24 55.9 ± 1.3
Whole Blood	33 72.1 ± 0.4	36 28.4 ± 0.2	34 38.1 ± 0.9	26 54.7 ± 1.7
Testis	34 71.0 ± 0.1	42 27.4 ± 0.2	39 35.8 ± 1.2	33 52.0 ± 0.2
Uterus Corpus	35 70.8 ± 2.0	5 34.2 ± 0.2	21 42.3 ± 1.1	41 48.9 ± 3.3
Tongue	36 70.1 ± 0.2	12 33.2 ± 0.7	29 39.4 ± 1.1	40 49.0 ± 0.4
Testis Seminiferous Tubule	37 69.5 ± 0.1	47 26.1 ± 0.1	46 33.8 ± 0.2	35 50.8 ± 0.1
Skeletal Muscle	38 69.1 ± 0.1	34 29.3 ± 0.1	36 36.6 ± 0.1	36 49.9 ± 0.1
Testis Interstitial	39 68.7 ± 0.3	49 25.3 ± 0.2	47 33.6 ± 0.2	38 49.2 ± 0.3
Ciliary Ganglion	40 68.42 ± 0.04	37 28.4 ± 0.1	42 35.44 ± 0.04	37 49.3 ± 0.1
Trachea	41 68.3 ± 0.8	43 27.3 ± 1.7	35 37.9 ± 1.3	49 46.5 ± 1.5
Superior Cervical Ganglion	42 68.3 ± 0.1	38 28.21 ± 0.01	38 35.8 ± 0.1	39 49.17 ± 0.03
Dorsal Root Ganglion	43 67.8 ± 0.1	45 27.0 ± 0.1	40 35.6 ± 0.1	43 48.5 ± 0.1
Whole Genome	44 67.79 ± 0.01	33 29.38 ± 0.03	43 35.031 ± 0.004	44 48.23 ± 0.01
Trigeminal Ganglion	45 67.53 ± 0.03	41 27.9 ± 0.1	41 35.5 ± 0.1	42 48.82 ± 0.01
Skin	46 67.5 ± 0.2	44 27.1 ± 0.5	45 34.4 ± 0.7	47 47.1 ± 0.4
Atrioventricular Node	47 67.401 ± 0.005	40 27.93 ± 0.01	44 35.0 ± 0.1	45 48.21 ± 0.01
Lymph Node	48 65.3 ± 1.1	46 27.0 ± 2.0	32 38.5 ± 3.4	48 47.0 ± 2.0
Fetal Liver	49 64.0 ± 0.3	48 25.9 ± 0.9	50 28.8 ± 1.3	46 47.4 ± 0.7
Appendix	50 63.17 ± 0.05	50 24.1 ± 0.1	49 30.5 ± 0.4	50 43.8 ± 0.8

* Ranking of G4 enrichment in TRRs of genes active in a specific tissue was performed by comparing the mean value of G4 enrichment in that tissue with other tissues.

Compared with other tissues, several tissues with high G4 enrichment in the full-length TRR (at whole TRR level) are also highly G4-enriched when partial regions of TRR were analysed, such as the lung, whole brain, caudate nucleus, heart, and cerebellum peduncles. The high G4 enrichment in these tissues may indicate that these tissues have evolved certain mechanisms to utilize G4s to regulate gene transcription (i.e. providing binding sites for tissue-restricted TFs, maintain the separation of coding and template strands, and so on), or at least to avoid possible negative effects caused by the formation of G4s in TRR (i.e. to resolve folding of G4 by helicases). For tissues with the lowest G4 enrichment at the whole TRR level, corresponding G4 enrichment in different regions of TRR is also among the lowest, such as trigeminal ganglion, skin, atrioventricular node, lymph node, fetal liver, and appendix (Table 3.1). Interestingly, for some tissues, although moderate or relatively low G4 enrichment was found at whole TRR level, G4s were found to be enriched in either distal or proximal promoters in the TRRs. For example, in smooth muscle, although the overall G4 enrichment (73.1 ± 1.5 %) is only moderately higher than that found at the whole genome level, its G4 enrichment in the proximal promoter region is among the highest (48.6 ± 2.3 %, the forth highest). Similar phenomena were found in tissues of adipose, olfactory bulb, and fetal thyroid. This may imply that, in these tissues, G4s are involved in regulating tissue-specific gene transcription by interacting with TF binding in the proximal promoter regions.

Although the candidate genes are rather limited, the high G4 content in TRRs of cardiac function-related genes (genes active in the heart) may indicate potential roles of G4s in cardiac functions and diseases.

According to the above findings, several points became apparent. Firstly, enrichment of G4 in TRR of genes has a strong tissue-specificity. For example, G4s are particularly enriched in the whole TRRs of genes active in the lung, whole brain, caudate nucleus, heart, and cerebellum peduncles. Secondly, in some tissues (smooth muscle, adipose tissue, olfactory bulb, and fetal thyroid), G4s are enriched specifically in certain regions of TRRs (location-specificity) reflecting possible tissue-related regulatory roles of G4s. These tissues, with either tissue- or location-specific enrichment of G4s in TRRs, are worthy of further investigation in the future.

From the studies described here, G4s appear to be highly abundant in TRRs of genes active in the cardiac function-related tissues, such as the heart and smooth muscle. To further investigate the biological significance of G4s in the heart, I analysed the abundance and location significance of G4s in TRRs of genes involved in different pathways described within the KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY database, especially those related to cardiac functions. The KEGG PATHWAY is an online databases records networks of molecular interactions in the cells [47]. The importance of G4s in each TRR was evaluated by two scores: G4 abundance score (*F* score, and its cumulative frequency: *CF*) and G4 location significance score (*Q* score, and its cumulative frequency *CQ*), which will be described separately in following sections.

3.3 G4 abundance in TRRs of transcripts (the *F* score)

As compared to duplex DNA, G4s are normally less stable (reviewed in [48-50]). Particularly in genomic DNA, the formation of would be even more difficult in the presence of adjacent duplex DNA or compact chromatins. Thus, high G4 content (abundance) in a particular TRR can enhance the competition of G4 DNA with adjacent duplex DNA, resulting in the increased probability of the involvement of G4s in regulating downstream gene expression. Based on this assumption, the *F* score representing the abundance of G4s in a particular TRR was used as the first indicator to evaluate the importance of G4 in the TRR of each gene. For each transcript, the *F* score was calculated based on the percentage of nucleotides involved in G4 formation in the TRR of that transcript (Eq. 2.26). By including all transcripts with redundancy, the coding strand, template strand (non-coding strand), and both strands were analysed separately to give the F_{cd}, F_{ncd}, and F_o scores, respectively. F_{cd} scores ranged from 0 % to 33.2 % with a median of 1.6 % and an average of 2.1 %, while F_{ncd} scores ranged from 0 % to 29.6 % with a median of 1.4 % and an average of 1.9 % (Table 3.2). Higher maximum, median and average F_{cd} values indicate that G4s are more enriched on the coding strand than on the template strand. The overall G4 abundance scores F_o ranged from 0 % to 33.2 % with a median of 2.2 % and an average of 2.9 % (Table 3.2).

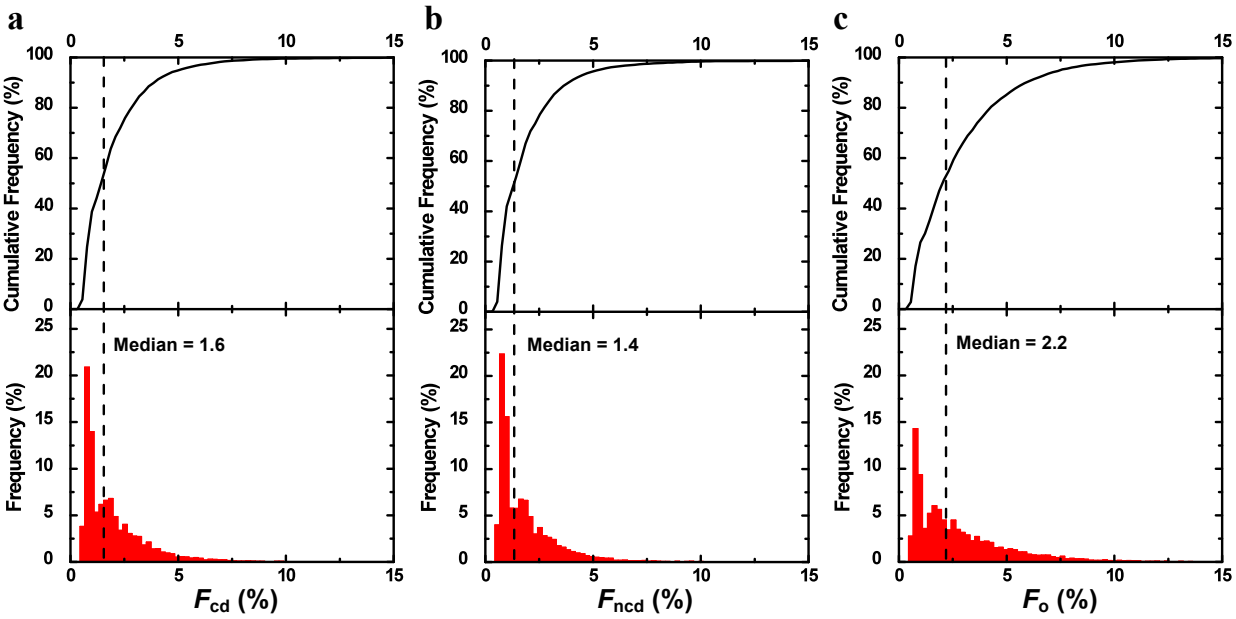


Figure 3.2: Distribution of G4 abundance scores (the *F* scores) and corresponding cumulative frequencies (the *CF* scores).

(a) Distribution of G4 abundance scores (F_{cd}, lower graph) and corresponding cumulative frequency (CF_{cd}, upper graph) by analysing the coding strand of all transcripts available in ENSEMBL database. The position and corresponding values of the first and second peak in the histogram are 0.8 % (peak position) and 20.9 % (peak value), and 1.9 % (peak position) and 6.8 % (peak values) respectively. **(b)** Distribution of G4 abundance scores (F_{ncd}, lower graph) and corresponding cumulative frequency (CF_{ncd}, upper graph) by analysing the template strand (noncoding strand) of all transcripts available in ENSEMBL database. The position and corresponding values of the first and second peak in the histogram are 0.8 % (peak position) and 22.4 % (peak value), and 1.7 % (peak position) and 6.8 % (peak values) respectively. **(c)** Distribution of G4 abundance scores (F_o, lower graph) and corresponding cumulative frequency (CF_o, upper graph) by analysing both of the coding and template strands of all transcripts available in ENSEMBL database. The position and corresponding values of the first and second peak in the histogram are 0.8 % (peak position) and 14.3 % (peak value), and 1.7 % (peak position) and 6.0 % (peak values) respectively. In the analysis, TRRs from 46,205 transcripts exported from ENSEMBL with redundancy were included. TRRs of transcripts without G4-forming motifs were excluded in calculating the distribution and corresponding cumulative frequency.

Figure 3.2 shows the distribution of the F_{cd}, F_{ncd}, and F_o scores obtained by including all transcripts with redundancy. A similar skewed distribution of G4 abundance was reported previously in an analysis showing G4 formation potential (G4P) in RefSeq of genes [43]. However, due to the higher resolution in this study (using the percentage of nucleotides involved rather than the percentage of a shifting searching window containing G4-forming motifs), two peaks were found in the distribution of G4

abundance scores. The first peak, centred at 0.8 %, is dominant in all histograms (Figure 3.2, lower graphs), which corresponds to a single G4-forming motif with a length of ~ 23 nucleotides. The second peak, centred at 1.9 % on the coding strand and 1.7 % on the template strand and both strands, respectively, may correspond to a region containing two G4-forming motifs with a length of 50 ~ 56 nucleotides. By comparing the peak position and corresponding values in the distribution histograms of G4 abundance on coding and template strand (Figure 3.2 a and b, lower graphs), a smaller first peak and a larger second peak were found in the histogram of the coding strand (Figure 3.2 a, lower graph). This also implies that G4-forming motifs are more enriched in the coding strand as compared to the template strand.

All values were re-calculated with 10 iterations with randomly chosen transcripts, and compared with values from transcripts with redundancy, and similar results were obtained (Table 3.2). The CF_{cd}, CF_{ncd}, and CF_{o} values for each transcript were calculated as the cumulative frequency according to the F_{cd}, F_{ncd}, and F_{o} histograms (Figure 3.2, upper graphs).

Table 3.2: Comparison of G4 abundance in TRRs of transcripts (the *F* score)

	Median Value	Mean Value	Max Value
Transcripts with Redundancy (46,205 Transcripts)			
$\boldsymbol{F_{cd}}$	1.6	2.1	33.2
$\boldsymbol{F_{ncd}}$	1.4	1.9	29.6
$\boldsymbol{F_{o}}$	2.2	2.9	33.2
Transcripts without Redundancy (19,633 Transcripts), 10 iterations *			
$\boldsymbol{F_{cd}}$	1.55 ± 0.02	2.03 ± 0.01	33.2
$\boldsymbol{F_{ncd}}$	1.42 ± 0.02	1.93 ± 0.01	25.88 ± 3.96
$\boldsymbol{F_{o}}$	2.13 ± 0.01	2.90 ± 0.01	33.2

* The F_{cd}, F_{ncd}, and F_{o} values from 1st iteration were used as templates and compared with those from following iterations by one-way ANOVA test. No differences among iterations were observed ($P > 0.5$).

3.4 G4 location significance in TRRs of transcripts (the *Q* score)

Besides the abundance, the locations of G4s in a particular TRR are also important in estimating the potential biological significance of G4. For example, Du *et al.* suggested that the colonization of G4-forming motifs with TF binding sites in the proximal promoters indicates potential regulatory roles of G4s [35]. Thus, the biased probability distribution of G4-forming motifs in TRRs at whole genome level may

reflect the preference of G4 location in TRRs through evolution. As shown in Figure 3.3 a, the probability distributions (Eq. 2.27) of G4 on coding and template strand are in general very similar, except that the coding strand has a higher probability at the peak regions (-50 bp ~ TSS) and regions downstream to the TSS. Maximum probability of G4s, 3.4 % and 3.0 %, were found at -55 bp and -57 bp relative to the TSS on the coding and template strand, respectively, which also indicates the higher enrichment of G4s in the proximal promoter region on the coding strand as compared to the template strand [30]. However, G4s are apparently more enriched on the coding strand downstream to the TSS (TSS ~ +1,000 bp). This may imply the functional differences of G4s on coding and template strand in regions upstream and downstream to the TSS. Similar probability distribution of G4s on the coding and template strand in regions upstream to the TSS indicates that promoter G4s on both strands possibly may have similar functions as molecular switches or TF binding sites to modulate downstream gene transcription. On the other hand, in regions downstream of the TSS, temporarily blocked rehybridization induced by G4s on coding strand have been shown to be favoured by the transcription process, while G4s on the template strand are detrimental [42]. Similar probability distribution of G4 around TSS has been reported in the genomes of other warm-blooded animals [34]. This location and strand polarity-dependent distribution of G4s therefore appears to reflect an evolutionary pressure, and thus it is necessary to take into account the location of G4 motifs when considering their importance in the TRR of a particular gene. Therefore, a location significance score Q was introduced in order to characterise the location significance of G4s. For each TRR, its G4 location significance scores on the coding (Q_{cd}) and template (Q_{ncd}) strand were calculated according to Eq. 2.27 and 2.28. The overall G4 location significance score (Q_o) was determined as the sum of Q_{cd} and Q_{ncd}. Histograms Q_{cd}, Q_{ncd}, and Q_o for TRRs of all transcripts with redundancy are shown in Figure 3.3 b ~ d (lower graphs). Q_{cd}, Q_{ncd}, and Q_o ranged from 0 to 26.9, 19.8, and 29.3, respectively. The median values of the three scores were 2.5 (Q_{cd}), 1.7 (Q_{ncd}), and 3.2 (Q_o), while average values were 3.3 (Q_{cd}), 2.3 (Q_{ncd}), and 4.2 (Q_o), respectively. The larger maximum, median, and average Q scores on the coding strand also reflect the higher location significance of G4s on the coding strand as compared to the template strand.

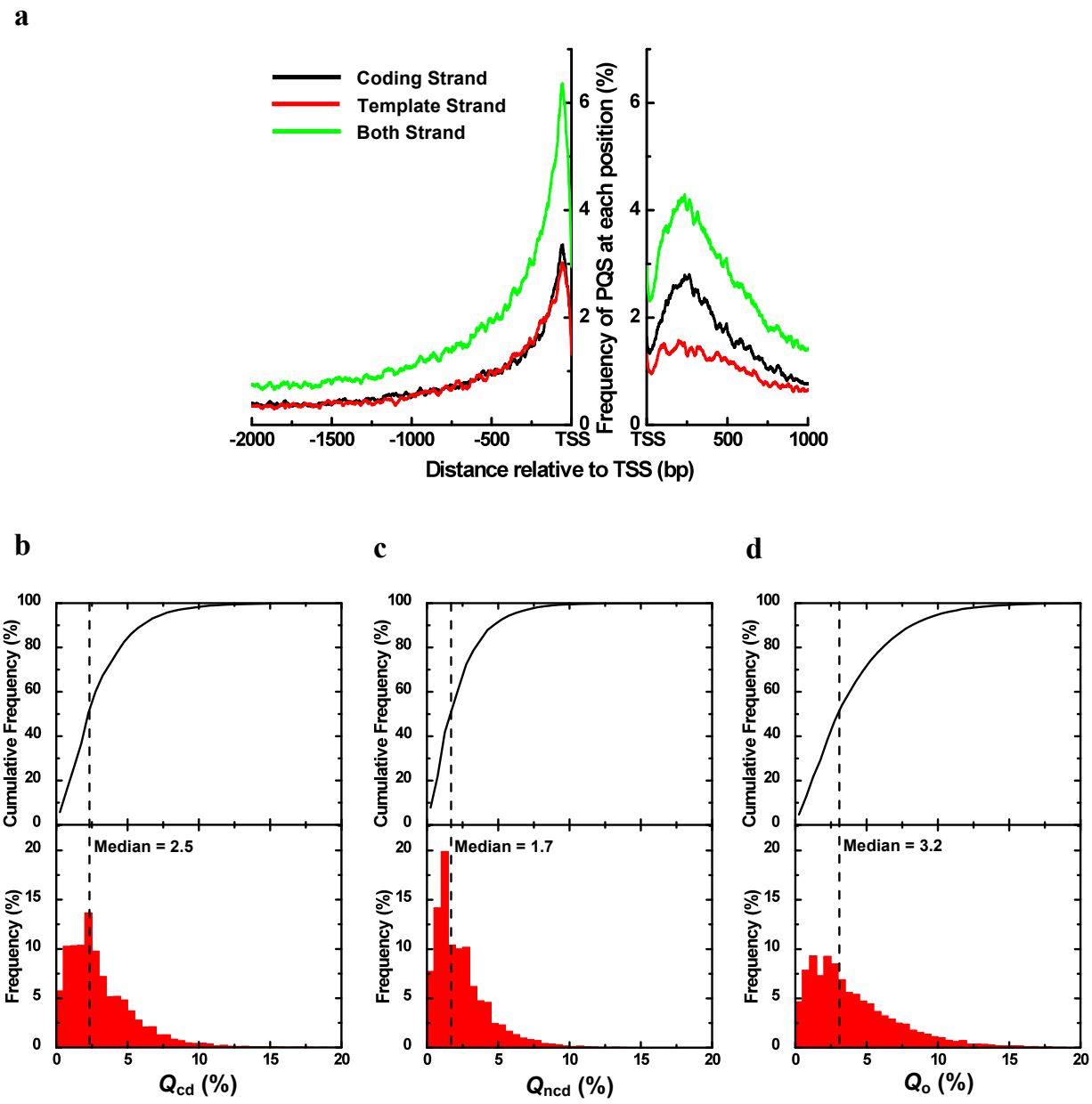


Figure 3.3: Probability distribution and location significance scores (the *Q* scores) distribution of G4s in TRRs of transcripts.
(a) Probability distribution of each nucleotide involved in G4 formation on coding strand (black), template strand (red), and both strands (green) in TRRs of all transcripts. TRRs from 46,205 transcripts exported from ENSEMBL with redundancy were included. **(b)** Distribution of G4 location significance (Q_{cd}, lower graph) and corresponding cumulative frequency (CQ_{cd}, upper graph) by analysing the coding strands of all transcripts available in ENSEMBL database. **(c)** Distribution of G4 location significance (Q_{ncd}, lower graph) and corresponding cumulative frequency (CQ_{ncd}, upper graph) by analysing the template strands of all transcripts available in ENSEMBL database. **(d)** Distribution of G4 location significance (Q_o, lower graph) and corresponding cumulative frequency (CQ_o, upper graph) by analysing both strands of all transcripts available in ENSEMBL database. TRRs of transcripts without G4-forming motifs were excluded in calculating the distribution and corresponding cumulative frequency.

Transcripts without redundancy were also analysed with 10 iterations of randomly chosen transcripts, and similar results were obtained as compared to the results from transcripts with redundancy (Table 3.3). The *CQ* values for each transcript were calculated as its cumulative frequency in *Q* histogram and are shown in Figure 3.3 b ~ d (upper graphs).

Table 3.3: Comparison of G4 location significance in TRRs of transcripts (the *Q* score)

	Median Value	Mean Value	Max Value
Transcripts with Redundancy (46,205 Transcripts)			
Q_{cd}	2.5	3.3	26.9
Q_{ncd}	1.7	2.3	19.8
Q_{o}	3.2	4.2	29.3
Transcripts without Redundancy (19,633 Transcripts), 10 iterations *			
Q_{cd}	2.48 ± 0.02	3.13 ± 0.02	23.4 ± 1.6
Q_{ncd}	1.85 ± 0.02	2.40 ± 0.02	19.7 ± 0.2
Q_{o}	3.20 ± 0.02	4.07± 0.01	28.3 ± 0.2

* The Q_{cd}, Q_{ncd}, and Q_{o} values from 1st iteration were used as templates and compared with those from following iterations by one-way ANOVA test. No differences among iterations were observed ($P > 0.23$).

3.5 Correlation between G4 abundance and G4 location significance scores

The correlation coefficient between *CF* and *CQ* were calculated to evaluate the correlation between abundance and location significance of G4s in TRRs. As shown in Figure 3.4, *CF* and *CQ* are plotted as correlation maps. In each map, TRRs with both values higher than 50 % are regarded as G4-important TRRs, while those with both values lower than 50 % are G4-less-important TRRs. According to the maps, *CF* values are positively correlated to *CQ* values. Correlation coefficients between *CF* and *CQ* on coding strand (CF_{cd} vs. CQ_{cd}), template strand (CF_{ncd} vs. CQ_{ncd}), and both strands (CF_{o} vs. CQ_{o}) are 0.74, 0.72, and 0.84, respectively. However, corresponding correlation coefficients for G4-important TRRs are 0.61 (coding strand), 0.58 (template strand), and 0.66 (both strands), respectively, while those for G4-less important TRRs are 0.10, 0.09, and 0.34 respectively. Apparently, in G4-important TRRs, their G4 abundance is more positively correlated to their G4 location significance. To investigate whether this positive correlation of G4-important TRRs is

only caused by the increase of G4 content or not, G4s found in different regions of TRRs were subjected to further investigation.

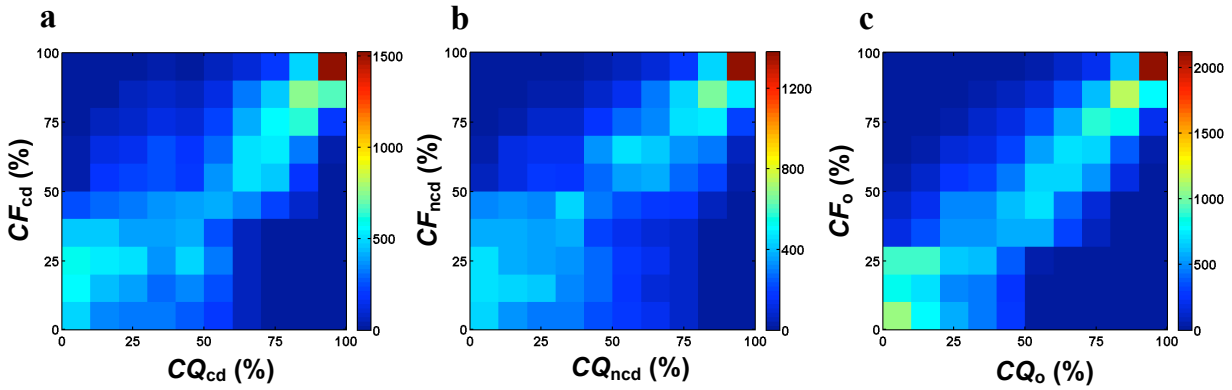


Figure 3.4: Correlation between the location significance (the *CQ* score) and abundance (the *CF* socre) of G4s identified in TRRs of all transcripts with redundancy.

(a) Correlation between G4 location significance and G4 abundance on the coding strand (CQ_{cd} vs. CF_{cd}). **(b)** Correlation between G4 location significance and G4 abundance on the template strand (CQ_{ncd} vs. CF_{ncd}). **(c)** Correlation between G4 location significance and G4 abundance on both strands (CQ_o vs. CF_o). In each map, transcripts with the TRRs *CQ* and *CF* scores larger than 50 % are identified as G4-important transcripts, and those with both values smaller than 50 % are G4-less-important (or less G4-important) transcripts. G4 abundance and its location significance in G4-important ranscripts are well correlated (correlation coefficients are 0.61, 0.58, and 0.66 in a, b c respectively). Correlation between those two values in G4-less-important transcripts are not significant (correlation coefficients are 0.10, 0.09, and 0.34 in a, b c respectively).

The TRR was divided into three regions according to their different W_k values and different biological functions. Region I, from -2,000 bp to -501 bp, represents the distal promoter region, which has the lowest W_k values and is possibly of the least importance with regard to downstream gene transcription. Region II, the proximal or key promoter region, is from -500 bp to -1, with the highest W_k values and the highest importance in regulating downstream gene expression. The region between TSS (+1) and +1,000 bp is denoted as Region III, with medium W_k values, and G4s found in this region function mainly to maintain the separation of coding and template strands. The total number of G4s identified in these regions in G4-important (*CF* and *CQ* values are higher than 50 %) and less G4-important (*CF* and *CQ* values are lower than 50 %) transcripts were counted and normalized to the length of different regions. If the increase of *CQ*, the location significance score, was solely dependent on the

increase of *CF*, the G4 abundance score, then the proportion of the normalized total number of G4 found in different regions should be in constant between G4-important and less G4-important TRRs. However, it was found that the proportion of G4s found in Region II, the proximal promoter region, was increased by 20 % from less G4-important TRRs to G4-important TRRs (Figure 3.5). This indicates that the larger *CQ* value in G4-important TRR reflects the combined effects of increased total number of G4s and particularly the increased number of G4 in key promoter regions (with the high W_k values). Most well-studied genes containing promoter G4s with potential regulatory functions are in the group of G4-important transcripts in the correlation map, including *MYC*, *VEGF-A*, *BCL-2*, *HIF1-α*, *MYB*, *PDGF-A*, *PDGFR-β*, *KRAS*, and *TERT* (Table 3.4). One interesting exception is *KIT*, with CF_o and CQ_o of 43.5 % and 73.4 %. This may suggest that G4 location might be more relevant than the abundance of G4 in evaluating potential biological significance of G4s in a TRR.

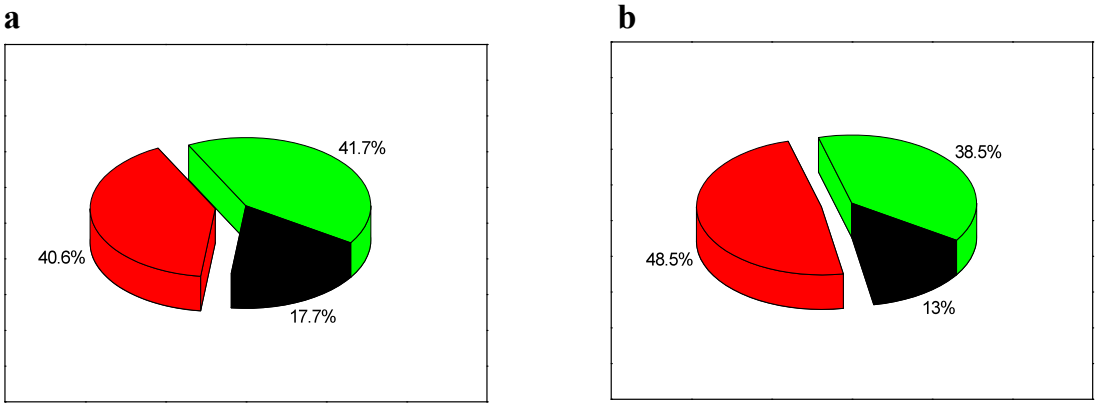


Figure 3.5: The normalized percentage of G4s identified in different regions of TRRs of G4-less-important and G4-importance transcripts.
(a) By normalizing to the total length of different regions, 17.7 %, 40.6 %, and 41.7 % G4s were found in the Region I (distal promoter, -2,000 ~ 500 bp, black), II (proximal promoter, -500 ~ -1 bp, red) and III (downstream region, TSS ~ 1,000 bp, green) in TRRs of G4-less-important transcripts. **(b)** By normalizing to the total length of different regions, 13.0 %, 48.5 %, and 38.5 % G4s were found in the Region I (distal promoter, black), II (proximal promoter, red) and III (downstream region, green) in TRRs of G4-important transcripts. The proportion of G4s found in Region II, the proximal promoter region, is increased by 20 % from G4-less-important to G4-important TRRs, indicating that larger *CQ* value in G4-important TRR is from combined effects of increased total number of G4s, and increased number of G4s in the proximal promoter region (with the high W_k values). This calculation includes TRRs of all 46,205 transcripts with redundancy.

Table 3.4: CF_o and CQ_o values of TRRs of genes containing well-studied promoter G4s

Gene Name	Ensembl ID	CF_o (%)	CQ_o (%)
BCL-2	ENSG00000171791	50.1	56.1
KIT	ENSG00000157404	43.5	73.4
MYC	ENSG00000136997	80.6	64.1
HIF1-α	ENSG00000100644	55.1	56.6
MYB	ENSG00000118513	82.9	77.5
PDGF-A	ENSG00000197461	96.2	78.0
PDGFR-β	ENSG00000113721	72.2	63.9
TERT	ENSG00000164362	68.7	52.9
VEGF-A	ENSG00000112715	75.3	68.9

3.6 G4-important pathways

All pathways composed of more than 50 genes available in KEGG PATHWAY database were extracted, and CF_o and CQ_o values of each gene in a particular pathway were calculated. For genes with multiple transcripts, one transcript was picked randomly, and genes with both values (CF_o and CQ_o) equaling to zero were excluded. Then, in each pathway, percentages of genes with TRRs having both CF_o and CQ_o values higher than 50 % were counted. Pathways with more than 50 % genes' transcripts identified as G4-important were regarded as G4-important pathways as listed in Table 3.5. Ten iterations were carried out to evaluate effects from transcripts redundancy. For all pathways, no significant difference was found in CF_o and CQ_o values among iterations of each individual pathway (one-way ANOVA test, $P > 0.5$). Based on this criterion, G4-important pathways could be grouped into six categories with distinct functions. As shown in Table 3.5, the first category is cancer-related pathways where G4s are enriched in proto-oncogenes, in accordance with previously reported results [43]. Given the fact that G4s are enriched in cancer-related pathways, it is not difficult to understand why the off-target effects of certain G4-binding ligands with anti-cancer potential could well bring some benefits in anti-cancer therapy, such as the G4-binding fluoroquinolone derivative Quarfloxin [51], and the naphthalene diimide derivative 1 [52]. The second category is signaling pathways such as those involving VEGF, Wnt, Hedgehog, and MAPK (mitogen-activated protein kinase), which have key component containing promoter G4s with possible regulatory functions. Pathways related to cell junction, neuron function, and bacterial infection

are also found to be G-important. The last category, heart function-related pathways, is found to be highly G4-important. There are seven pathways in the KEGG PATHWAY database involved in cardiac functions or heart diseases, while six of them are identified as highly G4-important. These pathways are not only involved in normal heart function such as calcium signaling and muscle contraction, but also in various cardiomyopathies, making targeting promoter G4s at pathway level rather than in single gene an attractive strategy (the multi-targeting strategy) in the treatment of certain cardiovascular diseases.

Table 3.5: KEGG pathways with high G4-importance

KEGG Pathway ID	Pathway Description	% of High CF_o and CQ_o	No. of Gene	% of Gene with G4 in TRR	P_{max} of CQ_o *	P_{max} of CF_o *
Cancer-related pathways						
hsa05210	Colorectal cancer	57.9 ± 4.5	62	86.3 ± 2.2	6E-4	0.001
hsa05211	Renal cell carcinoma	57.6 ± 3.3	71	83.9 ± 2.8	7E-4	0.003
hsa05213	Endometrial cancer	56.9 ± 3.7	52	84.8 ± 3.1	0.019	0.037
hsa05217	Basal cell carcinoma	53.7 ± 2.9	55	93.8 ± 1.3	4E-5	6E-6
hsa05200	Pathways in cancer	53.5 ± 1.3	327	82.5 ± 1.4	3E-11	3E-11
hsa05223	Non-small cell lung cancer	53.5 ± 2.6	54	86.9 ± 3.1	0.003	0.02
Signaling pathways with promoter G4s in key components						
hsa04340	Hedgehog signaling pathway	60.9 ± 3.4	56	87.3 ± 1.8	2E-4	1E-4
hsa04370	VEGF signaling pathway	60.7 ± 3.0	80	87.9 ± 2.1	8E-6	1E-4
hsa04010	MAPK signaling pathway	59.5 ± 1.5	271	83.5 ± 1.3	7E-16	1E-16
hsa04310	Wnt signaling pathway	58.9 ± 2.5	150	89.9 ± 0.9	2E-12	4E-12
Cell junction-related pathways						
hsa04520	Adherens junction	67.0 ± 4.2	72	81.3 ± 1.6	2E-5	6E-5
hsa04510	Focal adhesion	56.9 ± 1.3	199	80.4 ± 1.3	4E-8	8E-8
hsa04810	Regulation of actin cytoskeleton	55.7 ± 2.4	213	79.1 ± 2.3	3E-6	2E-5
hsa04530	Tight junction	53.9 ± 2.1	129	77.0 ± 1.8	8E-4	5E-4
hsa04540	Gap junction	53.8 ± 2.0	88	86.1 ± 1.9	8E-5	6E-5
Neural cell regulation-related pathways						
hsa04724	Glutamatergic synapse	61.8 ± 1.6	129	87.1 ± 1.6	5E-9	7E-9
hsa04360	Axon guidance	61.0 ± 2.7	128	86.7 ± 2.2	4E-9	4E-9
hsa04720	Long-term potentiation	60.2 ± 2.0	70	90.6 ± 1.2	1E-6	8E-6
hsa04722	Neurotrophin signaling pathway	55.6 ± 2.1	125	88.9 ± 1.9	4E-8	1E-7
hsa04912	GnRH signaling pathway	55.5 ± 1.3	104	85.9 ± 1.7	1E-6	1E-6
Bacterial infection-related pathways						
hsa05130	Pathogenic E.coli infection	58.6 ± 4.9	55	82.7 ± 3.8	0.004	0.008
hsa05131	Shigellosis	55.8 ± 3.4	62	86.3 ± 2.3	8E-5	2E-4
hsa05100	Bacterial invasion of epithelial cells	54.5 ± 2.8	71	82.4 ± 3.7	0.006	0.01
Heart function-related pathways						
hsa05412	Arrhythmogenic right ventricular cardiomyopathy	60.0 ± 3.5	73	80.1 ± 3.5	0.004	0.01
hsa05410	Hypertrophic cardiomyopathy	57.5 ± 3.1	83	77.1 ± 2.7	0.01	0.04
hsa05414	Dilated cardiomyopathy	57.2 ± 3.1	90	78.2 ± 2.8	0.009	0.02
hsa04020	Calcium signaling pathway	54.4 ± 2.1	178	83.1 ± 1.5	5E-6	5E-6
hsa04260	Cardiac muscle contraction	54.2 ± 3.0	72	73.2 ± 2.9	0.1	0.4
hsa04270	Vascular smooth muscle contraction	53.8 ± 1.0	121	85.2 ± 1.8	1E-7	9E-7

* P_{max} values are calculated by comparing the CQ_o and CF_o scores of genes from individual pathways with those from all transcripts without redundancy in each iteration by Wilcoxon rank sum test. $P \leq 0.05$ means the G4 abundance (CF_o) or location significance (CQ_o) of gene involved in the pathway is significantly different from those at genome level. For all pathways, no significant difference was found in CF_o and CQ_o values among iterations of each individual pathway (one-way ANOVA test, $P > 0.5$).

3.7 Discussion

The correlations between G4s and gene function, particularly of those involved in cancer development, has been investigated previously [43]. Furthermore, there is now compelling evidence to support the existence of G4 in living cells, such as G4-induced guanine-protection revealed by *in vitro* DMS footprinting [18] (reviewed in [53]). Thus, together with the identification of the potential regulatory roles of G4s in promoters of several proto-oncogenes [5, 6, 9, 17, 54-56], this has inspired the interest to utilize G4-binding ligands as potential anti-cancer drugs (as reviewed in [57, 58]). However, although it become clear that G4-forming motifs have a functional role in living cells and there are some good successes in designing specific G4-binding ligands, discriminating and targeting specific G4s via these small synthetic molecules is still very challenging at present due to several hurdles as mentioned in Section 1.5. Among all these hurdles, the selectivity of G4-binding ligands is the most notable one that warrants further improvement. For example, although there are some ligands reported to interact with G4s much more selectively with respect to duplex DNA, such as Mn^{III} porphyrin (10^4-fold) [59] and dimetallic terpyridine-based ligand (over 100-fold) [60], the selectivity of most ligands are in the range between 10- and 100-fold (reviewed in [58]). Furthermore, very few ligands have been reported to exert satisfactory selectivity between different G4s. Taking into account the facts that G4s may occur transiently in the human genome and are very much outnumbered by adjacent duplex DNA, G4-binding ligands with ideal selectivity and affinity to a particular promoter G4 structure would be difficult to achieve at the moment, and off-target effects would be inevitable. Thus, all these factors have limited the therapeutic potential of G4-binding ligands *in vivo*. On the other hand, accumulating evidence has shown that off-target effects or multi-targeting effects may be advantageous for G4-binding ligands and may enhance their therapeutic potential in particular instances. For example, the Quadruplex-binding fluoroquinolone derivative Quarfloxin (CX-3543), which is in clinical trials as an anti-cancer drug, has been shown to bind to *MYC* promoter G4 and to selectively disrupt nucleolin/rDNA quadruplex complexes [51]. Another small G4-binding molecule, naphthalene diimide derivative 1, can bind to *KIT* promoter G4 and telomeric G4, and can efficiently arrest proliferation of human gastrointestinal stromal tumor (GIST) cells [52]. According to these examples, it is reasonable to hypothesize that targeting multiple similar G4s with a similar

biological function is feasible and, of more importance, this multi-targeting strategy might provide an alternative to compromise the selectivity issue of G4-binding ligands, and enhance their therapeutic potential *in vivo*. To test this hypothesis, a systematic overview about the abundance of G4s in TRRs of genes involved in different biological processes would be necessary. Thus, in this chapter, G4-tissue and G4-pathway correlations in TRRs of genes were investigated.

3.7.1 Tissue-specific enrichment of G4s in TRRs of genes

In Section 3.2, the tissue-specific enrichment of G4s in TRRs of genes has been clearly shown (Figure 3.1, Table 3.1). For example, for genes active in the lung, as high as 82.8 ± 0.4 % of them contain at least one G4-forming motif in the TRRs, while in the appendix, only 63.17 ± 0.05 % genes contain at least one G4-forming motif in their TRRs. For most tissues, the G4 enrichment in different regions of TRRs is relatively consistent. For example, for genes active in the lung, whole brain, caudate nucleus, heart, and cerebellum peduncles, high G4 enrichment was found in the distal, proximal promoter, or the downstream regions. However, there are also some exceptions. For example, for genes active in the olfactory bulb, fetal thyroid, smooth muscle, and uterus corpus, G4s are highly enriched in either distal or proximal promoter regions, but not in the downstream regions. Furthermore, G4s are asymmetrically distributed on the coding and template strands in different regions of TRRs. Enrichment of G4s in different regions around TSS is known to associate with diverse biological processes. For example, G4s in the distal or proximal promoters of proto-oncogenes normally act as repressors in regulating downstream gene transcription (*MYC* promoter G4 [5, 6]). G4s in the downstream regions relative to TSS can modulate mRNA translation efficiency differently depending on the strand polarity [33, 42]. Thus, this tissue-specific, location-specific, and asymmetric G4 enrichment may also imply that tissues with high G4 abundance in TRRs or certain regions of TRRs have evolved certain tissue-specific mechanisms to tolerate or even to utilize G4s in the processes of transcription regulation.

Another uncertain point is that high G4 enrichment was found in most tissues as compared to the whole genome (G4 enrichment at whole genome level ranked 44, 33, 43, 44 in the full-length TRR, distal promoter, proximal promoter, and downstream

regions, Table 3.1). Although this may imply that there are a group of ubiquitous genes with high G4 abundance in their TRRs involved in different tissues, the actual reason is still unclear at this stage. Thus, this point will be further investigated in future work. For example, the overlap of genes containing G4s in their TRRs will be analysed and compared among different tissues.

3.7.2 G4-important pathways

To further evaluate G4-importance in TRRs of genes involved in different biological processes, the importance of G4s in the TRRs of genes in different pathways was also investigated. Previous genome-wide studies have found that promoter G4s have a strong positional bias towards TSS in different species (mentioned in Section 3.1), and this positional bias of promoter G4s is believed to be a result from evolution selection [29]. Furthermore, in the previous study, location-specific enrichment of G4s has been found in TRRs of genes active in several tissues. Thus, in addition to the G4 abundance (*F* and *CF* scores), the location significance of G4 (*Q* and *CQ* scores) is also considered to evaluate the biological importance of G4s in each TRR. By employing these two indicators, higher resolution of G4-importance in different genes could be achieved. Although a strong overall correlation between *F* and *Q* scores was found in TRRs, differences were also revealed in TRRs with different G4-importance and different biological functions. Generally, correlation between *F* and *Q* scores in G4-less-important TRRs (correlation coefficient: 0.34) is not as strong as that in G4-important TRRs (correlation coefficient: 0.66) (Figure 3.4). Furthermore, when G4-importance increases, the probability to find G4s in the proximal promoter regions (-500 bp ~ -1 bp) of TRR increases as well (Figure 3.5). Because the majority of promoter G4s with proven regulatory functions locate in the proximal promoter regions, it is highly likely that the accumulation of proximal promoter G4s found in pathways with high G4-importance is biologically relevant (i.e. G4-mediated mechanisms may exist in these pathways). In regions downstream to the TSS, however, the probability to find G4s in both G4-important and G4-less important genes is relatively constant, suggesting that the transcription induced G4 formation might be ubiquitous in different genes.

G4-important pathways show a significant correlation with their functions. For example, G4-important genes are concentrated in neuron function, cell junction and bacterial infection-related pathways. In accordance with previous findings, most cancer-related pathways are determined as G4-important, which may suggest that off-target effects of certain anti-cancer G4-binding ligands can provide extra benefits for cancer treatment. Signaling pathways with key components containing biologically relevant G4s in proximal promoters are identified as highly G4-important as well, such as Hedgehog, VEGF, Wnt, and MAPK signaling pathways. This may imply that in these pathways, G4-mediated regulation might be well tailored to be a common regulatory mechanism, further enhancing the potential of applying multi-targeting strategy in these pathways. Heart function-related pathways, especially those involved in various cardiomyopathies, are also found to be highly G4-important, indicating the therapeutic potential of G4s in cardiovascular diseases. Interestingly, less G4-important pathways also relate to their functions. For example, since porphyrin and its derivatives can interact with a variety of G4s via π- π stacking with the end G-tetrads, porphyrin and chlorophyll metabolism pathway (hsa00860; percentage of *CF* & *CQ* > 50 %: 17.4 ± 2.5; P_{max} < 4E-20) is particularly less G4-important.

Based on the their enrichment in TRRs of genes active in the heart, and their high importance in cardiac function-related pathways, G4s are very likely to function as a group of general regulatory elements in cardiac function-related genes. To validate this finding, G4s found in promoters of human myocyte enhancer factor-2D (*MEF2D*) and cardiac troponin I (*TnIc*) were chosen as model systems to investigate possible biological functions of cardiac function-related promoter G4s.

MEF2D belongs to the MEF2 family, which is involved in the development of cardiac hypertrophy, fibrosis, and fetal gene activation as mentioned in Section 1.8.1. The TRRs of all four members of the human MEF2 gene family are found to be highly G4-important (CF_{max} and CQ_{max} values are: *MEF2A*, 79.8 % and 85.1 %, *MEF2B*, 62.8 % and 74.0 %, *MEF2C*, 51.7 % and 61.5 %, and *MEF2D*, 69.0 % and 93.8 %). In the proximal promoter region of human *MEF2D*, four G4-forming motifs were identified, which are at 232, 169, 85, and 18 bases upstream to the TSS of human *MEF2D*. Particularly, the G4-forming motif starting at -169 bp on the template strand contains four runs of a consecutive GGGG repeat, and have similar pattern (the first and third loops are short, and the second loop is relatively long) as compared to G4-

forming motifs in the upstream TRR of *MYC*, *KIT*, *HIF-1α*, *KRAS*, *BCL-2*, *PDGF-A* and *VEGF* [61].

As one of the central components in cardiac muscle contraction pathway, TnIc is also involved in most cardiac function-related pathways. Furthermore, the TRR of *TnIc* is highly G4-important according to the bioinformatics analysis (CF_o 92.6 %, and CQ_o 95.2 %). In the promoter of human *TnIc*, two types of G4-forming motifs were identified, which are designated as "MNSG4" (between -528 and -319 bp upstream to the TSS), and "-80G4" (centred at -80bp to the TSS). Since human TnIc is exclusively expressed in human cardiomyocytes, these G4 elements could be used as ideal models to investigate cardiac-specific function of G4s.

Thus, as the first step to evaluate their biological potentials, the biophysical properties of these promoter G4s found in the human *MEF2D* and *TnIc* were first investigated, which will be described in next chapter.

3.8 References

1. Han, H., and Hurley, L. H. (2000) G-quadruplex DNA: a potential target for anti-cancer drug design, *Trends Pharmacol Sci 21*, 136-142.
2. Balasubramanian, S., Hurley, L. H., and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?, *Nat Rev Drug Discov 10*, 261-275.
3. Kennedy, G. C., German, M. S., and Rutter, W. J. (1995) The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription, *Nat Genet 9*, 293-298.
4. Lew, A., Rutter, W. J., and Kennedy, G. C. (2000) Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ, *Proc Natl Acad Sci U S A 97*, 12508-12512.
5. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, *Proc Natl Acad Sci U S A 99*, 11593-11598.
6. Grand, C. L., Han, H., Munoz, R. M., Weitman, S., Von Hoff, D. D., Hurley, L. H., and Bearss, D. J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo, *Mol Cancer Ther 1*, 565-573.
7. Cogoi, S., and Xodo, L. E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, *Nucleic Acids Res 34*, 2536-2549.
8. Cogoi, S., Paramasivam, M., Spolaore, B., and Xodo, L. E. (2008) Structural polymorphism within a regulatory element of the human KRAS promoter: formation of G4-DNA recognized by nuclear proteins, *Nucleic Acids Res 36*, 3765-3780.
9. Qin, Y., Rezler, E. M., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4, *Nucleic Acids Res 35*, 7698-7713.
10. Phan, A. T., Kuryavyi, V., Burge, S., Neidle, S., and Patel, D. J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter, *J Am Chem Soc 129*, 4386-4392.
11. Fernando, H., Reszka, A. P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A. R., Neidle, S., and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene, *Biochemistry 45*, 7854-7860.
12. Todd, A. K., Haider, S. M., Parkinson, G. N., and Neidle, S. (2007) Sequence occurrence and structural uniqueness of a G-quadruplex in the human c-kit promoter, *Nucleic Acids Res 35*, 5799-5808.
13. Hsu, S. T., Varnai, P., Bugaut, A., Reszka, A. P., Neidle, S., and Balasubramanian, S. (2009) A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics, *J Am Chem Soc 131*, 13399-13409.
14. Kuryavyi, V., Phan, A. T., and Patel, D. J. (2010) Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter, *Nucleic Acids Res 38*, 6757-6773.
15. Guo, K., Pourpak, A., Beetz-Rogers, K., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene, *J Am Chem Soc 129*, 10220-10228.
16. Sun, D., Guo, K., Rusche, J. J., and Hurley, L. H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents, *Nucleic Acids Res 33*, 6070-6080.

17. Sun, D., Liu, W. J., Guo, K., Rusche, J. J., Ebbinghaus, S., Gokhale, V., and Hurley, L. H. (2008) The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex-interactive agents, *Mol Cancer Ther 7*, 880-889.
18. Sun, D., Guo, K., and Shin, Y. J. (2011) Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene, *Nucleic Acids Res 39*, 1256-1265.
19. Sun, D. (2010) In vitro footprinting of promoter regions within supercoiled plasmid DNA, *Methods Mol Biol 613*, 223-233.
20. Murchie, A. I., and Lilley, D. M. (1992) Retinoblastoma susceptibility genes contain 5' sequences with a high propensity to form guanine-tetrad structures, *Nucleic Acids Res 20*, 49-53.
21. Xu, Y., and Sugiyama, H. (2006) Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Res 34*, 949-954.
22. Xu, Y., and Sugiyama, H. (2005) Structural and functional characterizations of the G-quartet and i-motif elements in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Symp Ser (Oxf)*, 177-178.
23. Dexheimer, T. S., Sun, D., and Hurley, L. H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter, *J Am Chem Soc 128*, 5404-5415.
24. Dai, J., Chen, D., Jones, R. A., Hurley, L. H., and Yang, D. (2006) NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region, *Nucleic Acids Res 34*, 5133-5144.
25. Dai, J., Dexheimer, T. S., Chen, D., Carver, M., Ambrus, A., Jones, R. A., and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution, *J Am Chem Soc 128*, 1096-1098.
26. De Armond, R., Wood, S., Sun, D., Hurley, L. H., and Ebbinghaus, S. W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter, *Biochemistry 44*, 16341-16350.
27. Todd, A. K., Johnston, M., and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA, *Nucleic Acids Res 33*, 2901-2907.
28. Huppert, J. L., and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome, *Nucleic Acids Res 33*, 2908-2916.
29. Huppert, J. L., and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome, *Nucleic Acids Res 35*, 406-413.
30. Huppert, J. L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res 36*, 6260-6268.
31. Eddy, J., and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, *Nucleic Acids Res 36*, 1321-1333.
32. Du, Z., Kong, P., Gao, Y., and Li, N. (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome, *Biochem Biophys Res Commun 354*, 1067-1070.
33. Verma, A., Halder, K., Halder, R., Yadav, V. K., Rawal, P., Thakur, R. K., Mohd, F., Sharma, A., and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species, *J Med Chem 51*, 5641-5649.
34. Zhao, Y., Du, Z., and Li, N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals, *FEBS Lett 581*, 1951-1956.
35. Du, Z., Zhao, Y., and Li, N. (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs, *Nucleic Acids Res 37*, 6784-6798.

36. Todd, A. K., and Neidle, S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements, *Nucleic Acids Res 36*, 2700-2704.
37. Kumar, P., Yadav, V. K., Baral, A., Saha, D., and Chowdhury, S. (2011) Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide, *Nucleic Acids Res 39*, 8005-8016.
38. Raiber, E. A., Kranaster, R., Lam, E., Nikan, M., and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro, *Nucleic Acids Res 40*, 1499-1508.
39. Walsh, K., and Gualberto, A. (1992) MyoD binds to the guanine tetrad nucleic acid structure, *J Biol Chem 267*, 13714-13718.
40. Yafe, A., Etzioni, S., Weisman-Shomer, P., and Fry, M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes, *Nucleic Acids Res 33*, 2887-2900.
41. Shklover, J., Weisman-Shomer, P., Yafe, A., and Fry, M. (2010) Quadruplex structures of muscle gene promoter sequences enhance in vivo MyoD-dependent gene expression, *Nucleic Acids Res 38*, 2369-2377.
42. Du, Z., Zhao, Y., and Li, N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription, *Genome Res 18*, 233-241.
43. Eddy, J., and Maizels, N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome, *Nucleic Acids Res 34*, 3887-3896.
44. Rawal, P., Kummarasetti, V. B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S. K., and Chowdhury, S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation, *Genome Res 16*, 644-655.
45. Hershman, S. G., Chen, Q., Lee, J. Y., Kozak, M. L., Yue, P., Wang, L. S., and Johnson, F. B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae, *Nucleic Acids Res 36*, 144-156.
46. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A 101*, 6062-6067.
47. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome, *Nucleic Acids Res 32*, D277-280.
48. Li, W., Wu, P., Ohmichi, T., and Sugimoto, N. (2002) Characterization and thermodynamic properties of quadruplex/duplex competition, *FEBS Lett 526*, 77-81.
49. Chaires, J. B. (2010) Human telomeric G-quadruplex: thermodynamic and kinetic studies of telomeric quadruplex stability, *FEBS J 277*, 1098-1106.
50. Kumar, N., and Maiti, S. (2008) A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes, *Nucleic Acids Res 36*, 5610-5622.
51. Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C. B., Proffitt, C., Trent, K., Whitten, J. P., Lim, J. K., Von Hoff, D., Anderes, K., and Rice, W. G. (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis, *Cancer Res 69*, 7653-7661.
52. Gunaratnam, M., Swank, S., Haider, S. M., Galesa, K., Reszka, A. P., Beltran, M., Cuenca, F., Fletcher, J. A., and Neidle, S. (2009) Targeting human gastrointestinal stromal tumor cells with a quadruplex-binding small molecule, *J Med Chem 52*, 3774-3783.
53. Lipps, H. J., and Rhodes, D. (2009) G-quadruplex structures: in vivo evidence and function, *Trends Cell Biol 19*, 414-422.
54. Douarre, C., Gomez, D., Morjani, H., Zahm, J. M., O'Donohue M, F., Eddabra, L., Mailliet, P., Riou, J. F., and Trentesaux, C. (2005) Overexpression of Bcl-2 is associated with apoptotic resistance to the G-quadruplex ligand 12459 but is not

sufficient to confer resistance to long-term senescence, *Nucleic Acids Res 33*, 2192-2203.
55. Bejugam, M., Sewitz, S., Shirude, P. S., Rodriguez, R., Shahid, R., and Balasubramanian, S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression, *J Am Chem Soc 129*, 12926-12927.
56. McLuckie, K. I., Waller, Z. A., Sanders, D. A., Alves, D., Rodriguez, R., Dash, J., McKenzie, G. J., Venkitaraman, A. R., and Balasubramanian, S. (2011) G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells, *J Am Chem Soc 133*, 2658-2663.
57. Neidle, S. (2010) Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer, *FEBS J 277*, 1118-1125.
58. Monchaud, D., and Teulade-Fichou, M. P. (2008) A hitchhiker's guide to G-quadruplex ligands, *Org Biomol Chem 6*, 627-636.
59. Dixon, I. M., Lopez, F., Tejera, A. M., Esteve, J. P., Blasco, M. A., Pratviel, G., and Meunier, B. (2007) A G-quadruplex ligand with 10000-fold selectivity over duplex DNA, *J Am Chem Soc 129*, 1502-1503.
60. Suntharalingam, K., White, A. J., and Vilar, R. (2010) Two metals are better than one: investigations on the interactions between dinuclear metal complexes and quadruplex DNA, *Inorg Chem 49*, 8371-8380.
61. Qin, Y., and Hurley, L. H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, *Biochimie 90*, 1149-1171.

Chapter 4 Biophysical Characterisations

4.1 Introduction

Knowing the biophysical properties of a G4 not only pave the way for subsequent *in vitro* studies, but also shed light on the design of novel G4-binding ligands. Thus, in solution biophysical characterisations have been regarded as an initial task to study potential biological functions of a G4. For example, in the field of studying cancer-related G4s, their biophysical profiles have been investigated in detail, such as G4s found in human telomere ([1-6], reviewed in [7] and in the first chapter of this thesis), and in promoters of *MYC* [8, 9], and *KIT* [10-13] proto-oncogenes. However, as compared to those thought to be related to and/or involved in cancers, G4s found in other genes have been insufficiently studied to date. Given the fact that G4s are identified in more than 40 % of promoter regions within the human genome [14], and promoter G4s are particularly enriched in cardiac function-related genes as have been discussed in the Introduction and Chapter 3, promoter G4s found in cardiac function-related genes are worth studying. Thus, G4-forming sequences found in the promoters of two genes, the human *MEF2D* and *TnIc* genes, were chosen and subjected to biophysical studies as described in this chapter.

MEF2 factors belong to the MADS box family of TFs. This TF family is evolutionarily conserved in eukaryotic organisms ranging from yeast to human [15]. Four members coded by separate genes have been identified in the vertebrate MEF2 family, which are MEF2A, MEF2B, MEF2C, and MEF2D (reviewed in [16]). Although the four vertebrate *Mef2* genes are largely expressed in striated muscles and brain, expression of MEF2 proteins are also found in the differential processes of many other cells. In particular, they are the only myogenic TF that are known to be associated with the differentiation of all three muscle cell types (cardiac, skeletal, and smooth) [15, 16]. Interestingly, the TRR regions of all four human MEF2 genes are found to be highly enriched in G4s as determined in Chapter 3 (CF_{max} and CQ_{max} values are: MEF2A, 79.8 % and 85.1 %, MEF2B, 62.8 % and 74.0 %, MEF2C, 51.7 % and 61.5 %, and MEF2D, 69.0 % and 93.8 %). In the TRR of MEF2D, several G4-forming sequences are found around the TSS. Among all these G4-forming

sequences, the one named as “MEF2DG4”, locating between -169 and -140 bp relative to the TSS on the template strand is of particular interest. First, it is comprised of four consecutive GGGG repeats, and therefore this region of the *MEF2D* upstream promoter may adopt a very stable four-tetrad G4 and compete strongly with adjacent duplexes. Secondly, the MEF2DG4 locates very close to the TSS, and potentially could have a strong influence on downstream gene transcription once formed. Furthermore, 2 bp downstream to this G4-forming sequence there is the sole E-box in the proximal promoter of human *MEF2D*.

As compared to the wide expression patterns of MEF2 family, human TnIc is the only sarcomeric protein found to be exclusively expressed in the cardiac myocyte, which makes it an ideal model to investigate cardiac-specific gene expression and in turn to provide new approaches to the construction of cardiac-specific vectors in gene therapy. Although TnIc has been regarded as a preferred biomarker in MI diagnosis for more than a decade [17, 18], the molecular mechanisms of human *TnIc* transcription regulation are still poorly understood, and only a few TFs have been found to be involved in regulating the transcription of *TnIc* [19-21]. This is partially due to the difficulties in obtaining suitable cell models, and possibly due to the extremely enriched G4-forming sequences in its promoter. TRR of *TnIc* is highly G4-important according to the bioinformatics analysis (*CF* 92.6 %, and *CQ* 95.2 %). Multiple G4-forming sequences have been identified in the proximal promoter of human *TnIc*, and two types of G4-forming motifs were selected for detailed studies, which are designated as “MNSG4” (between -528 and -319 bp upstream to the TSS), and “-80G4” (centred at -80bp to the TSS). The MNSG4 motif is composed of 6 repeats of chromosome 19 specific G-rich MNS, which are highly conserved in primates. Although conserved chromosome 19 specific MNS repeats are also found in neuronally expressed Shc adaptor homolog *SCK1/SLI* [22], coding region of an EFG-like module-containing mucin-like receptor, *EMR3* [23], the third intro of human apolipoprotein C-II gene [24], and the first intron of human *TnT* [25], the actual functions of this widely distributed G4-forming sequence are still unclear. Thus, biophysical characterisations of the MNSG4 motif in the human *TnIc* promoter may provide us enlightenment in understanding the molecular mechanisms and the potential biological functions of these MNS repeats in other genes. On the other hand, similar as the KIT1 G4-forming sequence, the -80G4 is found to be unique in the

human genome [26]. More importantly, multiple putative or proven TF binding sites are clustered around the -80G4-forming sequences, such as a CACC-box, an Sp1 binding site, an E-box, and a GATA biding site [20]. Thus, the biophysical studies of this -80G4 are necessary to understand its potential biological roles in regulating downstream gene expression.

Biophysical characterisations investigate the formation, structural diversity, and thermodynamic/kinetic properties of G4s. Due to the more compact structure of a G4 as compared to a randomly coiled oligo, the formation of a particular G4 can be characterised by its faster mobility in EMSA experiments as compared to a control oligo with the same length. This simple technique opened up G4 studies as Henderson *et al.* first observed a fast migrating band that corresponds to human telomeric G4 in 1987 [27]. CD spectroscopy, an empirical approach, is also a useful method to probe the formation of G4s, and preliminarily investigate the structural diversity of G4s (i.e. parallel, and anti-parallel conformations). As described in Chapter 1, by comparing CD spectra to corresponding X-ray or NMR-resolved G4 structures, few types of distinct spectra have been well correlated with certain G4 structures [28-32]. Generally, positive signals at 265 nm and 295 nm indicate the existence of parallel and anti-parallel G4s respectively. By resolving guanines involved in Hoogsteen hydrogen bond formation in G4s, DMS footprinting is routinely used to demonstrate formation (VEGF G4 [33], HIF-1α G4 [34]) and loop arrangements (KRAS G4 [35]) of G4s, and sometimes even to derive their folding patterns (MYC G4 [36]). Thermodynamic parameters of G4 folding/unfolding are usually measured by traditional bulk methods, such as UV melting (reviewed in [37]), CD melting, (Variable Temperature-CD, VT-CD) [38], and so on. Single molecule techniques are recently emerging as a group of sophisticated methods to study the formation and structural properties of G4s. Due to their much higher spatial-temporal sensitivity and signal-to-noise ratio, single molecule techniques can record signals from individual molecules, and thus reclaim information about conformational heterogeneity, and thermodynamic/kinetic properties of certain G4 subpopulations, which are likely to be averaged when measured by bulk experiments. For example, Ying *et al.* revealed a parallel and an anti-parallel telomeric G4 structures, and investigated their interchanges by using a combination of smFRET technique and molecular modelling approach [4]. By using single molecule force spectroscopy, Yu *et al.* demonstrated the

coexistence of two stable G4s (parallel and anti-parallel) formed by the predominant ILPR sequence [39]. On the other hand, the kinetic parameters of G4 unfolding are normally determined by FRET signal changes at bulk or single molecule level, in which unfolded G4s are trapped by excess complimentary C-rich DNA strand [4, 5], or PNA strand [6]. Although very sensitive, results obtained by FRET experiments are sometimes dependent on the labelling sites of fluorophores, which may alter the folding pattern and stability of G4s.

In this chapter, biophysical characterisations of the MEF2DG4, and TnIc single/two-repeat of MNSG4 and -80G4 are described. First, the formation of G4s was examined by EMSA and CD spectroscopy experiments. Next, the conformational heterogeneity of individual G4 was investigated by combination of CD and smFRET spectroscopy. The thermodynamic and kinetic parameters of the unfolding of G4s were also determined by CD melting and FRET experiments. Finally, the interactions between TnIc G4s and a G4-binding ligand, a di-copper complex, were described.

4.2 Formation of G4s observed by EMSA

EMSA was first used to probe the formation of MEF2D and TnIc G4s. For each G4-forming sequence, its C-rich complimentary strand with the same length was used as a control. Although C-rich strands are able to adopt i-motifs, the quick annealing step and slightly basic pH (pH 8.0) in TBE buffer in my experiments eliminated the possibility of formation of i-motifs. All DNA oligos were incubated at 5 μM concentration in 10 mM Tris-HCl (pH 7.4) buffer containing 100 mM KCl, and diluted to 100 nM before loading to the gels.

For MEF2DG4, Cy3 was internally labeled in the middle of the second loop (CGG^{Cy3}AGCC, oligo **MEF-I**) to assess the formation of G4 in the presence of a fluorophore, and a Cy5 internally-labeled C-rich complimentary strand was used as a control (oligo **MEF-III**). As shown in Figure 4.1 a, on 12 % native PAGE, the band corresponding to oligo **MEF-I** (lane 2) migrates much faster than the band representing oligo **MEF-III** (lane 1). In contrast, on a 12 % denaturing PAGE containing 7 M urea these two bands have similar mobility (lane 3 and 4 correspond to **MEF-III** and **MEF-I** respectively). These results indicate that oligo **MEF-I** can adopt a much more compact structure(s) than **MEF-III** under native conditions (10 mM Tris-HCl buffer containing 100 mM KCl). Therefore, the fast migrating band of oligo **MEF-I** was regarded as evidence of the formation of G4(s). Meanwhile, for the fluorophore-labeled oligo **MEF-I**, no slow migrating band was observed on native gels when the oligo was incubated at 5 μM concentration, which indicates that the MEF2DG4-forming sequence would not adopt intermolecular G4.

EMSA experiments were also carried out to probe the formation of G4s in the unlabeled MNS and -80 G4-forming sequences from the promoter of human TnIc. As shown in Figure 4.1 b, on a native gel the band corresponding to a single repeat of MNSG4 (**Tr$_{MNS}$-I**, lane 2) migrates much faster than its C-rich complimentary strand (**Tr$_{MNS}$-II**, lane 1), suggesting the formation of intramolecular G4(s). In lane 2, no bands with slow mobility were observed, which indicates that at 5 μM oligo concentration, intramolecular G4 is the major form of MNSG4. Both oligos exhibit similar mobility on a control denaturing gel (lane 3 and 4 correspond to oligo **Tr$_{MNS}$-II** and **Tr$_{MNS}$-I** respectively). Results for the -80G4 were slightly different (Figure 4.1 c), as both fast and slow migrating bands were found in the -80G4 G-rich strand

(oligo **Tr$_{-80}$-I**, lane 2) on a native gel, while on a denaturing gel, no mobility difference was observed between the G-rich and C-rich strand (oligos **Tr$_{-80}$-II** and **Tr$_{-80}$-I** correspond to lane 3 and 4, respectively). This indicates that the G-rich strand of -80G4 can adopt a mixture of intramolecular (the major form, ~ 60 % according to band intensity) and intermolecular G4s at 5 μM concentration.

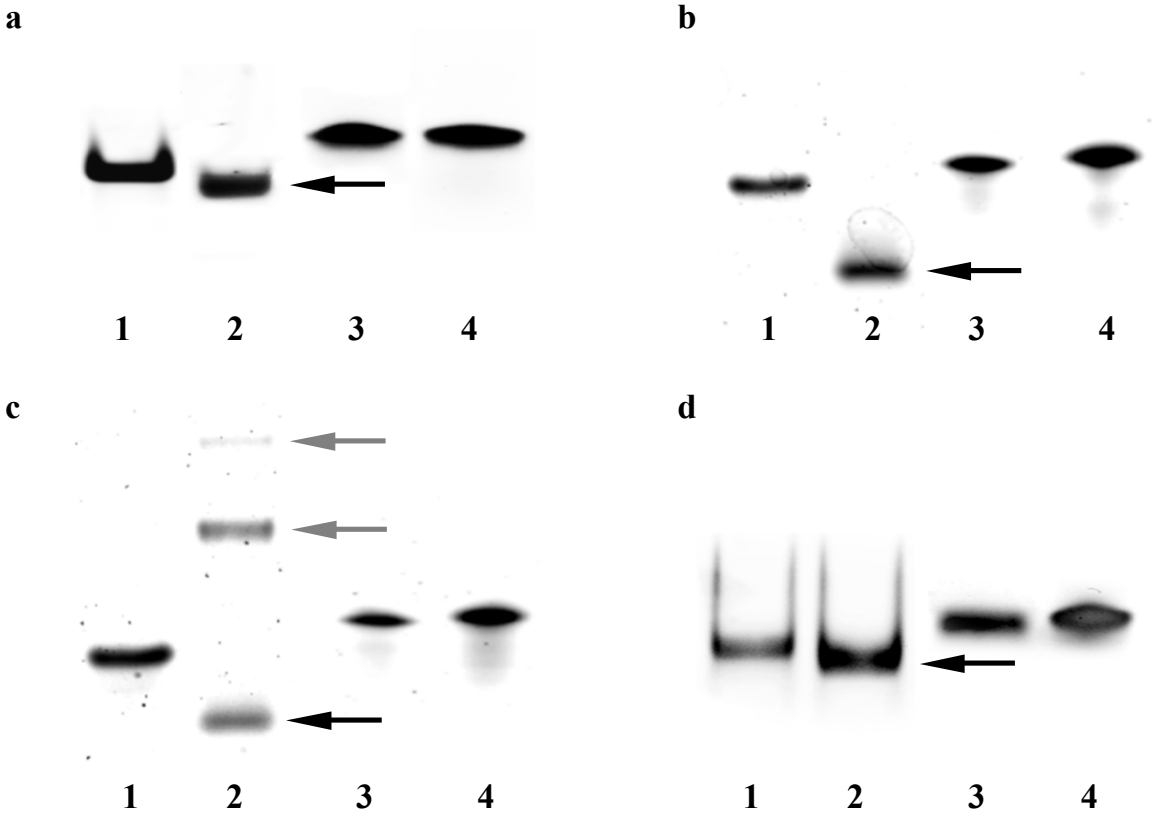


Figure 4.1: Native and denaturing gel electrophoresis indicating the formation of G4s.

(a) MEF2DG4: Lane1 and 2 are **MEF-III** and **MEF-I** (*MEF2D* promoter G4 C-rich and G-rich oligos, respectively) run in a 12 % native PAGE containing 100 mM KCl. Lane 3 and 4 are **MEF-III** and **MEF-I** run in a 12 % denaturing PAGE containing 7 M urea as controls. **(b)** TnIc MNSG4, 1 repeat: Lane1 and 2 are **Tr$_{MNS}$-II** (TnIc MNSG4 C-rich oligo) and **Tr$_{MNS}$-I** (TnIc MNSG4 G-rich oligo) run in a native gel. Lane 3 and 4 are oligos (**Tr$_{MNS}$-II** and **Tr$_{MNS}$-I**, respectively) run in a denaturing gel as controls. **(c)** TnIc -80G4: Lane1 and 2 are **Tr$_{-80}$-II** (-80G4 C-rich oligo) and **Tr$_{-80}$-I** (-80G4 G-rich oligo) run in a native gel. Lane 3 and 4 are oligos run in a denaturing gel as controls. **(b)** TnIc two-repeat MNSG4: Lane1 and 2 are **Tr$_{2\times MNS}$-II** (two-repeat MNSG4 C-rich oligo) and **Tr$_{2\times MNS}$-I** (two-repeat MNSG4 G-rich oligo) run in a native gel. Lane 3 and 4 are oligos run in a denaturing gel as controls. Formation of intramolecular G4s is proved by species with fast mobility under native conditions (indicated by black arrows) but not under denaturing conditions. Intermolecular G4s formed by the -80G4 forming sequence are indicated by gray arrows in (c).

Since multiple repeats of G4-forming MNS were found in the promoter of human *TnIc*, formation of G4(s) on the oligo containing two-repeat MNSG4 was also investigated by using EMSA experiments. In Figure 4.1 d, G-rich oligo containing two-repeat MNSG4 (oligo **$Tr_{2\times MNS}$-I**, lane 2) migrates faster than its C-rich complimentary strand (oligo **$Tr_{2\times MNS}$-II**, lane 1) in a native gel, while in a denaturing gel, both of them exhibited similar mobility (lane 3 and 4, oligos **$Tr_{2\times MNS}$-II** and **$Tr_{2\times MNS}$-I**, respectively). However, it is still unclear whether two G4s coexist on the same oligo or not.

To further investigate the formation, conformational heterogeneity, and thermal stability of these G4s, CD spectroscopy and CD melting experiments were subsequently carried out.

4.3 Formation and thermal stability of G4s determined by CD spectroscopy

Comparative CD spectra analysis was used to provide primary evidence showing the formation and general conformational heterogeneity of both MEF2D and TnIc G4s in solutions containing 100 mM different salts. CD melting experiments were used to investigate the thermal stability of these G4s.

Figure 4.2 a shows that the MEF2DG4-forming sequences (oligo **MEF-V**) incubated in 10 mM Tris-HCl (pH 7.4) buffer containing either 100 mM KCl (black) or LiCl (blue) both exhibited a CD spectrum characterised by a positive peak at 263 nm, a negative peak at 240 nm, and a pronounced positive shoulder between 280 nm and 310 nm. This spectral pattern is similar to that of the hybrid parallel/anti-parallel G4 structure of BCL-2 Pu39WT, whose structure has been solved previously by NMR [31], and that of a mixture of parallel and anti-parallel G4s of 31-mer ILPR [39]. Thus, these data suggest that the oligo **MEF-V** may form hybrid parallel and anti-parallel G4(s), or a mixture of pure parallel and anti-parallel G4s. In comparison to Pu48-mer, the 4-tetrad G4 forming sequence critical to PDGF-A expression in the promoter of *PDGF-A*, the G4 here seems to be easier to form under less favourable conditions. This suggests that MEFF2DG4 might be thermally more stable than the PDGF-A G4. As shown in Figure 4.2 a, the positive ellipticity at 266 nm in oligo

MEF-V in 100 mM LiCl decreases only ~ 1.5 fold compared to that in 100 mM KCl, while the same peak in Pu48-mer decreases more than 2.5 fold when changing from 100 mM KCl to 100 mM NaCl (Fig. 2C in reference [40]). Indeed, as compared to the PDGF-A G4 ($T_{m\ PDGF\text{-}A}$ = 85 ± 2 °C, 50 mM K^+ [41]), this MEF2DG4 has a higher melting temperature as determined by subsequent CD melting experiments (Table 4.1). CD melting experiments proved the formation of a hybrid of thermally stable parallel/anti-parallel G4(s). As shown in Figure 4.2 b, the melting curves obtained at 263 nm and 295 nm decreased homogeneously with increasing temperature, and the two curves possessed similar melting temperatures (T_{m263} = 92.9 ± 1.7 °C and T_{m295} = 93.1 ± 3.7 °C as determined by Eq. 2.4 and 2.8, Table 4.1; T_{m263} and ΔH was also determined by fitting the first derivative of melting curve obtained at 263 nm, and no significant difference was observed. Figure 4.2 c). This melting profile is similar to that of a hybrid parallel/anti-parallel human telomeric G4 as observed elsewhere (Figure 8C, in reference [42]). The unfolding enthalpy and entropy of MEF2DG4 were determined as 22.5 ± 0.4 kcal mol^{-1} and 60.7 ± 3.2 cal mol^{-1} K^{-1}, respectively (Table 4.1).

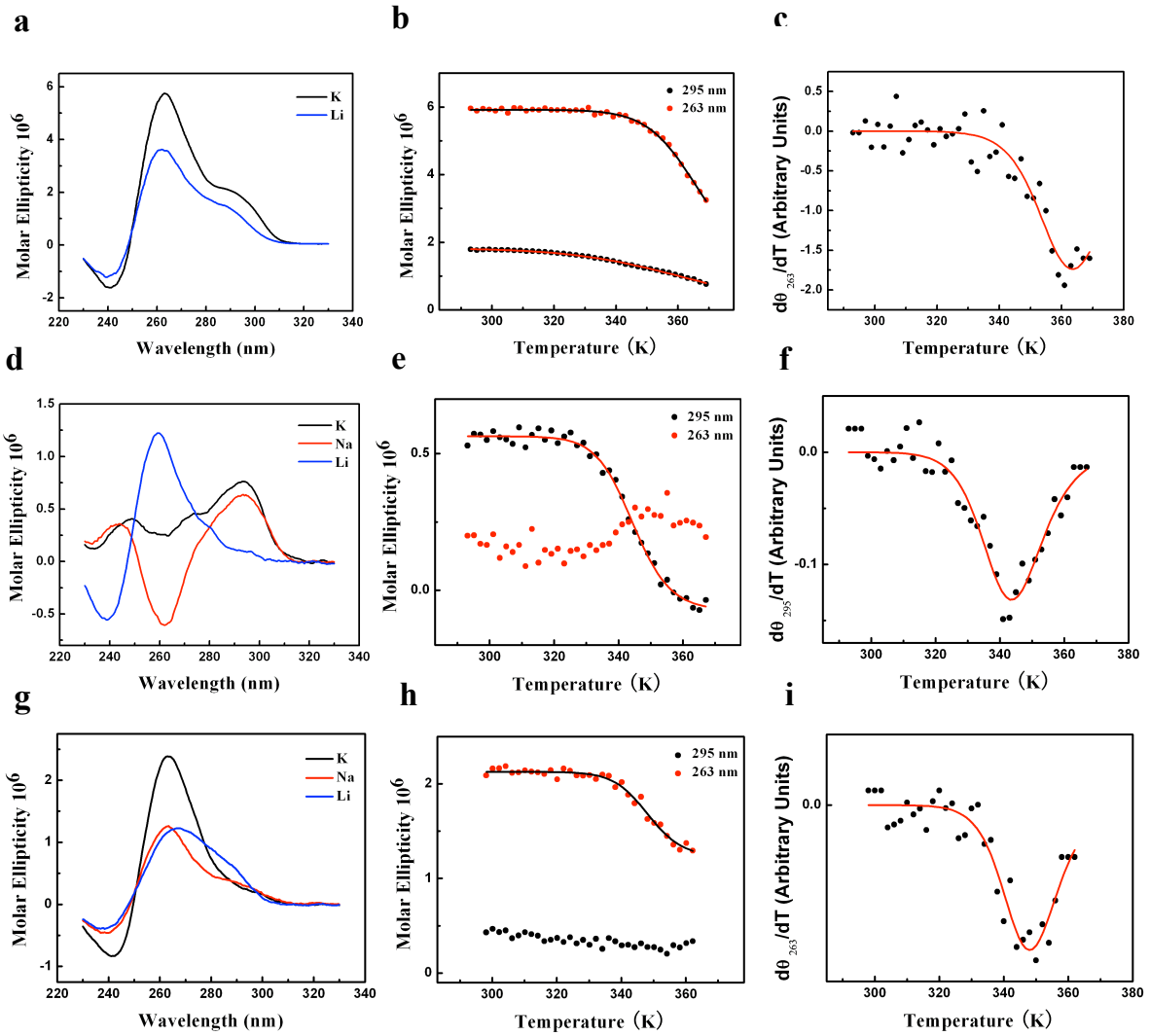
Figure 4.2 d shows the CD spectra of a single repeat of MNSG4 (oligo **Tr_{MNS}-I**) incubated in 10 mM Tris-HCl (pH 7.4) containing 100 mM K^+, Na^+, or Li^+. In the presence of LiCl, the CD spectrum of MNSG4 exhibits a positive peak at 262 nm and a negative trough at 240 nm. This spectrum is consistent with the CD signatures from parallel MYC Pu27mer G4, an HIV aptamer T30695, and a heptad-tetrad parallel structure determined by NMR [28, 43]. In 100 mM NaCl and KCl, the CD spectra give a minor positive absorption at around 250 nm, a maximum absorption peak at 295 nm, and two valleys at 230 and 260 nm. These spectra are in accordance with that from a thrombin biding aptamer (TBA) [28, 44], whose anti-parallel chair-type G4 structure has been determined by X-ray [45] and NMR studies [46]. Thus, it is very likely that in KCl solution, the majority of single repeat of MNSG4 adopts a similar folding pattern as the TBA and forms an anti-parallel G4 (or G4s). However, as compared to the CD spectra of TBA [28, 44] and MNSG4 measured in Na^+ which are characterised by a very negative valley at around 263 nm, this valley of MNSG4 CD spectrum in K^+ is positive, indicating that a small amount of parallel G4 may form in the presence of K^+. The melting curve at 263 nm, corresponding to possible parallel structures, only showed a slight increase at around 60 °C (Figure 4.2 e, red dots).

Although a similar phenomenon has been reported previously, and was explained as an accumulation of unstructured ssDNA by Yu *et al.* (Figure 4a, [39]), the possibility of parallel G4 formation still cannot be fully ruled out. The decrease of CD signals at 263 nm induced by the unfolding of parallel G4 might be compensated by the increase of CD signals at this wavelength because of the unfolding of chair-type anti-parallel G4. By fitting the melting curve obtained at 295 nm, the melting temperature and unfolding enthalpy and entropy of the chair-type anti-parallel MNSG4 were derived from Eq. 2.4 and 2.8 to be 71.4 ± 0.5 °C, 40.7 ± 2.6 kcal mol^{-1} and 118.1 cal mol^{-1} K^{-1} respectively (Figure 4.2 e, Table 4.1). Similar thermodynamic parameters were recalculated by fitting the first derivatives of the melting curve at 295 nm (Figure 4.2 f**)**, and no differences were identified.

Figure 4.2 g shows the CD spectra for the -80G4 (oligo **Tr_{-80}-I**). In 100 mM LiCl solution, the spectrum is characterised by a negative trough at 240 nm, a positive peak at 262 nm, and a pronounced positive shoulder between 280 and 300 nm. This is very similar to the spectrum of BCL-2 Pu39WT with a hybrid parallel/anti-parallel G4 structure solved by NMR [31]. In both 100mM NaCl and KCl solution, the spectra only showed a positive peak at 262 nm and a negative trough at 240 nm, a signature of parallel G4 structures as mentioned above. In addition, the higher CD signal at 260 nm obtained in KCl solution indicates that K^+ facilitates the formation of G4 as expected. Based on the CD melting experiments, there might be two or more parallel G4s in 100 mM K^+ solution. As shown in Figure 4.2 h and i, the parallel G4 with lower thermal stability, which accounts for 57.3 % of CD signals at 263 nm, was almost fully denatured at 90 °C, while no denaturing was observed for the other parallel species with high thermal stability. Since it has been shown by EMSA experiments that about 40 % of the -80G4 is able to adopt an intermolecular G4 conformation at 5 μM concentration, it is likely that the undenatured parallel species corresponds to intermolecular parallel G4s.

Melting profiles of the oligo containing two-repeat MNSG4 (oligo **$Tr_{2\times MNS}$-I**) were also preliminarily investigated, and a complicated melting behaviour was observed. Generally, strong signals at both 263 nm and 290 nm indicated that the existence of parallel and anti-parallel structures is very likely in both K^+ and Na^+ solution, while the K^+ cation promotes formation of the parallel one as shown in Figure 4.2 j. Melting experiments further revealed the cooperative unfolding processes of the parallel and

anti-parallel G4s (Figure 4.2 k and l). Basically, a parallel G4 with T_m = 46.7 ± 0.1 °C started to unfold first, while afterwards, an anti-parallel G4 with T_m = 68.3 ± 0.1 °C began to unfold, and finally, unfolding of another parallel G4 with T_m = 89.2 ± 0.4 °C was observed. As shown by the first derivative curves of melting profile (Figure 4.2 l), the negatively related melting processes of parallel and anti-parallel species are clearly demonstrated. Thus, according to these cooperative melting processes, it is very likely that both parallel and anti-parallel G4s coexist on the same strand. Thermodynamic parameters of individual G4 species are listed in Table 4.1, which were determined according to Eq. 2.4 ~ 2.8.



(Figure 4.2 to be continued on the next page)

(continued Figure 4.2)

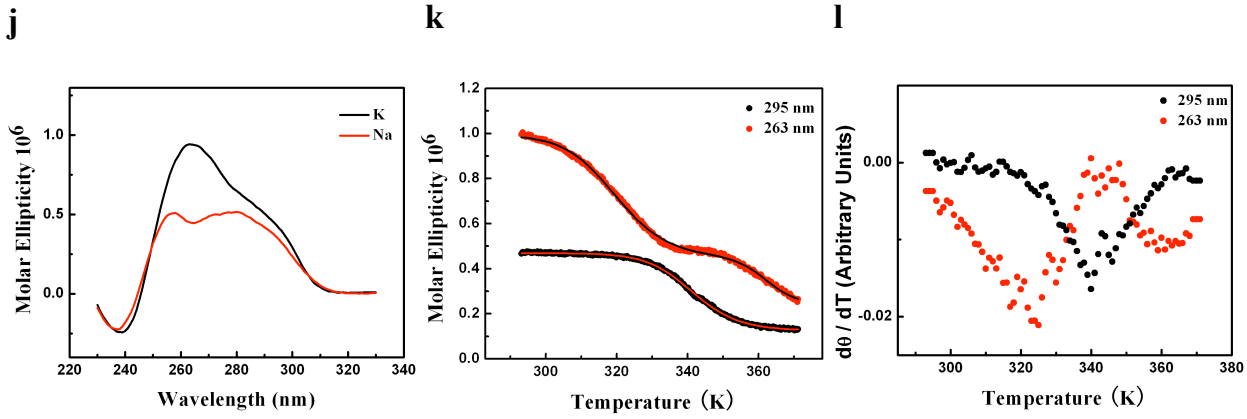


Figure 4.2: CD spectra and CD melting results of MEF2DG4 and TnIc G4s.

(a) CD spectra of oligo **MEF-V** in the presence of 100 mM K^+ (black curve) or Li^+ (blue curve) at room temperature. **(b)** CD signal changes at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, both of which were fitted by Eq. 2.4 and 2.8 (indicated by red line and black line respectively). **(c)** First derivative of CD melting curve at 263 nm (black dots), which were fitted by a model as described previously (Eq. 2.9) [47]. **(d)** CD spectra of oligo **Tr_{MNS}-I** in the presence of 100 mM K^+ (black curve), Na^+ (red curve) or Li^+ (blue curve) at room temperature. **(e)** CD signal changes at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, signal changes at 295 nm wavelength were fitted by Eq. 2.4 and 2.8 (red line). **(f)** First derivative of CD melting curve at 295 nm (black dots), which were fitted by a model as described previously. **(g)** CD spectra of oligo **Tr_{-80}-I** in the presence of 100 mM K^+ (black curve), Na^+ (red curve) or Li^+ (blue curve) at room temperature. **(h)** CD signal changes at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, signal changes at 263 nm wavelength were fitted by Eq. 2.4 and 2.8 (black line). **(i)** First derivative of CD melting curve at 263 nm (black dots), which were fitted by a model as described previously. **(j)** CD spectra of oligo **$Tr_{2\times MNS}$-I** in the presence of 100 mM K^+ (black curve) or Na^+ (red curve) at room temperature. **(k)** CD signal changes at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, both of which were fitted by Eq. 2.4 and 2.8 (indicated by red line and black line respectively). **(l)** First derivative of CD melting curve at 295 nm (black dots), and 263 nm (red dots).

All experiments were carried out in 10 mM Tris-HCl buffer (pH 7.4).

Table 4.1: Energetics of MEF2DG4 and TnIc MNSG4 and -80G4 unfolding in solutions determined by CD melting.

Oligo Names	Wavelength / nm	T_m / °C	ΔH / kcal mol^{-1}	ΔS / cal mol^{-1} K^{-1}	$\Delta G_{(310\ K)}$ / kcal mol^{-1}
MEF-V (MEF2DG4)	263	92.9 ± 1.7	33.0 ± 1.8	90.2 ± 4.9	5.1 ± 0.4
	295	93.1 ± 3.7	13.1 ± 0.7	35.8 ± 1.9	2.0 ± 0.1
	Overall	97.6 ± 0.2	22.5 ± 0.4	60.7 ± 3.2	3.7 ± 0.2
Tr_{MNS}-I (TnIc MNSG4)	263	N/A	N/A	N/A	N/A
	295	71.4 ± 0.5	40.7 ± 2.6	118.1 ± 7.5	4.1 ± 0.3
Tr_{-80}-I (TnIc -80G4)	263	75.2 ± 1.0	44.8 ± 5.2	128.6 ± 14.9	4.9 ± 0.4
	295	N/A	N/A	N/A	N/A
$Tr_{2\times MNS}$-I (TnIc two-repeat MNSG4)	263	46.7 ± 0.1	27.7 ± 0.4	86.6 ± 1.3	0.9 ± 0.1
		89.2 ± 0.4	60.0 ± 3.4	165.6 ± 9.4	8.7 ± 0.4
	295	68.3 ± 0.1	35.7 ± 0.2	104.6 ± 0.6	3.3 ± 0.2

Experiments were carried out in solutions containing 10 mM Tris-HCl (pH 7.4) and 100 mM K^+.

4.4 Conformational analysis G4 by smFRET

4.4.1 MEF2DG4*

The internal and end dual-labeled G4 systems composed of oligos **MEF-I/IV** and **MEF-II/IV** are shown in Figure 4.3 a and b respectively. The G4-forming sequence is connected to a 30 bp duplex on its 5' end. In both internal and end dual-labeled constructs, the acceptor, Cy5, is labeled to an adenine via an NHS ester 6 bp from the G4. In the internal dual-labeled system, the donor, Cy3, is labeled on a thymine in the third side loop of G4 (Figure 4.3 a), and in the end-labeled system, Cy3 is labeled on the last guanine involved in G4 (Figure 4.3 b). In the folded states of both internal and end dual-labeled systems, the donor-acceptor distances are within estimated Förster radius of Cy3/Cy5 pair, 5.3 nm [49]. In the presence of either **MEF-III** or **VI**, if G4 hybridises into duplex state it would lead to a large decrease of FRET efficiency from ~ 70 % to less than 15 % for the internal dual-labeled system, and negligible (< 1 %) for the end dual-labeled system as a result of the increased separation between donor and accepter fluorophores in the duplex state (Figure 4.3 c and d).

* Part of this Section was adopted from my previous publication [48].

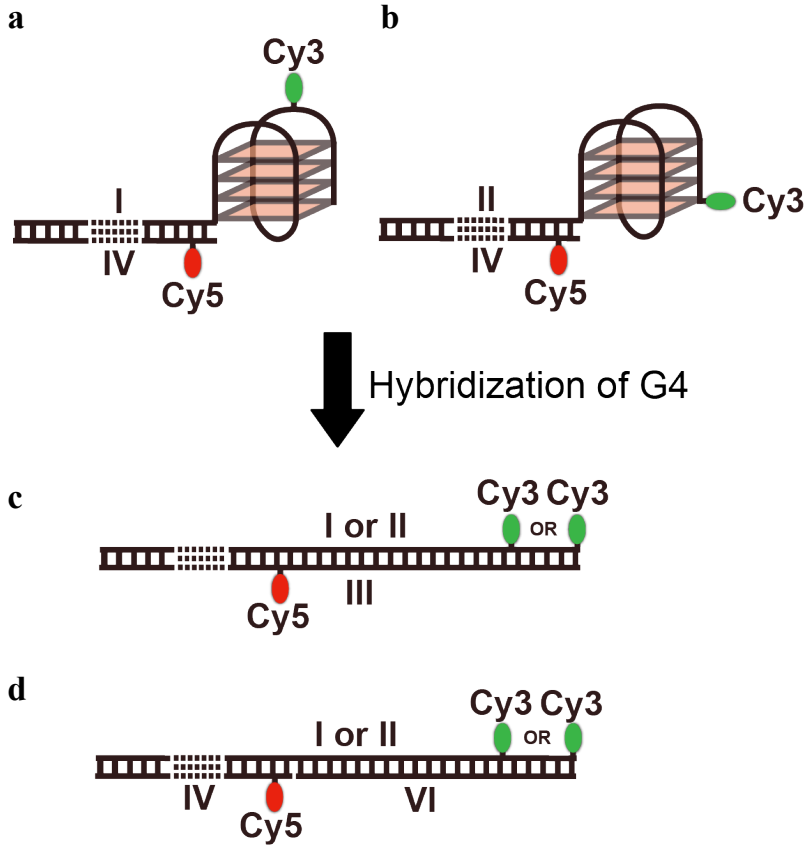


Figure 4.3: Schematic illustration of the dual-labeled MEF2DG4 systems for single molecule and bulk FRET studies.
(a) Internal dual-labeled G4 system formed between oligos **MEF-I** and **IV**. **(b)** end dual-labeled G4 system formed between oligos **MEF-II** and **IV**. **(c)** duplex formed in the presence of oligo **MEF-III**. **(d)** duplex formed in the presence of oligos **MEF-IV** and **VI**. These systems were designed to show high FRET in folded states (a and b), and low or no FRET in duplex states (c and d).

As the formation of G4 is dependent on metal ion coordination, initial smFRET studies were performed by using the Cy3/Cy5 dual-labeled G4 systems to probe the formation of G4 in different salt solutions. Figure 4.4 a ~ d show the smFRET histograms of the internal dual-labeled system (oligos **MEF-I** and **IV**) obtained at 25 °C in solutions without salt and containing 100 mM Li^+, Na^+, and K^+ respectively. In these histograms, four subpopulations were identified by global fittings. **S1** around zero is due to inactive acceptor [4] or unfolded DNA. **S2**, the broadest peak centered about 0.37, is likely the ensemble of partially folded intermediates. **S3** and **S4**, with their FRET efficiencies centered on 0.66 and 0.87, respectively, are assigned to two distinct G4 conformations, both of which are likely to adopt hybrid parallel/anti-parallel conformations. Low FRET **S1** and **S2** dominate in the histogram obtained in

the solution without salt (Figure 4.4 a). In the absence of cations, the repulsion of central oxygen atoms would overwhelm the strength of Hoogsteen bond at N7 position of guanines involved in maintaining the overall stability of G4 [50], while the coordination of cations between G-tetrads can reduce this repulsion and hence stabilise overall G4 structures. A general trend among alkali ions from the most stable to least is as follows: $K^+ > Na^+ > Li^+ >$ no salt. Thus, without central metal cations bound to the tetrads, unfolded and partially folded states would be dominant. In 100 mM Li^+ solution, most of the unfolded states previously observed in the absence of salt (Figure 4.4 a, **S1**) were folded into high FRET G4s, which result in a smaller **S1** and larger **S3**, **S4** (Figure 4.4 b). Although the stabilising effect of lithium cation is very weak, it can still promote G4 structures, as evidenced by an increase of high FRET species in Figure 4.4 b. Histograms obtained in 100 mM Na^+ and K^+ solutions (Figure 4.4 c and d) were identical, with fewer **S2** and more **S4**. The larger increase of **S4** over **S3** is probably due to the different selectivity of the cation to the specific folded states.

To learn the effect of labeling position on the folding and stability of G4, similar smFRET experiments were also carried out on the end dual-labeled G4 system. The results are shown in Figure 4.4 e ~ h. Global fitting did not work in this case as different G4s formed under different salt conditions. Again, **S1** is assigned to the inactive Cy5 or unfolded state, while **S2** is an ensemble of partially folded intermediates as described above. Without salt, only one high FRET subpopulation can be identified and **S1** predominates (Figure 4.4 e). In comparison to that of internal dual-labeled system in the absence of salt (Figure 4.4 a), **S2** population in the end dual-labeled system (Figure 4.4 e) is very small while a much larger and slightly shifted **S1** ("zero" peak) is observed, which is possibly caused by further separation between donor and acceptor in this system. The partially folded subpopulation may be merged into the "zero" peak and is difficult to resolve by fitting alone. Histograms obtained from Li^+ and Na^+ solutions are almost identical (Figure 4.4 f and g), indicating Li^+ and Na^+ may have similar effects on the structure and stability of these G4s. FRET histogram obtained in K^+ solution differs substantially from those obtained in either Na^+ or Li^+, suggesting that K^+ favours the formation of a particular G4 conformation in **S3**.

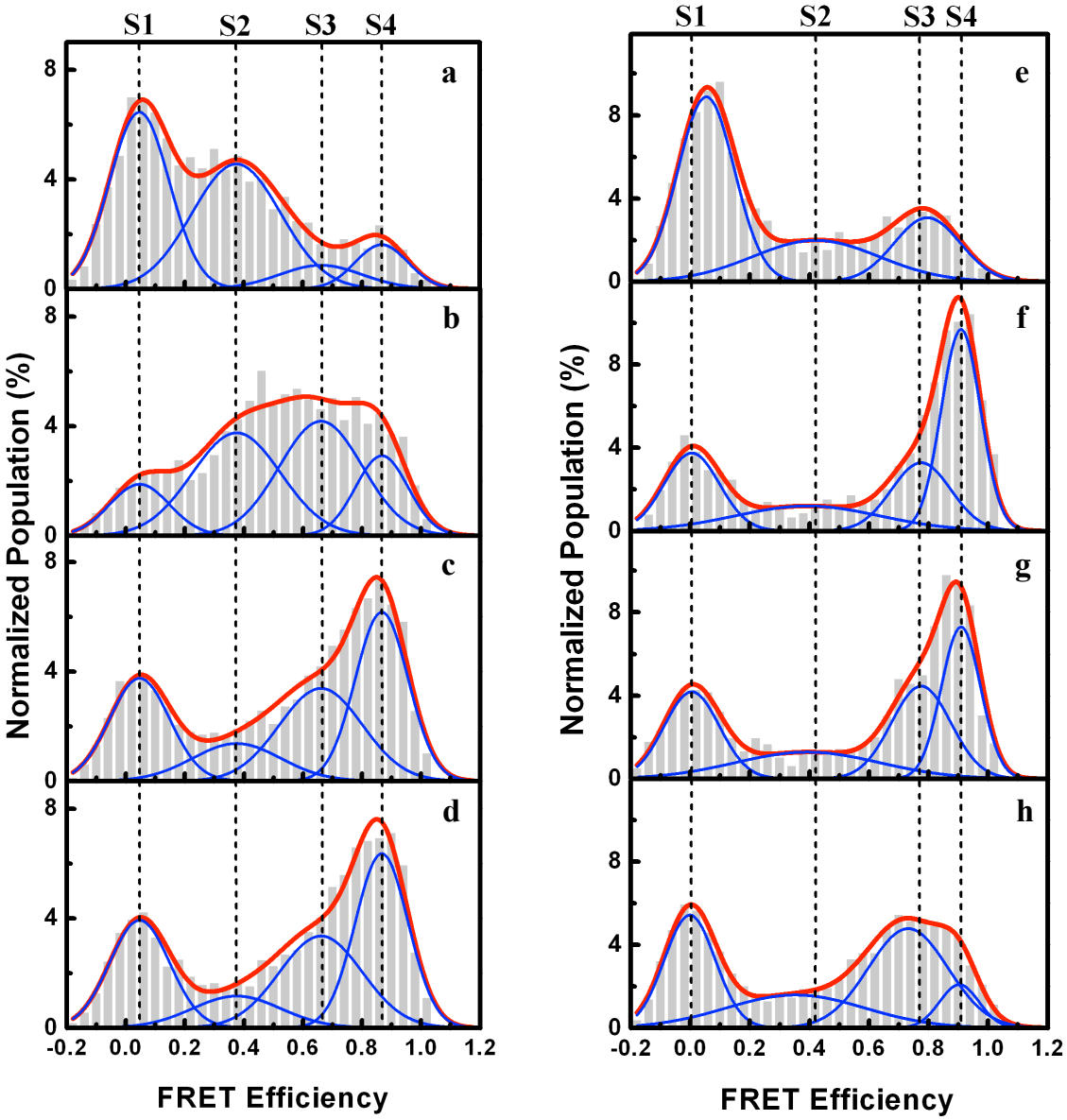


Figure 4.4: Single molecule FRET histograms in MEF2DG4 conformational analysis.

(a) ~ **(d)** Histograms in smFRET conformational analysis using the internal dual-labeled system (oligos **MEF-I** and **IV**) obtained in the absence of salt, 100 mM Li^+, Na^+, and K^+ solutions respectively. **(e)** ~ **(h)** Histograms in smFRET conformational analysis using end dual-labeled system (oligos **MEF-II** and **IV**) obtained in the absence of salt, 100 mM Li^+, Na^+, and K^+ solutions respectively. Experiments were carried out at 25 °C. The donor, Cy3, was exited by a 150 μW collimated laser beam at 514 nm. All solutions contain 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % (φ) Tween 20. Blue lines represent individual subpopulations fitted by a multi-peak Gaussian distribution model, and red lines are the sum of subpopulations.

4.4.2 TnIc MNSG4 and -80G4

As illustrated in Figure 4.5 a, double-stranded (ds) dual-labeled G4 systems were designed to probe the formation and conformational heterogeneity of MNSG4 and -80G4 by smFRET experiments. For both MNSG4 and -80G4, one protecting nucleotide is added onto the 5' of the G4-forming sequence (thymine for MNSG4, and cytosine for -80G4; oligo **Tr_{MNS}-III** and **Tr_{-80}-III,** Table 2.1), on which the donor, Cy3, is labeled, and a 35-mer artificial linker is attached to the 3' end of the G4-forming sequence. After annealing to the Cy5-labeled strand complimentary to the linker region (oligo **Comp-Cy5**), the acceptor is 8 nucleotides from the 3' of the G4-forming sequence. One nucleotide gap is left between the G4-forming sequence and the annealed duplex linker to enable free gyration of the system, and recognition of the C-rich complimentary strand. Similar to the end dual-labeled MEF2DG4 system, once fully hybridised by the C-rich strand (oligo **Tr_{MNS}-II** or **Tr_{-80}-II**), the FRET efficiency of this ds dual-labeled system decreased from approximately 70 % (G4 folded states) to less than 1 % due to the separation of the donor and acceptor on the fully hybridised duplex.

By using this ds dual-labeled system (oligos **Tr_{MNS}-III** and **Comp-Cy5**), the cation-dependent conformational heterogeneity of one repeat of MNSG4 was first investigated by smFRET experiments. As shown in Figure 4.6 a, no high FRET species were observed in the absence of cations. In the smFRET histogram, the **S1** and **S2**, centred at 0 and 0.11, respectively, correspond to the zero peak and unfolded states of MNSG4, while the minor population of **S3** is likely from FRET signals of randomly coiled oligos, or simply fluorescent background. In the presence of 100 mM Na^+, a dramatic decrease of unfolded state (**S2**) and the emergence of a high FRET subpopulation centred around 0.66 (**S3**) were observed (Figure 4.6 b). Furthermore, a small population with very high FRET efficiency (**S4**, centred at 0.94) was also revealed by global fittings to three independent experiments. When incubated in 100 mM K^+ (Figure 4.6 c), the smFRET histogram is characterised by a further reduced zero peak (**S1**), decreased unfolded state (**S2**), and enlarged high FRET subpopulations (**S3**, and **S4**, centred at 0.80 and 0.94 respectively), which reflects that the majority of the oligo is able to adopt compact G4 structures in the presence of K^+. Due to its high FRET efficiency, the subpopulation **S4** observed in both Na^+ and K^+ solutions were designated to a chair-type anti-parallel G4 (will be discussed later),

while the **S3** is likely a parallel structure as proposed in CD experiments (Figure 4.2 d and e).

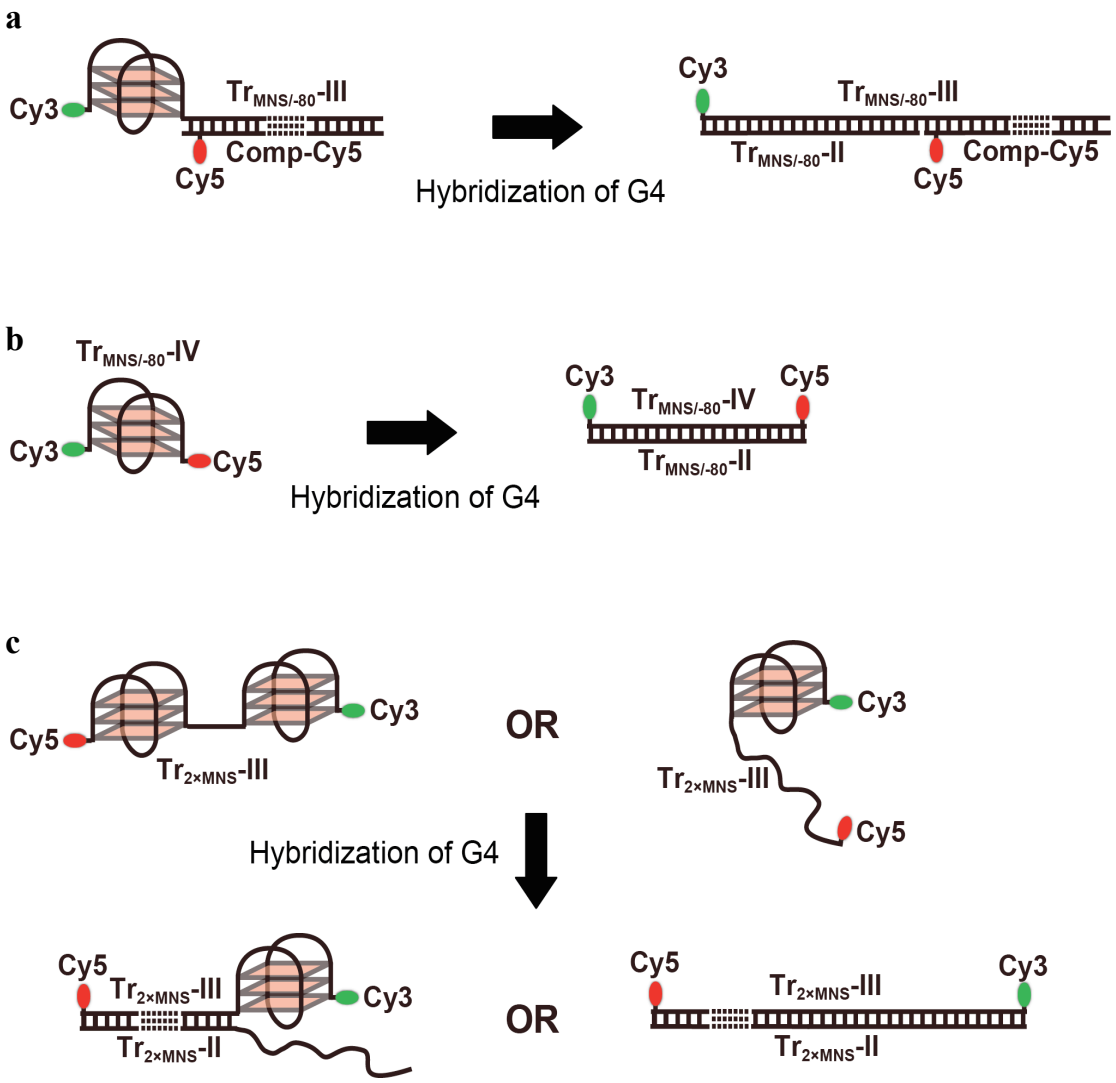


Figure 4.5: Schematic illustration of the dual-labeled TnIc G4s systems for single molecule and bulk FRET studies.

(a) Double-stranded (ds) dual-labeled G4 system formed between oligos **Tr$_{MNS}$-III** or **Tr$_{-80}$-III** and **Comp-Cy5**, which can be fully hybridised by **Tr$_{MNS}$-II** or **Tr$_{-80}$-II**, respectively. **(b)** Single-stranded (ss) dual-labeled G4 system formed within oligos **Tr$_{MNS}$-IV** or **Tr$_{-80}$-IV.** The presence of **Tr$_{MNS}$-II** or **Tr$_{-80}$-II** can fully unfold the G4 structures through hybridisation. In cases as illustrated in **a** and **b**, folding of G4 would lead to high FRET signals, while hybridisation would eliminate FRET signals. **(c)** Ss dual-labeled G4 system containing double repeats of MNSG4 (oligo **Tr$_{2×MNS}$-III**). In this system, formation of two or one G4 (double/single G4-folded states) within the oligo would show considerable FRET signals (due to free gyration of single-stranded region), while hybridisation of duplex leads to low or no FRET signals.

Cation-dependent conformational heterogeneity of the TnIc -80G4 was also examined by using a similar ds dual-labeled system (oligos **Tr$_{-80}$-III** and **Comp-Cy5**). Similarly, without the stabilising effects from cation ions, G4 structure cannot fold (Figure 4.7 a). In the presence of 100 mM Na^+ and K^+, however, significant amount of high FRET subpopulation was observed (**S3** in Figure 4.7 b, and **S4** in Figure 4.7 c), suggesting the formation of G4 structures. It is worth noting that as in the CD experiments, only parallel G4 was found in Na^+ and K^+ solutions for the -80G4-forming sequence, and the propeller-type structure is the only folding pattern for an intramolecular parallel G4. Thus, in smFRET histograms, only one high FRET subpopulation was resolved in both Na^+ and K^+ solutions. As compared to in the presence of Na^+, the larger zero peak (**S1**) and the smaller high FRET subpopulation (**S4**) observed in 100 mM K^+ implies that the actual conformations of -80G4 are different in Na^+ and K^+ solutions, possibly due to different relative orientations between the G4 and duplex linker, or different loop arrangements. Furthermore, a larger and broader peak (**S2**) corresponding to partially unfolded structures was observed in Na^+ solution, reflecting less favourable G4-stabilising ability of Na^+ as compared to K^+.

To further explore the conformational state(s) of TnIc MNSG4 and -80G4, a simplified model, single-stranded (ss) dual-labeled G4 system was designed. As shown in Figure 4.5 b, in the ss dual-labeled G4 system, one protecting nucleotide was added to both 5' and 3' ends of G4-forming sequences of MNSG4 and -80G4, which were labeled by Cy3 and Cy5, respectively. Folding of G4 structures would bring donor and acceptor into close proximity and hence give high FRET signals, while full hybridisation would result in a decrease in FRET efficiency to less than 5 %. However, as shown in Figure 4.6 d ~ f and Figure 4.7 d ~ f, only one skewed high FRET peak was observed in each histogram, which is caused by over-juxtaposition of the donor and acceptor in the folded states. Although most of the conformational information was lost in condensed high FRET signals, this experiment still yielded an important conclusion, namely that although in the absence of cation ions, both MNSG4 and -80G4 cannot form in the ds dual-labeled systems (Figure 4.6 a, and Figure 4.7 a), high FRET subpopulation corresponding to folded G4 structures were observed in the ss dual-labeled systems under similar conditions (Figure 4.6 d, and

Figure 4.7 d), suggesting that adjacent duplex DNA in ds dual-labeled systems may interfere with the formation of both MNSG4 and -80G4 in solution.

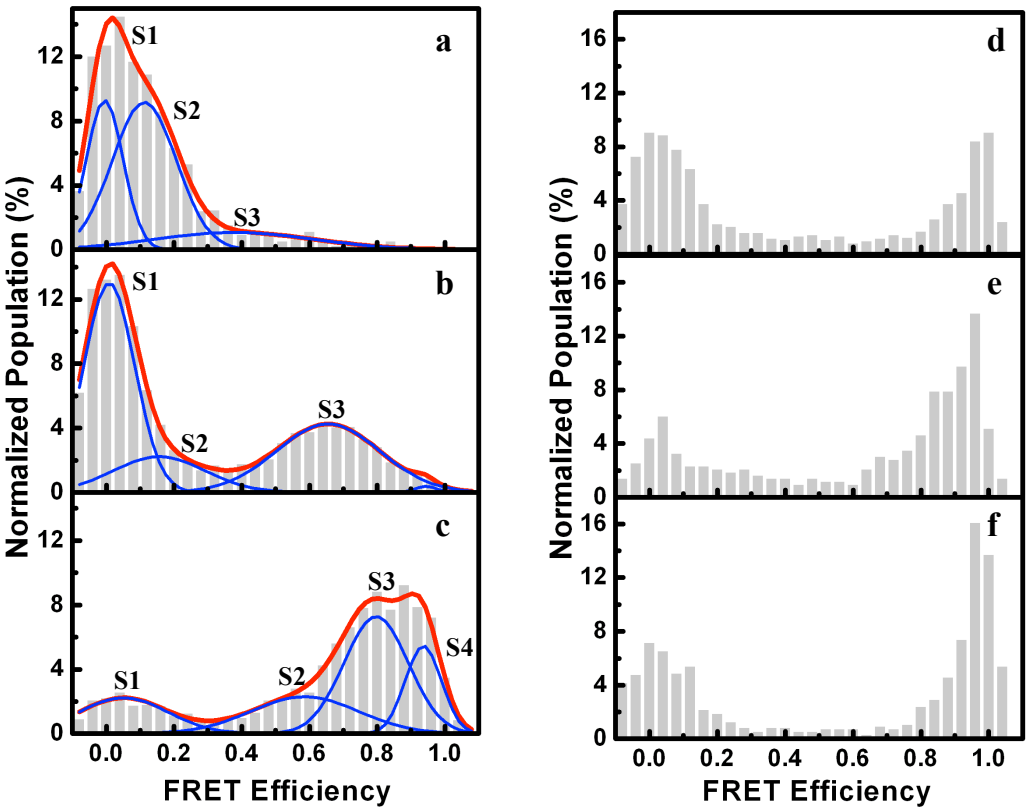


Figure 4.6: Single molecule FRET histograms in TnIc MNSG4 conformational analysis.

(a)~(c) Histograms in smFRET conformational analysis using ds dual-labeled system (oligos **Tr_{MNS}-III** and **Comp-Cy5**) obtained in the absence of salt, 100 mM Na^+, and K^+ solutions respectively. **(d)~(f)** Histograms in smFRET conformational analysis using ss dual-labeled system (oligo **Tr_{MNS}-IV** alone) obtained in the absence of salt, 100 mM Na^+, and K^+ solutions respectively. Experiments were carried out at 25 °C. The donor, Cy3, was exited by a 150 μW collimated laser beam at 514 nm. All solutions contain 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % (φ) Tween 20. Blue lines represent individual subpopulations fitted by a multi-peak Gaussian distribution model, and red lines are the sum of subpopulations.

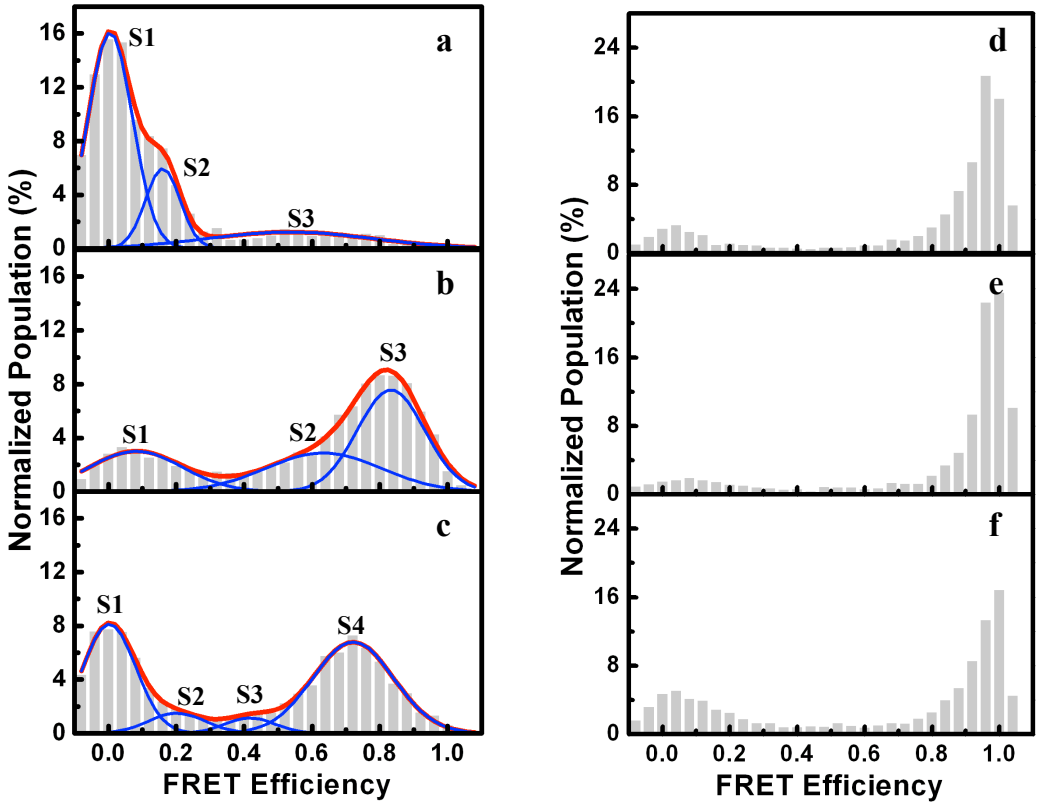


Figure 4.7: Single molecule FRET histograms in TnIc -80G4 conformational analysis.

(a)~(c) Histograms in smFRET conformational analysis using ds dual-labeled system (oligos **Tr$_{-80}$-III** and **Comp-Cy5**) obtained in the absence of salt, 100 mM Na^+, and K^+ solutions respectively. **(d)~(f)** Histograms in smFRET conformational analysis using ss dual-labeled system (oligo **Tr$_{-80}$-IV** alone) obtained in the absence of salt, 100 mM Na^+, and K^+ solutions respectively. Experiments were carried out at 25 °C. The donor, Cy3, was exited by a 150 μW collimated laser beam at 514 nm. All solutions contain 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % (φ) Tween 20. Blue lines represent individual subpopulations fitted by a multi-peak Gaussian distribution model, and red lines are the sum of subpopulations.

4.4.3 TnIc two-repeat MNSG4

The ss dual-labeled two-repeat MNSG4 system was designed as shown in Figure 4.5 c. Cy5 and Cy3 are labeled to the first protecting thymine and the last protecting cytosine via NHS esters, respectively (oligo **Tr$_{2\times MNS}$-III**). Two 26mer consensus MNSG4-forming sequences are linked by an 11mer linker in the middle. In double G4-folded states, the donor-acceptor distances are within estimated Förster radius of the Cy3/Cy5 pair, 5.3 nm [49]. In the single G4-folded states, a certain amount of

FRET signals might be possible due to the free gyration of the single-stranded part. However, once annealed to its complimentary strand (oligo **$Tr_{2\times MNS}$-II**), the FRET efficiency of both partial duplex and full duplex would decrease to less than 1 % due to the increased separation between donor and acceptor fluorophores.

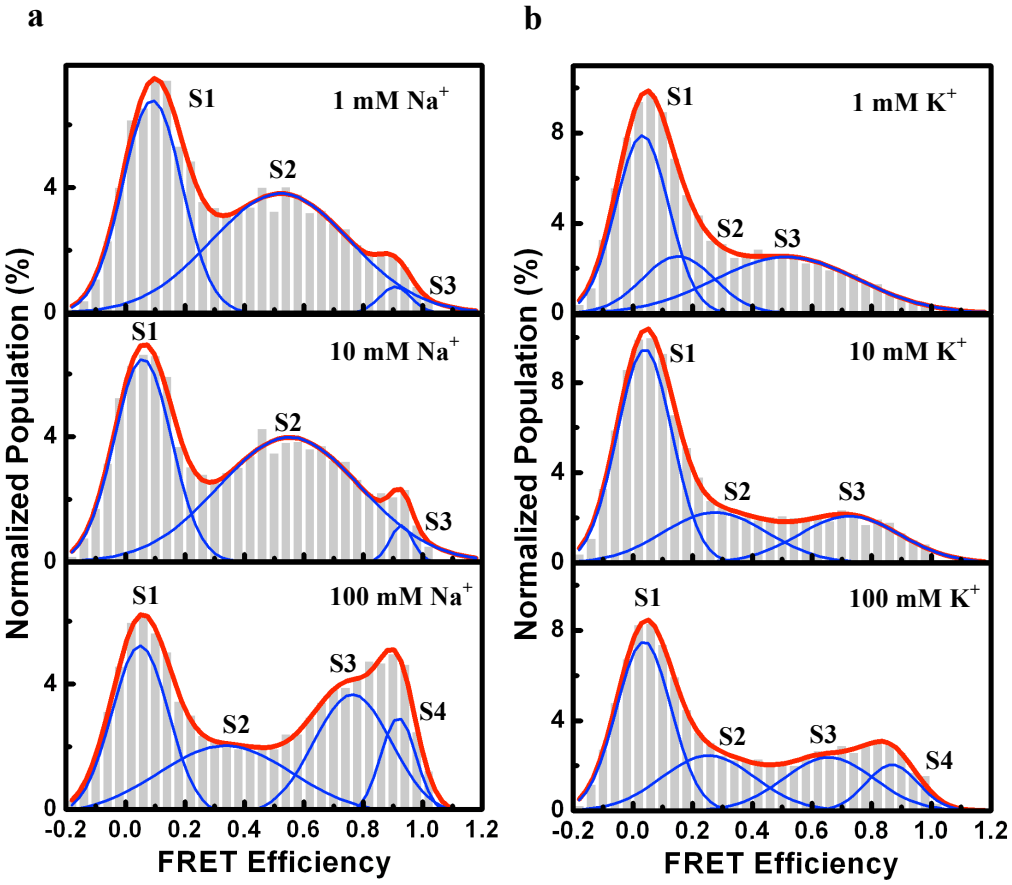


Figure 4.8: Single molecule FRET histograms in TnIc two-repeat MNSG4 conformational analysis.

(a) Histograms in smFRET conformational analysis using ss dual-labeled system (oligo **$Tr_{2\times MNS}$-III**) obtained in 1 mM, 10 mM, and 100 mM Na^+ solutions. **(b)** Analogous histograms obtained in 1 mM, 10 mM, and 100 mM K^+ solutions respectively. Experiments were carried out at 25 °C. The donor, Cy3, was exited by a 150 μW collimated laser beam at 514 nm. All solutions contain 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % (φ) Tween 20. Blue lines represent individual subpopulations fitted by a multi-peak Gaussian distribution model, and red lines are the sum of subpopulations.

Conformational analysis was performed in Na^+ and K^+ solutions with different concentrations (Figure 4.8). Although the two-repeat MNSG4 system cannot adopt folded structure(s) in the absence of cations, high FRET subpopulation (**S3**, centred at 0.91, Figure 4.8 a**,** top) was found in solution with low concentration of Na^+ (1 mM). When the Na^+ concentration was increased from 1 mM to 10 mM, this high FRET

subpopulation became more apparent, while the subpopulation corresponding to randomly coiled oligos or partially folded structures, **S2**, gradually compressed. In 100 mM Na^+, two high FRET subpopulations emerged (**S3** and **S4,** Figure 4.8 a, bottom). According to the donor-acceptor distances of subpopulations **S3** and **S4** (43.7 ± 1.5 Å and 35.2 ± 1.1 Å), they both represent the formation of two G4 structures on the same strand. Similarly, high FRET subpopulations accumulate as the K^+ concentration increases (Figure 4.8 b). Obscurely, as compared to subpopulations identified in solution containing 100 mM Na^+, a larger faction of zero peak and a smaller fraction of subpopulation with high FRET signals were identified. This may relate to the local environment differences such as G4-G4 stacking, or unfavourable fluorophores' dipole-dipole orientation.

4.5 G4/duplex competition assay using smFRET

4.5.1 MEF2DG4

To probe the kinetic stability of these G4 structures in the presence of their C-rich complementary strands, G4/duplex competition assays were carried out using smFRET in 100 mM K^+ solution. The Cy3 internal and end dual-labeled G-rich strands (oligos **MEF-I** and **II**) were incubated with 1.5-fold of either **MEF-III** (2-component, Figure 4.3 c), or **MEF-IV/VI** (3-component, Figure 4.3 d) Cy5-labeled C-rich complementary strands to allow the formation of full duplex. Comparing with the 3-component system, the 2-component system is better at unfolding G4, because the formation of duplex in the 5' flanking region of the G-rich strand would also bring the C-rich strand close to the G4, which would increase the local concentration of the C-rich strand.

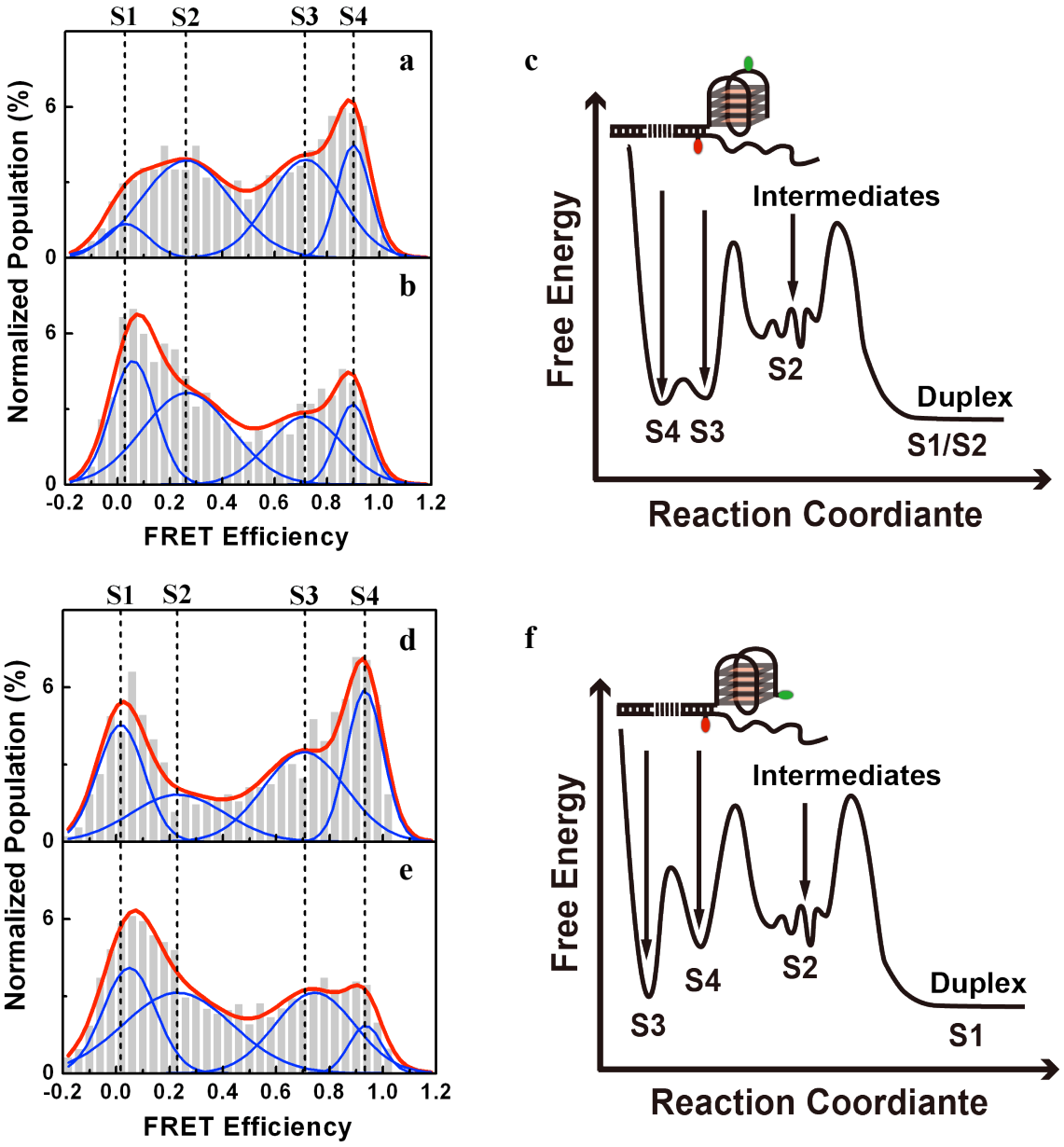


Figure 4.9: Single molecule FRET histograms in MEF2D G4/duplex competition assays and the proposed free energy landscapes.

Histograms obtained in smFRET MEF2D G4/duplex competition assay of the internal dual-labeled system composed of **(a)** 3-component (oligos **MEF-I**, **IV**, and **VI**) and **(b)** 2-component (oligos **MEF-I**, and **III**). **(c)** Proposed free energy landscapes illustrating unfolding of internal dual-labeled MEF2DG4. Histograms obtained in smFRET MEF2D G4/duplex competition assay of the end dual-labeled system composed of **(d)** 3-component (oligos **MEF-II**, **IV**, and **VI**) and **(e)** 2-component (oligos **MEF-II**, and **III**). **(f)** Proposed free energy landscapes illustrating unfolding of end dual-labeled MEF2DG4.. All experiments were carried out at 25 °C in solutions containing 100 mM K^+, 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % (v/v) Tween 20; **S3**, **S4** are two G4 structures, **S2** are mainly intermediates between G4 and duplex and **S1** is the full duplex. Blue lines represent individual subpopulations fitted by a multi-peak Gaussian distribution model, and red lines are the sum of subpopulations.

In the internal dual-labeled model, the presence of C-rich strands leads to an increase of **S2**, which are intermediates with partially unfolded structures (comparing Figure 4.9 a and b to Figure 4.4 d). This is likely due to the partial unfolding of G4 structures by the incomplete hybridisation. In the 3-component system, both high FRET folded states **S3** and **S4** decreased, while the fraction of **S1** is very small (Figure 4.9 a). This demonstrates that the presence of the short C-rich complimentary strand (oligo **MEF-VI**) can only partially unfold the G4 structures. In the 2-component system, **S3** and **S4** further decreased, and **S1** dramatically increased (Figure 4.9 b), which indicates that some of the folded states **S3** and **S4** are unfolded by the long C-rich strand (oligo **MEF-III**) and subsequently form duplex. The simultaneous decrease of **S3** and **S4** in 3-component and 2-component system shows that these two folded states may have similar kinetic stability. The proposed energy landscape is shown in Figure 4.9 c**.**

In the end dual-labeled model, observations are different. In the 3-component system, a high FRET folded state **S4** comes back and the other subpopulations remain relatively constant (comparing Figure 4.9 d to Figure 4.4 h). In the 2-component system, the folded state **S4** partially unfolds into intermediate **S2**, while the other folded state **S3** still remains constant (Figure 4.9 e). Thus, a G4/duplex competition energy landscape of the end-labeled system can be proposed. As shown in Figure 4.9 f, on encounter of an adjacent C-rich complementary strand, the more stable G4 structure **S3** first rearranges to a less stable G4 structure **S4**, which further unfolds into an intermediate before ending up as a full duplex. In this model, the rate-limiting step is the unfolding of G4 into duplex.

4.5.2 TnIc MNSG4 and -80G4s

To examine the stability of TnIc G4s, similar G4/duplex competition assays were carried out by incubating the ds or ss dual-labeled G4 systems in the presence of 1.5-fold of corresponding C-rich complimentary strands (as illustrated in Figure 4.5). Histograms obtained in smFRET experiments showed that all high FRET subpopulations of TnIc MNSG4, -80G4 and two-repeat MNSG4 were depleted by C-rich complimentary strands. This agrees to the findings that the melting temperatures of both MNSG4 (T_m = 71.2 °C) and -80G4 (T_m = 75.0 °C) are lower as compared to that of the MEF2DG4 (T_m > 85 °C). In the cooling process to allow G4 formation, it

is possible that duplexes already well-formed before the temperature reaches low enough for oligos to adopt stable G4 structures. On the other hand, although the melting temperature of the parallel G4 adopted by the two-repeat MNSG4 is relatively high (T_m = 87 °C), the 11mer linker between two MNSG4s can provide an attaching site for the C-rich complimentary strand even when two G4s are formed on the G-rich oligo. This partially annealed system could then bring the C-rich strand close to the G4 structures (increased local concentration of C-rich oligo), and accelerate the unfolding process of MNSG4.

To further study the unfolding kinetics of both the MEF2DG4 and TnIc G4s, the G4 hybridisation experiments were subsequently carried out by using bulk and single molecule FRET approaches.

4.6 G4 hybridisation and unfolding kinetics

4.6.1 MEF2DG4*

In light of the energy landscape for the unfolding of the end dual-labeled G4 system in K^+ solution, a kinetic scheme for the competition between G4 and duplex in the presence of C-rich complementary strands was proposed, as shown in Figure 4.10. In this model, **GQ** is the oligo **MEF-II** in G4 form, and C-Rich refers to the random-coiled complementary strand oligo **MEF-III**. When oligos **MEF-II** and **III** are incubated together, a partial duplex will quickly form in the 30-base long linker region. These G4/duplex hybrids would initially maintain the most stable (lowest free energy) G4 structure **GQC1**, corresponding to the **S3** species in Figure 4.4 h and Figure 4.9 d and e. Following the progress of the reaction, some **GQC1** would interconvert, via conformational rearrangement, into **GQC2**, a less stable G4 structure corresponding to the **S4** species in Figure 4.4 h and Figure 4.9 d and e. Afterwards, **GQC2** unfolds, leaving a partially-formed duplex (tagged as **P-Duplex**) that eventually anneals into a full duplex (tagged as **Duplex**).

* Part of this Section was adopted from my previous publication [48].

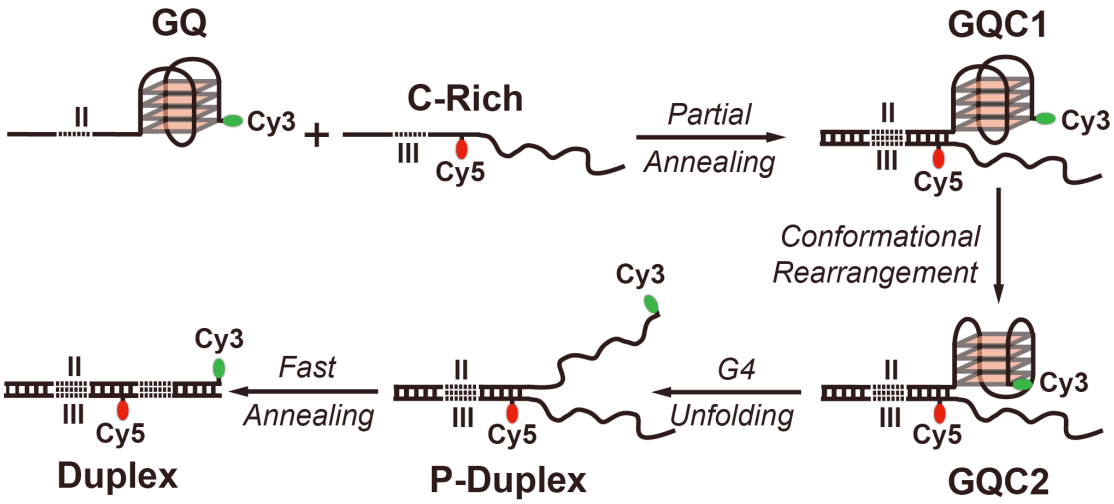


Figure 4.10: A model for the competition between MEF2DG4 and duplex in the presence of full length C-rich complimentary strand.
GQ is a G4 structure formed in the single stranded oligo **MEF-II**; **C-Rich** is the denatured single stranded oligo **MEF-III**; **GQC1** and **GQC2** are schematic representations of two G4 structure corresponding to **S3**, **S4** respectively; **P-Duplex** is a partially-formed duplex; and **Duplex** is the fully-formed duplex.

Based on the facts that, *i)* **GQC1** is more stable than **GQC2**; *ii)* the **GQC2** population is much smaller than **GQC1** in the presence of K^+ (Figure 4.4 h and Figure 4.9 e); and *iii)* the formation of the **P-Duplex** is very fast due to the existence of the adjacent complementary strand, the model proposed above can be simplified to,

$$\mathrm{GQ} + \mathrm{C} \xrightarrow{k_1} \mathrm{GQC} \xrightarrow{k_2} \mathrm{Duplex} \qquad \text{Eq. 4.1}$$

where k_1 and k_2 are the partial annealing rate (duplex hybridisation rate) and the G4 unfolding rate respectively. From this reaction scheme, it is not difficult to obtain the rate equation against species **GQC** as follows,

$$\frac{d[\mathrm{GQC}]}{dt} = k_1(c_0 - [\mathrm{GQC}])^2 - k_2[\mathrm{GQC}] \qquad \text{Eq. 4.2}$$

where c_0 is the initial concentration of both **GQ** and **C**. The analytical solution of Eq. 4.2 can be used to obtain the duplex hybridisation rate and the subsequent G4 unfolding rate from fitting to the experimental data, as shown in Figure 4.11. The fitting was carried out in MatLab 2009 by a built-in curve-fitting tool. The

hybridisation rate constant and unfolding rate constant were obtained to be 3.8 ± 0.2 × 10^4 $M^{-1}s^{-1}$ and 8.3 ± 0.4 × 10^{-6} s^{-1} respectively. In comparison to the hybridisation rate constant of 1.37 × 10^5 $M^{-1}s^{-1}$ between a 31-base G4-forming strand and its complementary from the human *MYC* promoter [51], the hybridisation rate constant measured here is slower but well within the typical range of DNA hybridisation. Remarkably, the unfolding rate constant of the G4 studied here is two to three magnitude slower than that of 3-tetrad G4s, with literature values 7.9 × 10^{-3} s^{-1} for human MYC G4 (150 mM K^+, 25 °C) [51] and 2.34 × 10^{-2} s^{-1} (100 mM Na^+, 25 °C) for human telomeric G4 [5], respectively. In addition to the results from single molecule duplex/G4 competition assay, the unfolding kinetics confirmed the exceptional stability of the four-tetrad G4, which might form easily in the promoter during gene transcription.

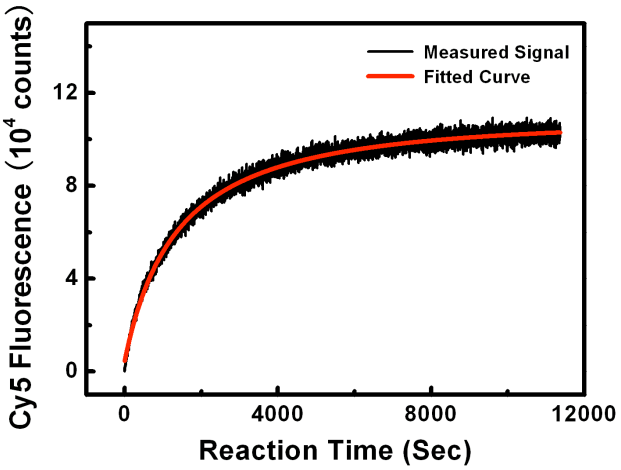


Figure 4.11: MEF2DG4 unfolding curve and corresponding fitted curve.
Hybridisation time trace recorded by acceptor Cy5 fluorescence (black) and the fit (red) based on the solution of Eq. 4.2. The hybridisation rate constant and unfolding rate constant were obtained to be 3.8 ± 0.2 × 10^4 $M^{-1}s^{-1}$ and 8.3 ± 0.4 × 10^{-6} s^{-1} respectively. The reaction was carried out at 25 °C in 100 mM KCl solution containing 10 mM Tris-HCl (pH 7.4).

4.6.2 TnIc MNSG4 and -80G4

For both TnIc MNSG4 and -80G4, solutions containing 100 nM G4-forming oligos (MNSG4: oligos **Tr_{MNS}-III** and **Comp-Cy5**; and -80G4: oligos **Tr_{-80}-III** and **Comp-Cy5**) were mixed with 2 μM C-rich complimentary strands (MNSG4: oligo **Tr_{MNS}-II**; and -80G4: **Tr_{-80}-II**) at time t = 0, and the decay of Cy5 fluorescence were measured. The resulting traces were fitted to double exponential curves, and the inverse of fitted

time constants were taken as the hybridisation or unfolding rate-constants. Temperature-dependent unfolding kinetics was examined by performing the hybridisation at different temperatures.

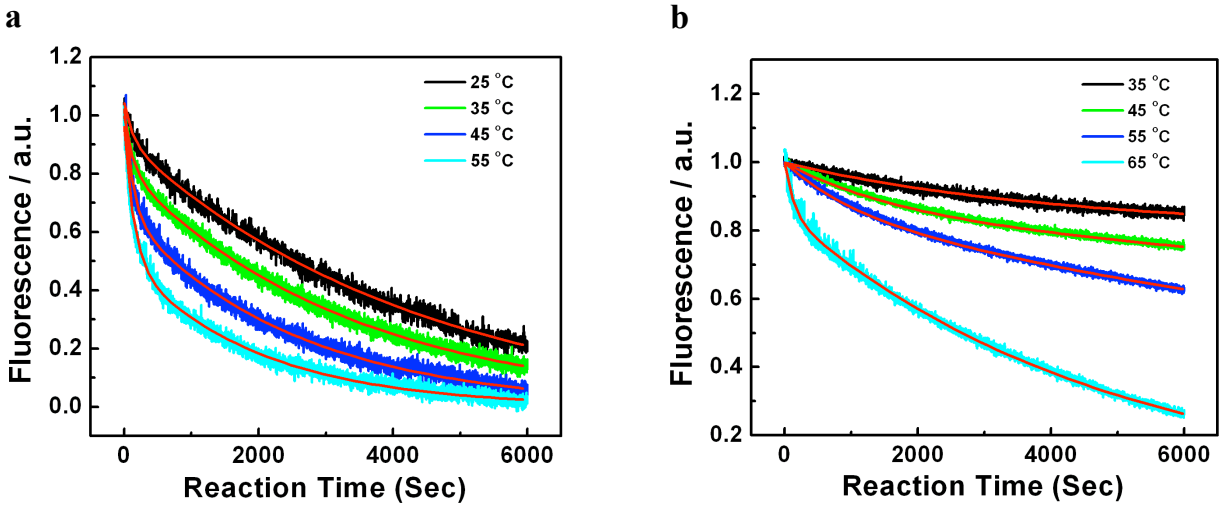


Figure 4.12: TnIc G4s unfolding curves and corresponding fitted curves.
(a) Decrease of Cy5 fluorescence recorded at 664 nm (5 nm split) in the process of MNSG4 (formed between oligos **Tr_{MNS}-III** and **Comp-Cy5**) opened by its C-rich complimentary strand (oligo **Tr_{MNS}-II**) at different temperature. Black, green, blue and cyan curves are from experimental data obtained at 25, 35, 45, and 55 °C, respectively, and red curves are fitted results by a double exponential decay model. **(b)** Decrease of Cy5 fluorescence recorded at 664 nm (5 nm split) in the process of -80G4 (formed between oligos **Tr_{-80}-III** and **Comp-Cy5**) opened by its C-rich complimentary strand (oligo **Tr_{-80}-II**) at different temperature. Black, green, blue and cyan curves are from experimental data obtained at 35, 45, 55, and 65 °C, respectively, and red curves are fitted results by a double exponential decay model. Hybridisation was performed between 100 nM G4 forming oligos and 2 μM complimentary C-rich oligo. The reaction was carried out in 100 mM KCl solution containing 10 mM Tris-HCl (pH 7.4).

The time constants of the fast decaying component for MNSG4 were determined as 126.7 ± 6.2, 128.5 ± 3.4, 127.4 ± 2.2, and 140.2 ± 1.6 s at 25, 35, 45, and 55 °C respectively (Figure 4.12 a and Table 4.2). The time constant of this fast decaying subpopulation maintained unchanged until 55 °C when prolonged time constant was observed. Thus, this fast hybridised subpopulation is believed to be partially folded or randomly coiled oligos with hybridisation rate constants around 0.4×10^4 $M^{-1}s^{-1}$. As compared to a typical hybridisation rate constant, which is around 10^5 $M^{-1}s^{-1}$, the much slower hybridisation rate constants observed here clearly indicated that the C-rich oligos still need to unfold certain unstable structures before fully hybridising the

partially unfolding G-rich strand. The slow decaying component represents the unfolding process of MNSG4, whose time constant decreases with increased temperature (Table 4.2). As compared to the KIT1 G4 studies that used a similar system [13], the MNSG4 is more kinetically stable at high temperature (time constants of KIT1 G4 at 45 and 50 °C are 2,300 and 1,000 s respectively), which echoes the higher thermal stability of MNSG4 in single-stranded form. The activation energy of the unfolding of MNSG4 was determined as 22.1 ± 0.4 kJ mol^{-1} by Arrhenius equation (Figure 4.13 a and Table 4.2), and the activation enthalpy and entropy were derived by Eyring equation as 19.6 ± 0.4 kJ mol^{-1}, and -249.2 ± 1.2 J mol^{-1} respectively. Similar as the unfolding of human telomeric G4 studied via an analogous strategy [4] and the unfolding of TBA studied in sodium solution [52], this small activation enthalpy and very negative entropy indicate that the unfolding of MNSG4 is entropically driven, and can occur very easily in the presence of C-rich complimentary strand. This relatively low activation energy/enthalpy and very negative entropy may suggest that the MNSG4 unfolds through a "pre-organized" transition state, which possibly relates to the long loops in the folded MNSG4 (two 5-nucleotide adjacent loops). In the process of MNSG4 opening, these fluctuating long loops may partially anneal to the C-rich strand, resulting in a series of relatively "ordered" intermediates with negative entropy, and the invasion of C-rich strand may further destabilise the folded G4 structures without transiting G4 into other partially unfolded structures (more disordered structures with positive entropies). Thus, in the absence of C-rich complimentary strand the MNSG4 can be thermally stable (relatively high melting temperature), while the presence of excess C-rich complimentary strand can open the folded MNSG4 relatively easily.

Similar experiments were carried out on the TnIc -80G4, and two G4s (fast and slow unfolding components) were revealed with different kinetic behaviours (Figure 4.13 b and Table 4.2). The fast-unfolding -80G4 (-80G4-F) exhibits comparable kinetic stability to the MNSG4, while the slow-unfolding -80G4 (-80G4-S) is at least 30-fold more stable than the -80G4-F, and of similar stability as the KIT2 G4 studied previously by using an identical method [13]. Both -80G4-F and -80G4-S were characterised by a large activation enthalpy and small unfavoured activation entropy, suggesting the existence of more significant enthalpic barriers ($\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ for -80G4-F and -80G4-S are 83.4 ± 23.6 kJ mol^{-1} and -44.1 ± 7.4 J mol^{-1}, and 98.0 ± 6.5

kJ mol^{-1} and -27.1 ± 20.0 J mol^{-1}, respectively). As compared to the MNSG4, only one long loop was found in the -80G4 (the first loop, length between 10 ~ 7 nucleotides), and correspondingly less negative activation entropies were identified in unfolding processes of the -80G4. As the -80G4-F has a more negative $\Delta S^{\ddagger}$ than the -80G4-S, the destabilising contribution from the loop (partially annealed more ordered states) of -80G4-F is more prominent than that from -80G4-S, and thus less energy is required for the -80G4-F to reach the transition state for the full hybridisation to happen. Indeed, a smaller $\Delta H^{\ddagger}$ of -80G4-F was also found.

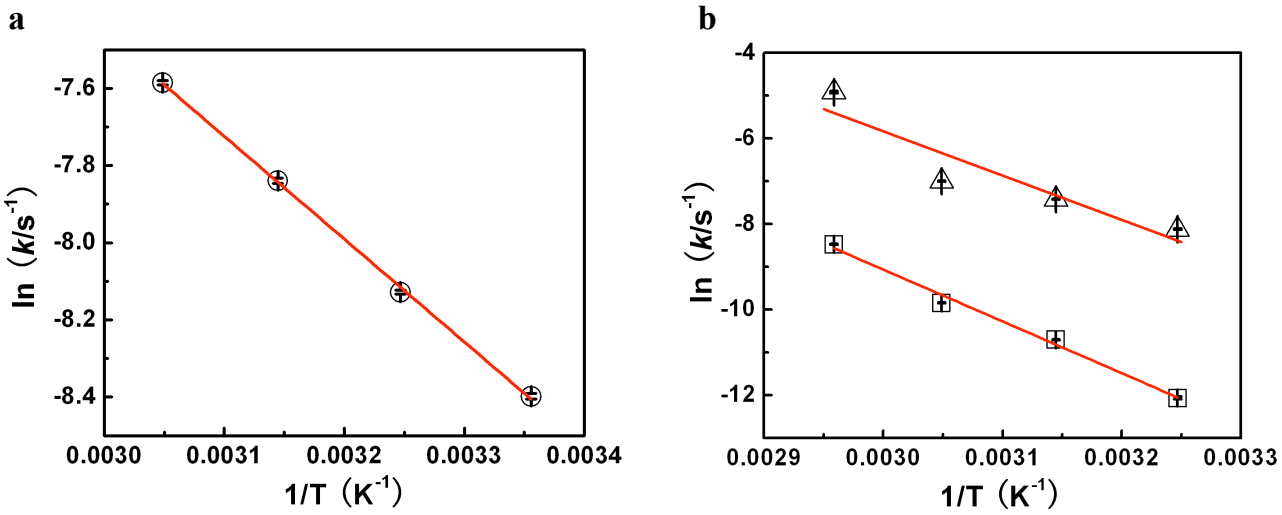


Figure 4.13: Arrhenius plotting for unfolding of TnIc G4s.
(a) Arrhenius plotting for the slow decaying component identified in the unfolding process of TnIc MNSG4. Empty circles (O) represent opening rates measured at different temperatures, and red line is the result fitted by Arrhenius equation. The activation energy is determined to be 22.1 ± 0.4 kJ mol^{-1}. **(b)** Arrhenius plotting for the slow and fast decaying components identified in the unfolding process of TnIc -80G4. Empty triangles (Δ) and squares (□) represent opening rates of the fast and slow decaying components, respectively, and red lines are results fitted by Arrhenius equation. The activation energies for the fast and slow decaying components are 86.1 ± 23.7 and 100.6 ± 6.57 kJ mol^{-1}, respectively, as determined by the fits.

Table 4.2: Parameters obtained from curve fittings of TnIc MNSG4 and -80G4 unfolding in solution.

Temperature / °C	$<\tau>$ / s	k / 10^{-5} s^{-1}	Ea / kJ mol^{-1}	$\Delta H^{\ddagger}$ / kJ mol^{-1}	$\Delta S^{\ddagger}$ / J mol^{-1}
TnIc MNSG4 (fast decaying component)					
25	126.7 ± 6.2	789.3 ± 38.6	--	--	--
35	128.5 ± 3.4	778.2 ± 20.6			
45	127.4 ± 2.2	784.9 ± 13.6			
55	140.2 ± 1.6	713.3 ± 8.1			
TnIc MNSG4 (slow decaying component)					
25	4,438.0 ± 32.1	22.5 ± 0.2	22.1 ± 0.4	19.6 ± 0.4	-249.2 ± 1.2
35	3,388.9 ± 16.3	29.5 ± 0.1			
45	2,538.9 ± 16.2	39.4 ± 0.3			
55	1,969.5 ± 11.1	50.8 ± 0.3			
TnIc -80G4-F (fast decaying component)					
35	3,366.6 ± 46.7	29.7 ± 0.4	86.1 ± 23.7	83.4 ± 23.6	-44.1 ± 7.4
45	1,669.2 ± 18.9	59.9 ± 0.7			
55	1,094.0 ± 13.6	91.4 ± 1.1			
65	136.5 ± 3.4	732.1 ± 18.4			
TnIc -80G4-S (slow decaying component)					
35	173,840.9 ± 5,542.6	0.6 ± 0.1	100.6 ± 6.6	98.0 ± 6.5	-27.1 ± 20.0
45	44,424.4 ± 482.3	2.3 ± 0.1			
55	18,811.2 ± 136.0	5.3 ± 0.1			
65	4,789.9 ± 32.8	20.9 ± 0.2			

G4 unfolding was measured by the decrease of Cy5 fluorescence when hybridised by C-rich complimentary strand. Experiments were carried out in solution containing 10 mM Tris-HCl (pH 7.4) and 100 mM K^+ at different temperatures.

4.6.3 TnIc two-repeat MNSG4

Unfolding kinetics of ss dual-labeled two-repeat MNSG4 trapped by duplex (as shown in Figure 4.5 c) was first measured at bulk level. Hybridisation of excess oligo **$Tr_{2\times MNS}$-II** (20-fold, 2 μM) to **$Tr_{2\times MNS}$-III** (100 nM) was performed in 10 mM Tris-HCl (pH 7.4) buffer containing 100 mM K^+ at 25 °C, and concomitant decrease of Cy5 fluorescence were monitored as unfolding of G4s. Figure 4.14 illustrates the unfolding curve of two-repeat MNSG4 (black curve) fitted with double exponential decay as a best fit (red curve). Similar to those found in the single repeat of MNSG4 unfolding kinetics studies, a fast decaying component with time constant of 69.2 ± 0.7 s was identified as representing partially unfolded or randomly coiled structures. The time constant of a slow decaying component corresponding to the averaged unfolding of MNSG4s is 1,673.9 ± 9.0 s. As compared to the single repeat of MNSG4, both

time constants for the two-repeat MNSG4 are smaller. This is due to the existence of an 11mer linker between individual MNSG4 in the system, which provides annealing sites for the C-rich complimentary strand, and further destabilises adjacent G4 structures.

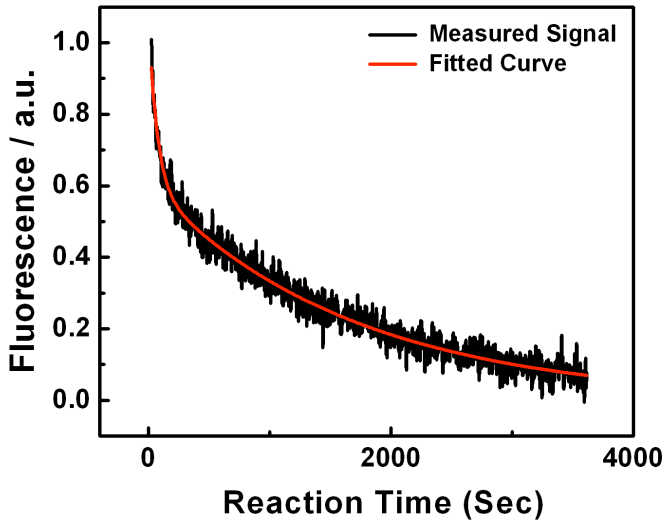


Figure 4.14: TnIc two-repeat MNSG4 unfolding curve and corresponding fitted curve.
Decrease of Cy5 fluorescence recorded at 664 nm (5 nm split) in the process of two-repeat MNSG4 (formed by oligo **$Tr_{2\times MNS}$-III**) opened by its C-rich complimentary strand (oligo **$Tr_{2\times MNS}$-II**) at 25 °C. The black curve is from experimental data, and the red curve is the fitted result by a double exponential decay model. Hybridisation was performed between 100 nM G4 forming oligos and 2 μM complimentary C-rich oligo. The reaction was carried out in 100 mM KCl solution containing 10 mM Tris-HCl (pH 7.4).

In previous G4 conformational analyses on the two-repeat MNSG4 system (Section 4.4.3), 3 different high FRET subpopulations (**S2**~**S4**) were identified in 100 mM K^+ (Figure 4.8 b). Thus, in order to investigate the kinetic stability of individual subpopulation, its unfolding kinetics was also measured as single molecule level. 2 nM oligo **$Tr_{2\times MNS}$-II** (20-fold as compared to the two-repeat MNSG4-forming sequence, oligo **$Tr_{2\times MNS}$-III**) was mixed with 100 pM oligo **$Tr_{2\times MNS}$-III** at time = 0, and the unfolding reaction was monitor for 7,900 s. smFRET histograms were generated every 160 s after the initiation of hybridisation, and globally fitted subsequently. As shown in Figure 4.15 a, four subpopulations (**S1** ~ **S4**) were revealed, and the fraction of high FRET ones (**S2** ~ **S4**) gradually decreased during the unfolding processes. Fractional changes of each subpopulation were plotted

against reaction time, and fitted with either double exponential (**S1** and **S3**) or single exponential decay (**S2** and **S4**) according to the best fit (Figure 4.15 b ~ e). **S1**, the zero peak, was the only subpopulation whose FRET fraction increased during the reaction, which was due to the increase of fully hybridised duplex with FRET signal around 0. Although measured at single molecule level, since the increase of **S1** is due to the decrease of other high FRET subpopulations, the kinetic parameters of the increase of **S1** should be consistent with those measured in bulk studies if the pseudo first-order model is valid. As shown in Table 4.3, compared to the parameters determined in the bulk FRET experiment, similar fractions of the slow and fast decaying components of **S1** were revealed by the smFRET experiment. Furthermore, the time constants of both components measured by smFRET are 3-fold larger than their counterparts in bulk FRET experiments (ratios of 2.9 ± 0.4 and 3.2 ± 0.2 for slow and fast component respectively). The prolonged time constant at single molecule level is consistent with previously reported results from a similar human telomeric G4 system (unfolding time constant: bulk, 116.3 s [5]; single molecule, 791 s [4]). **S3** might be largely composed of partially-unfolded G4 structures, which are hybridised into duplex very fast (time constant 139.5 ± 35.1 s), and contribute to the fast increase of **S1**. It is interesting that **S3** also contains a small amount slow decaying components (time constant 2,226.9 ± 1093.1 s) indicating the presence of certain stable structures. **S2** and **S4** are best fitted by single exponential decay with unfolding time constants 5,047.9 ± 234.6 s^{-1}, and 3,056.8 ± 1250.8 s^{-1} respectively. According to their donor-acceptor distances in the presence of C-rich complimentary strand, **S2** and **S4** imply the formation of two G4s on the same strand (otherwise FRET signals would disappear if only one G4 is formed).

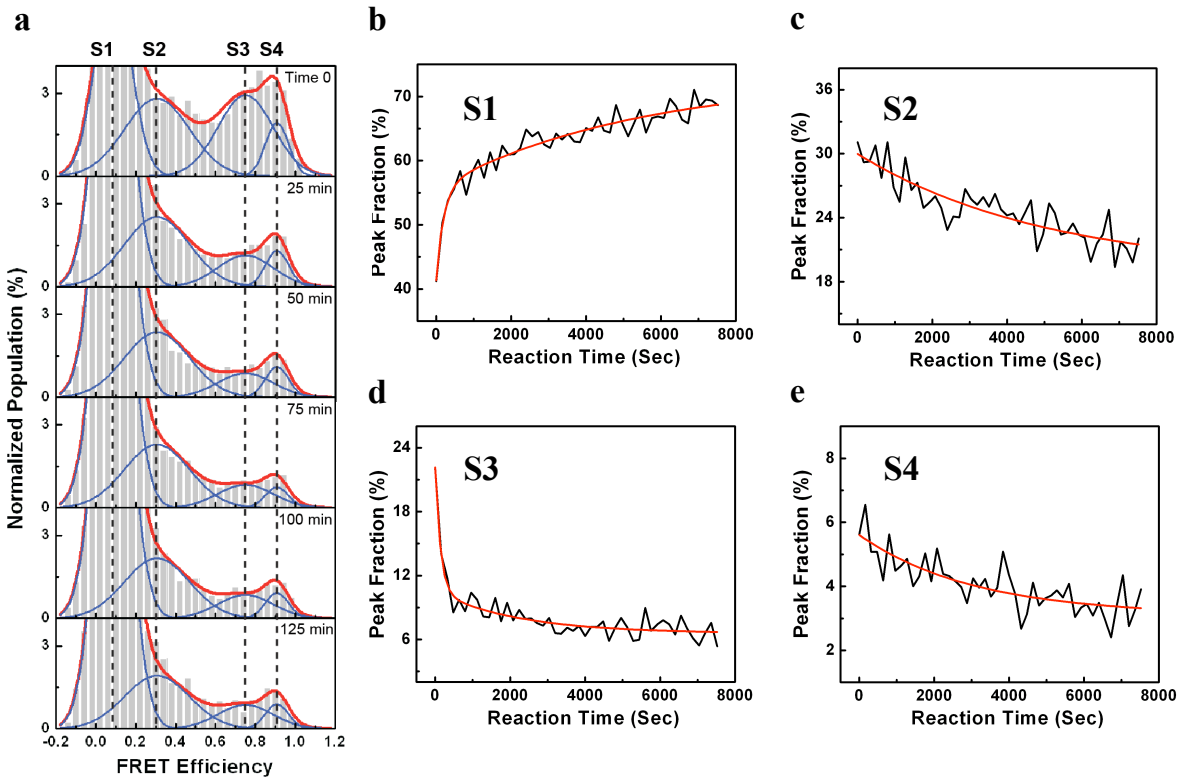


Figure 4.15: TnIc two-repeat MNSG4 unfolding measured by smFRET and corresponding fitted curves.

(a) smFRET histograms of two-repeat MNSG4 (formed by oligo **$Tr_{2\times MNS}$-III**) opened by hybridisation of its C-rich complimentary strand (oligo **$Tr_{2\times MNS}$-II**). 4 subpopulations (**S1** ~ **S4**) were identified by global fitting. **(b)** ~ **(e)** Fraction changes of subpopulation versus reaction time of each subpopulation (black curves, **S1** ~ **S4** respectively). Each curve was fitted by either two or one component exponential decay model. Hybridisation was performed between 100 pM G4 forming oligo and 2 nM complimentary C-rich oligo. The reaction was carried out at 25 °C in 100 mM KCl solution containing 10 mM Tris-HCl (pH 7.4), 200 μM Vitamin C, and 0.05 % Tween 20.

Table 4.3: Parameters obtained from curve fittings of two-repeat TnIc MNSG4 unfolding measured by bulk and single molecule FRET experiments in solution.

Subpopulation	Fraction (A) / %	$\langle\tau\rangle$ / s	k / 10^{-5} s^{-1}
Bulk FRET measurement			
Averaged [a]	A_1 43.7 ± 1.1 A_2 56.3 ± 1.6	τ_1 69.2 ± 0.7 τ_2 1,673.9 ± 9.0	k_1 1,445.1 ± 13.7 k_2 59.7 ± 0.3
Single molecule FRET measurements			
S1	A_1 45.6 ± 0.8 A_2 54.4 ± 1.2	τ_1 197.3 ± 28.8 τ_2 5,303.8 ± 248.9	k_1 506.8 ± 74.0 k_2 18.9 ± 0.9
S2[b]	--	τ_1 5,047.9 ± 234.6	k_1 19.8 ± 0.9
S3	A_1 74.7 ± 2.5 A_2 25.3 ± 1.9	τ_1 139.5 ± 35.1 τ_2 2,226.9 ± 1,093.1	k_1 716.8 ± 180.4 k_2 44.9 ± 21.0
S4	--	τ_1 3,056.8 ± 1,250.8	k_1 32.7 ± 13.4

Experiments were carried out in solution containing 10 mM Tris-HCl (pH 7.4) and 100 mM K+ at 25 °C.

[a] when curves are fitted by a two-component exponential decay model, A_1, τ_1, and k_1 are parameters for the fast decaying species, while A_2, τ_2, and k_2 are parameters for the slow decaying species.

[b] when curves are fitted by a one-component exponential decay model, only τ_1, and k_1 are calculated and listed.

4.7 DMS footprinting of TnIc MNSG4 and -80G4

Because the N-7 positions of guanines are protected from DMS methylation when the Hoogsteen hydrogen bonds exist, DMS footprinting can reveal guanines involved in G4 formation. To determine those guanines that are involved in G4 formation, bands with fast mobility on native gels (bands indicated by black arrows in Figure 4.1 b and c for MNSG4 and -80G4, respectively) were cut out and DMS footprinting was carried out. The result is shown as Figure 4.16.

In both MNSG4 (oligo **Tr_{MNS}-I**) and -80G4 (oligo **Tr_{-80}-I**), four runs of three contiguous guanines were protected from DMS methylation, showing both of them are composed of three layers of stacked G-tetrads. In MNSG4 (Figure 4.16 a), strong DMS protection is observed at G1~G3, G5~G7, G11~G13, and G14~G15. Two guanines at the 3' of MNSG4 (G17~18, oligo **Tr_{MNS}-I**) are not involved in G4 formation. Compared to other guanines involved in loops (G4, G8, and G9), G10 is partially protected from DMS. Because G13 is fully protected from DMS, this partial protection of G10 may be due to the gyration hindrance from the adjacent G4 structure, which blocks its access by DMS. The DMS footprinting pattern of -80G4 is

more difficult to interpret (Figure 4.16 b). In the first 5' run of guanines (G1 ~ G5), partial protection of guanines is apparent. Meanwhile, in the second and third runs of guanines (G9~G12 and G13~G16), partially protected flanking guanines in each G-run are also found (G9/G12 and G12/G16) with two fully protected guanines in the middle (G10~G11 and G14~G15).

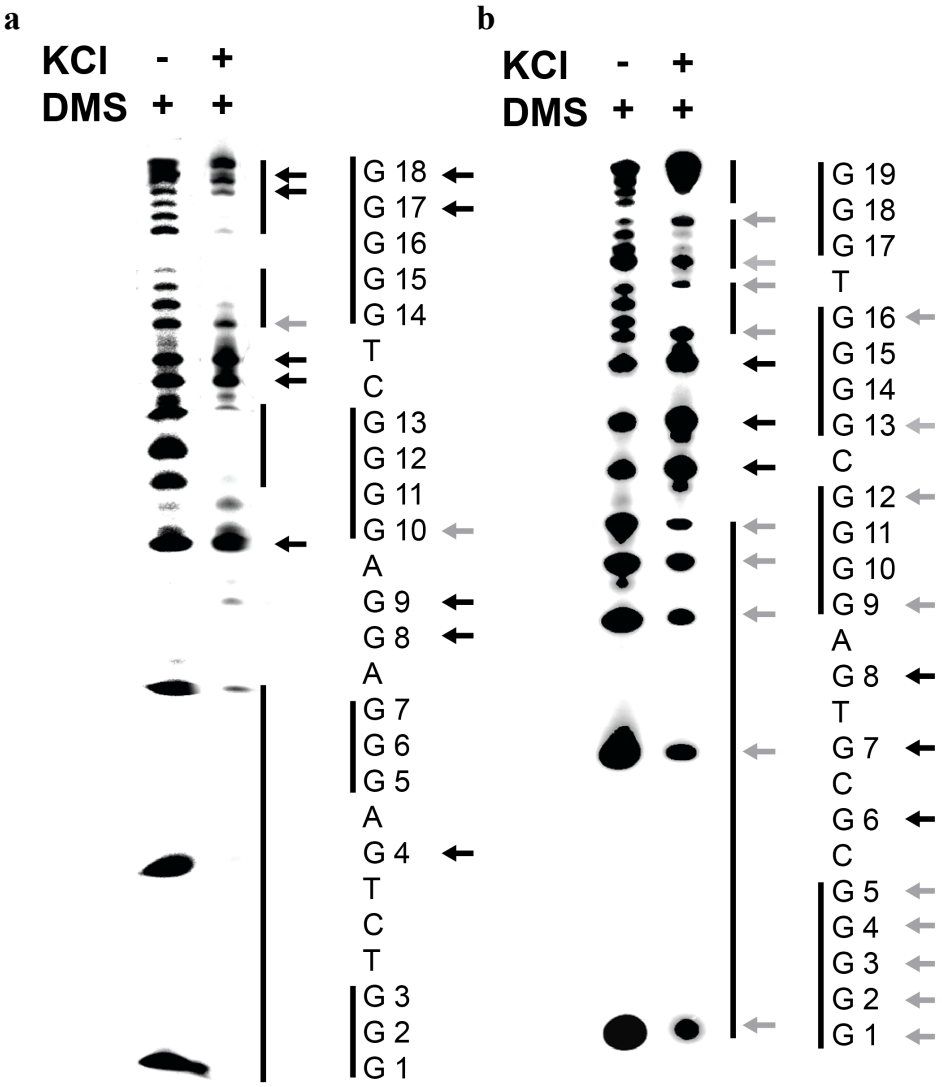


Figure 4.16: DMS footprinting results of TnIc MNSG4 and -80G4.
(a) DMS footprinting of MNSG4 (oligo **Tr_{MNS}-I**). **(b)** DMS footprinting of -80G4 (oligo **Tr_{MNS}-I**).

This poorly defined footprinting pattern of -80G4 element is possibly due to the long first loop (at least 7 nucleotides) which induces flexibility of the G4 conformations. Together with CD and smFRET experiments, it's highly likely that this sequence can fold into a mixture of exchanging parallel intramolecular G4s with different loop

arrangements in KCl solution. Furthermore, in the second and third G-runs, it's clear that G12 and G13 are better protected from DMS than G9 and G16, which may reflect the preference of different intramolecular G4s. According to the different protection effects in DMS footprinting, a preferred folding pattern of MNSG4 is proposed as $G_3N_5G_3N_5G_3N_2G_3$, while for the -80G4 multiple folding patterns may coexist with first loop ranging between 7 ~ 10 nucleotides, and second and third loops ranging between 1 ~ 2 nucleotides. Both of them are distinct from the general folding pattern of promoter G4s indentified in cancer-related genes. Although the exact G4 structures of MNSG4 and -80G4 cannot be proposed based on DMS footprinting results alone, those fully protected guanines in both sequences clearly indicate the formation of G4s.

4.8 Stabilising TnIc MNSG4 and -80G4 by a G4-binding ligand

Once it has been proved that these G4-forming sequences could adopt stable G4 structures in solution, it was of interest to investigate whether small G4-binding molecules could further stabilise these G4 structures. As has been demonstrated by Vilar *et al.* over the past few years, a group of metallo-organic complexes, especially those with two linked metal moieties, can act as excellent G4-binding ligands with relatively high affinity and selectivity [53, 54]. X-ray studies also have shown that the di-metal compounds are strong G4 binders via a combination of π-π stacking on the end of G-tetrad, and electrostatic interactions with adjacent loops and grooves of G4s. Amongst these metallo-organic complexes, the di-copper complex (compound 3 as shown in Figure 5.6) was found to possess highest selectivity and affinity to G4s over duplex (over 100-fold as determined by FID assay) [53]. Indeed, by screening 20 different metallo-organic complexes in cells, complex 3 exhibited the best cellular performance (discussed in next chapter). Thus, I further investigated the ability of this complex (compound 3) in binding to and stabilising the TnIc MNSG4 and -80G4 in solution.

4.8.1 Stabilising ability of di-copper complex (compound 3) determined by CD melting

The CD melting experiment is a well-established method for assessing the stabilising ability of G4-binding ligands to certain G4 structures by measuring the changes of G4 melting temperature (i.e. ΔT_m) in the absence and presence of the ligands being studied. As described in the previous section, in the absence of G4-binding ligand, MNSG4 folds into anti-parallel structures (parallel structures may also be present) in 100 mM K^+ solution with T_m = 71.2 ± 0.4 °C. However, in 10 μM complex 3, the CD spectrum of MNSG4 dramatically changed. The spectrum, which is shown in Figure 4.17 a, is characterised by a positive peak at 263 nm, and a pronounced shoulder between 280 and 300 nm, consistent with the formation of hybrid parallel/anti-parallel G4 structures. The melting temperatures measured at 263 nm and 295 nm are 90.4 ± 0.7 °C and 86.2 ± 2.1 °C, respectively (Figure 4.17 b). It is likely that the di-copper complex preferentially binds to parallel MNSG4 structures, and thus promotes the transition of MNSG4 from anti-parallel to parallel structures. For the -80G4, which has been shown to adopt largely parallel structures in solution containing 100 mM K^+ in the absence of di-copper complex, the presence of the complex did not change its CD spectrum (Figure 4.17 c). Due to the strong stabilising ability of the complex, only a small fraction of the parallel -80G4 was melted at 93 °C, and based on the fitted curve obtained at 263 nm, the melting temperature is estimated to be over 100 °C (Figure 4.17 d). Based on the high selectivity (100-fold G4 over duplex) and proposed G4 stabilising mechanisms of this complex (i.e. π-π stacking and electrostatic interactions), only intramolecular G4s can be efficiently stabilised by the complex. Thus, although -80G4 may adopt intermolecular G4s, the elevated melting temperature observed here indicates the stabilised intramolecular parallel -80G4. The binding preference to parallel G4s observed here might be due to the long loops in both MNSG4 and -80G4. Long lateral or diagonal loops may partially or fully cover the end G-tetrad, and impede the π-π stacking, while the long external loops could leave enough space for the π-π stacking at the end G-tetrad for the first metal moiety, and even provide ideal electrostatic interactions sites for the second metal moiety.

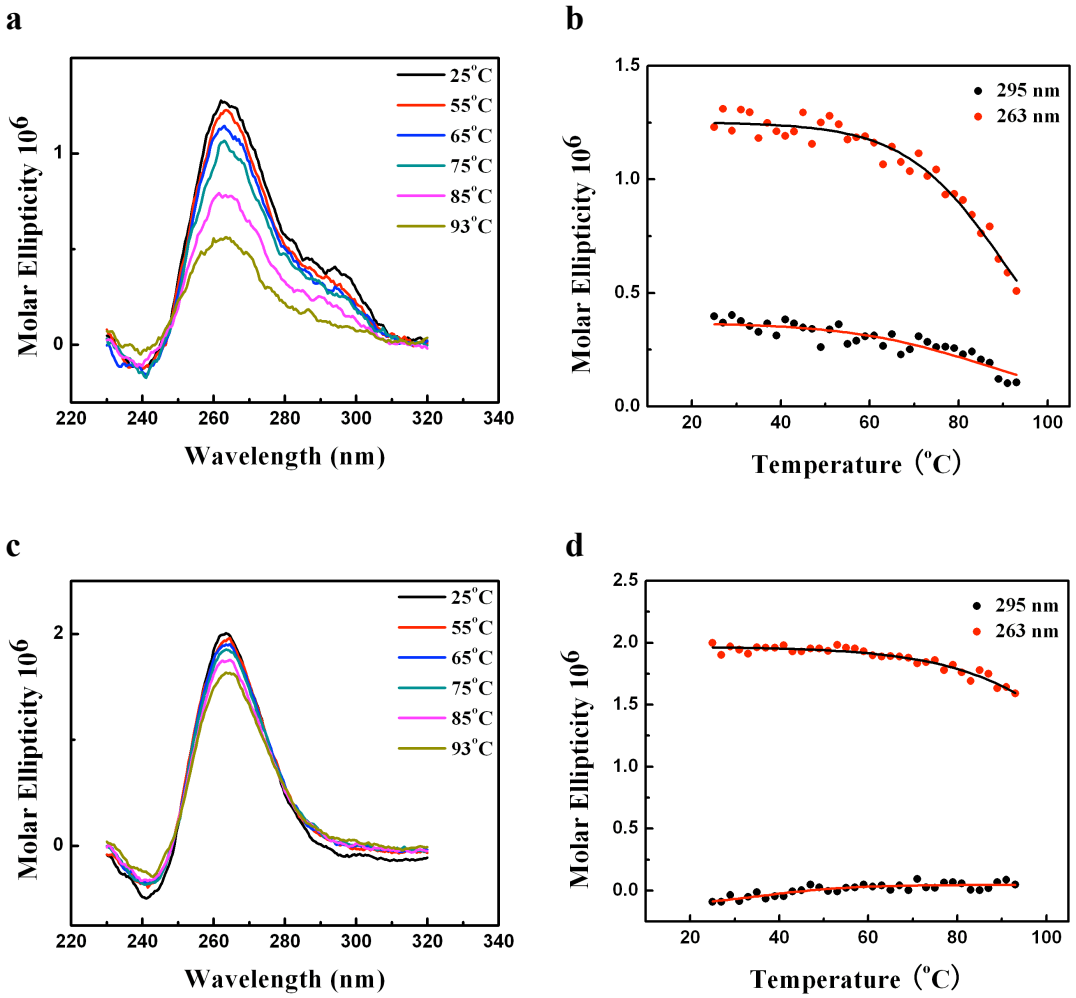


Figure 4.17: CD melting of TnIc MNSG4 and -80G4 in the presence of the G4-binding ligand complex 3.

(a) CD melting spectra of TnIc MNSG4 (oligo **Tr_{MNS}-I**, 5 μM**)** in the presence of 10 μM complex 3. **(b)** CD signal changes of TnIc MNSG4 (oligo **Tr_{MNS}-I)** at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, both of which were fitted by a sigmoidal model (indicated by red line and black line respectively). The melting temperature measured at 265 nm and 295 nm are 90.4 ± 0.7 °C and 86.2 ± 2.1 °C respectively **(c)** CD melting spectra of TnIc -80G4 (oligo **Tr_{-80}-I**, 5 μM**)** in the presence of 10 μM complex 3. **(d)** CD signal changes of TnIc -80G4 (oligo **Tr_{-80}-I)** at 295 nm (black dots) and 263 (red dots) nm wavelengths in response to temperature changes, both of which were fitted by a sigmoidal model (indicated by red line and black line respectively). All experiments were carried out in 10 mM Tris-HCl buffer (pH 7.4) containing 100mM K^+.

4.8.2 Binding affinity of di-copper complex to TnIc MNSG4 and -80G4

In the biophysical characterisation, the di-copper complex (compound 3) was found to be a strong fluorescent quencher to both Cy3 and Cy5. This property of the complex provides an efficient method to study the binding behaviours of the complex to fluorophore-labeled G4s. As mentioned previously, although the ss dual-labeled MNSG4 and -80G4 systems are not suitable to study structural heterogeneity of G4 due to the short donor-acceptor distance in folded states, these systems are of particularly value in studies of ligand-G4 interactions. As shown in Figure 4.18 a and b, the decreases of Cy5 fluorescence of MNSG4 and -80G4 quenched by increases of the di-copper complex (compound 3) were measured, and the ligand-binding curves were fitted by Eq. 2.21 assuming the binding stoichiometry equaling to one (one complex binds to one G4 model). The affinity (association) constants of the complex to the MNSG4 and -80G4 were determined as $2.0 \pm 0.3 \times 10^6$ M^{-1}, and $5.7 \pm 0.2 \times 10^6$ M^{-1}, respectively, in agreement with a similar affinity constant of the complex to human telomeric G4 has been reported previously (7.97×10^6 M^{-1}) [53]. Based on these results, the complex shows reasonably good affinity to both G4s, and is a better binder to -80G4 as compared to the MNSG4. In FID assays by using the same but unlabeled MNSG4 and -80G4 sequences, Vilar *et al.* have shown that the di-copper complex indeed has higher affinity to -80G4 as compared to the MNSG4 ($^{MNSG4}DC_{50}$ = 0.31, and $^{-80G4}DC_{50}$ = 0.19).

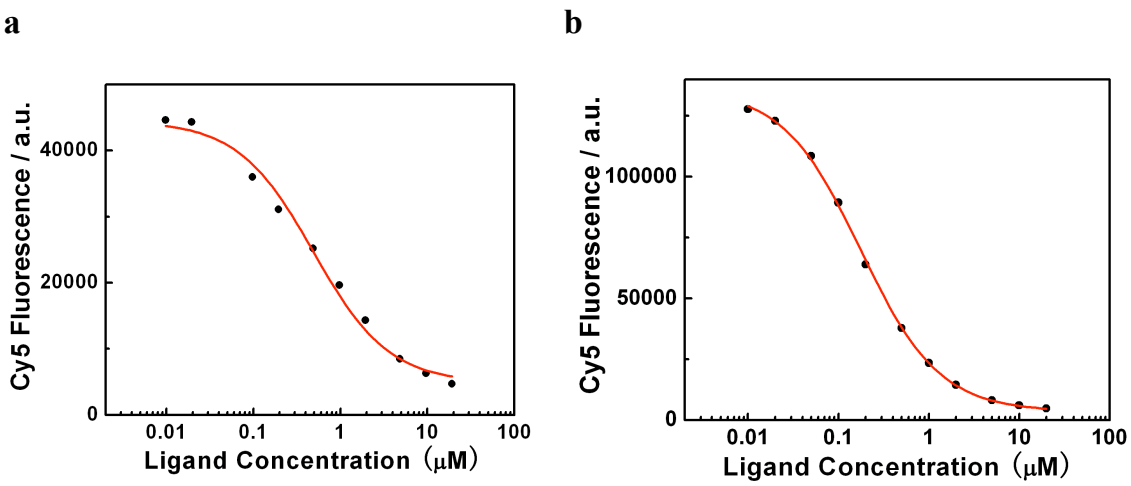


Figure 4.18: Ligand-binding curves for the ss dual-labeled MNSG4 and -80G4 systems as a function of the ligand concentration.

(a) Quenched Cy5 fluorescence of ss dual-labeled MNSG4 system (oligo **Tr_{MNS}-IV**, 100 nM) at different concentration of complex 3. **(b)** Quenched Cy5 fluorescence of ss dual-labeled -80G4 system (oligo **Tr_{-80}-IV**, 100 nM) at different concentration of complex 3. Experiments were carried out in 10 mM Tris-HCl buffer (pH 7.4) containing 100mM K^+ at 25 °C. Data were fitted by Eq. 2.21 assuming the binding stoichiometry equaling to one. Affinity (association) constants of complex 3 with MNSG4 and -80G4 were determined as $2.0 \pm 0.3 \times 10^6$ M^{-1}, and $5.7 \pm 0.2 \times 10^6$ M^{-1}, respectively.

4.9 Discussion

By following bioinformatics approaches, genes involved in cardiac function-related pathways, together with some other pathways, were found to be highly enriched in G4-forming motifs within their TRRs (Chapter 3). To further evaluate the validity of these findings, two genes, the human *MEF2D* and *TnIc*, with a high degree of G4s in their TRRs were chosen as models. The G4-forming sequences from the promoter of human *MEF2D* (MEF2DG4: between -169 and -140 bp), and *TnIc* (MNSG4: consensus sequence between -528 and -319 bp, and -80G4: centred at -80 bp) were selected for biophysical characterisations. In this chapter, formation of MEF2DG4 and TnIc G4s was proved by EMSA. Their structural diversities were assessed by combination of CD spectroscopy, smFRET experiments, and DMS footprinting. Thermodynamic and kinetic parameters of individual G4s were also determined by a group of bulk and single molecule experiments. All results indicate that the MEF2D and TnIc G4-forming sequences can adopt relatively stable G4 structures in solution.

4.9.1 Formation of MEF2D and TnIc G4s in solution

Due to its compact structure as compared to randomly coiled oligos with similar length, formation of G4 can be determined by the fast mobility on a native polyacrylamide gel. Similar to those observed by Henderson *et al.* in 1987 [27], formation of the MEF2DG4 and TnIc G4s were determined by the existence of fast migrating bands in EMSA experiments (Figure 4.1, indicated by black arrows). For the TnIc -80G4, besides the fast migrating intramolecular G4, slow migrating bands were also observed as intermolecular G4s (Figure 4.1 c, lane 2, indicated by grey arrows), which is similar to EMSA results of the PDGF-A G4-forming sequence [40]. Based on the intensity of each bands, the intramolecular G4 is still the major conformation (~ 60 %) formed by the TnIc -80G4 in solution containing 100 mM K^+. This result agrees well with the observation that around 40 % of CD signals at 263 nm come from thermally-stable intermolecular G4s in CD melting experiments (Figure 4.2 h). Besides the fast migrating species observed by EMSA experiments, CD spectra signatures, and high FRET subpopulations in smFRET experiments all

provide compelling evidence for the formation of G4 structures from the MEF2D and TnIc G4-forming sequences, as will be discussed later.

4.9.2 Conformational polymorphism of MEF2D and TnIc G4s

Due to the highly dynamic properties, normally a single G4-forming sequence is able to adopt diverse conformations with different stability and biological importance. For example, the MYC G4-forming sequence can adopt "basket" and "chair" types of G4s [36], and a KIT G4-forming sequence can also adopt multiple conformations with varied length of loop 2 [55]. Furthermore, the formation of a particular G4 conformation can be model- or condition-dependent, which further increases the complexity in determining the actual conformation of G4. For example, for the humane telomeric G4, propeller-type parallel G4s were determined by X-ray crystallography [56], while hybrid of parallel/anti-parallel types of G4s have been reported by NMR studies [57, 58].

CD spectroscopy

CD spectroscopy and CD melting are simple but invaluable experiments to probe the formation and general structures of G4s in solution. Several CD spectra corresponding to specific G4s have been well documented as mentioned previously (illustrated in [59]). Furthermore, by following CD melting curves at 263 and 295 nm, formation of a hybrid of parallel/anti-parallel G4s or mixture of parallel/anti-parallel G4s could be clarified [42], and less commonly, the transition between parallel and anti-parallel G4s could be observed [39]. Thus, CD spectroscopy and melting experiments were also used to investigate the structural polymorphism of MEF2D and TnIc G4s. The MEF2DG4-forming sequence was found to adopt a hybrid of parallel/anti-parallel G4(s). The similar spectra obtained in solutions containing K^+ and Li^+ indicate that the stabilising effect from cations is minor, while the stacking energy of G-tetrads dominates the formation of this G4 (Figure 4.2 a). For the TnIc G4s, although it is clear that in solutions containing 100 mM K^+ the MNSG4 and -80G4 adopt anti-parallel (possibly mixed with small amount of parallel G4) and parallel structures, respectively, their spectra are varied depending on the type of cation involved (MNSG4: Figure 4.2 d, and -80G4: Figure 4.2 e). Varied CD spectra

for the MNSG4 and -80G4 indicate that the stabilising effects from particular cations may have a strong influence upon the actual conformation of those G4s. For the two-repeat MNSG4, due to the negatively correlated melting curves obtained at 263 and 295 nm (Figure 4.2 k and l), it is very likely that one parallel and one anti-parallel G4 coexist in the same oligo.

Single-molecule FRET analysis

Due to its high sensitivity and resolution, the smFRET technique has been proved as a valuable method in studying the conformational heterogeneity of G4 [3, 13, 60, 61]. Although different labeling strategies were applied to probe the G4-forming sequences in single molecule experiments, high FRET subpopulations were observed in the MEF2D and TnIc G4-forming sequences, indicating the formation of G4 structures (Figure 4.4, and 4.6 ~ 4.9). Although it has been reported that the labeled fluorophores may introduce certain variables to the conformation and stability of G4s, the smFRET results still correlated well with the CD results. For example, in the case of MEF2DG4, the smFRET histograms of the internal dual-labeled MEF2DG4 system (as illustrated in Figure 4.3 a) showed identical subpopulations compositions in solutions containing different salts (the same peak position and width as fitted by a four-peak Gaussian model, Figure 4.4 a ~ d), which represents a weak structural dependence on the type of cation involved. This is in accordance with the identical CD spectra of MEF2DG4 observed in solutions containing K^+ and Li^+ (Figure 4.2 a). However, for the end dual-labeled MEF2DG4 system (as illustrated in Figure 4.3 b), global fitting did not work as different G4s formed under different salt conditions (shifted **S1** in the absence of salt, Figure 4.4 e, and **S3** in 100 mM K^+, Figure 4.4 h). These differences were caused by the different labeling sites. In the internal dual-labeled system, the donor Cy3 was conjugated via NHS ester in the middle of a long loop (the third loop, 7 nucleotides), which protrudes outside the G4 core structure without interfering with its formation. For the end dual-labeled system on the other hand, Cy3 was labeled to the last guanine (3') involved in G4 formation, which may impede its participation in the G-tetrad and therefore alter the stability of G4. Thus, the internal dual-labeled system has similar conformation and stability as the unlabeled system measured in CD experiments, while the end dual-labeled system requires the incorporation K^+ cations to compensate the instability induced by the Cy3. Due to this reason, fluorophores were labeled either on a linker or on a

protective nucleotide attached to the 5' or 3' end of the G4-forming sequences in subsequent dual-labeled G4 systems. By following this strategy, for the TnIc MNSG4 and -80G4, whose conformations are strongly dependent on cation species as shown by CD spectra (Figure 4.2 d, g, and j), different subpopulation compositions were resolved in the smFRET histograms when measured in different solutions (Figure 4.6 ~ 4.8). Another piece of interesting evidence showing the consistence between CD results and smFRET histograms comes from the ds dual-labeled MNSG4 system. According to the CD spectra (Figure 4.2 d), the MNSG4-forming sequence adopts a chair-type anti-parallel structure in solutions containing Na^+ and K^+. If the ds dual-labeled MNSG4 also adopts a chair-type anti-parallel structure, the donor and acceptor would be very close to each other, because the 5' and 3' ends of the G4-forming sequence exit from the same G-tetrad in the same direction. Indeed, the subpopulation **S4** reveal in the ds dual-labeled MNSG4 system presents the highest FRET efficiency ($E_{\mathbf{S4}} = 0.94 \pm 0.1$, Figure 4.6 b and c) as compared to subpopulations observed in other dual-labeled systems (i.e. ds dual-labeled MEF2DG4, -80G4, and 2×MNSG4 systems). Furthermore, the experimentally determined donor-acceptor distance of the chair-type anti-parallel G4 is 33.8 ± 1.2 Å, which is very similar to the theoretical donor-acceptor distance, 36 Å, based on a simple model of chair-type anti-parallel G4. Besides MEF2DG4 and MNSG4, smFRET histograms of the ds dual-labeled -80G4 are also consistent to its CD results. As compared to the smFRET histograms of other dual-labeled G4 systems obtained in K^+, only one high FRET subpopulation was revealed in the -80G4 system, which echoes the CD result indicating the parallel -80G4 in the presence of K^+ (Figure 4.2 g and h). For the two-repeat MNSG4, a highly cooperative melting behaviour was observed by the CD melting experiments (Figure 4.2 k and l), which indicated the presence of parallel and anti-parallel G4s on the same oligo. Accordingly, high FRET subpopulations representing formation of two G4s were also observed in smFRET experiments (Figure 4.8).

A folding pattern of $G_3N_5G_3N_5G_3N_2G_3$ was determined in the MNSG4 by DMS footprinting, while the folding pattern of the -80G4 is not straightforward to estimate, where multiple partially protected guanines were identified in DMS footprinting. These partially-protected guanines in -80G4 indicate the formation of a mixture of different G4 structures with a long first loop (7 ~ 10 nucletoides) in KCl solution. It is

clear that the general folding pattern of both MNSG4 and -80G4 are distinct from that of proto-oncogenes: $G_3N_{1\text{-}3}G_3N_{2\text{-}9}G_3N_1G_3$ [59]. Different folding patterns between these cardiac promoter G4s and proto-oncogene promoter G4s may relate to their distinct biological functions and therefore warrants further investigation by NMR.

4.9.3 Unfolding thermodynamics and kinetics of MEF2D and TnIc G4s

Stability is one of the key properties of G4 structures that need to be addressed in biophysical characterisations. The stability of a G4 is usually evaluated by its unfolding process either thermodynamically or kinetically. Although the thermodynamic and kinetic properties of G4 unfolding are closely related to each other, they are actually two different aspects reflecting the stability of the G4. In its simplest form, thermodynamic parameters reflect the energy differences (free energy, enthalpy, and entropy) between the folded and unfolded states, and thus reveal whether the unfolding process will happen under certain conditions, and to what extent it will happen. On the other hand, kinetics parameters reflect the rate of unfolding process under certain conditions. It is possible that a G4 with marginal stability in the view thermodynamics (i.e. a small free energy at 37 °C) can be kinetically stable at that temperature if a huge energy barrier existed in its unfolding pathway. Thus, thermodynamic and kinetic measurements were both carried out on the MEF2D and TnIc G4s.

MEF2DG4

Unfolding thermodynamics of MEF2DG4 was firstly investigated by CD thermal denaturation (melting) experiments. The thermodynamic parameters were listed in Table 4.1. The MEF2DG4 was found to adopt a hybrid of parallel/anti-parallel structure(s) according to its CD spectra, and similar melting temperatures were determined by analysing the denaturation transition curves (melting curves) obtained at 263 and 295 nm. However, van't Hoff analysis also revealed significantly different thermodynamic parameters (ΔH, ΔS, and $\Delta G_{(310K)}$) corresponding to the melting curves obtained at these two wavelengths. Furthermore, according to the melting temperature of the MEF2DG4 (T_m ~ 93 °C), it should be more thermodynamically stable than most of other G4 reported so far. Contrarily, according to the van't Hoff

analysis, it only has a moderate thermal stability ($\Delta G_{(310K)}$ ~ 5.1 kcal mol^{-1}) as compared to other G4s with similar melting temperature measured under similar condition (100 mM K^+), such as the human telomeric G4 (T_m ~ 82 °C, $\Delta G_{(310K)}$ ~ 8.4 kcal mol^{-1}) [62] and MYB G4 (T_m ~ 87 °C, $\Delta G_{(310K)}$ ~ 10.0 kcal mol^{-1}) [63]. According to Eq. 2.5 and 2.8, $\Delta G_{(310K)}$ is positively related to the melting temperature (T_m) and ΔH, both of which are directly obtained by van't Hoff analysis. Thus, the small $\Delta G_{(310K)}$ value of MEF2DG4 is mainly caused by the small ΔH. Indeed, as shown in Table 4.1, the ΔH values of MEF2DG4 are found to be smaller than the TnIc G4s, even though TnIc G4s have relatively low melting temperatures. It has been noticed that ΔH obtained from van't Hoff analysis may not be equal to the true (calorimetric) ΔH, when multiple unfolding intermediates are present or slow unfolding kinetics are involved [64, 65]. It is common to assume that the thermal denaturation of G4 is a simple two-state process without any intermediates. In real cases though, the folded G4 may transit into multiple intermediates prior to be fully denatured. Such transitions may have small enthalpy values, resulting in a broad melting curve (width of melting curve is negatively related to the ΔH value obtained from van't Hoff analysis). A dual-wavelength parametric test for the simple two-state model was carried out [66]. In this test, strict linear behaviour should be observed when CD signals obtained at two wavelengths were plotted against each other, while any deviations from the linearity indicates violation of the two-state model. When MEF2DG4 CD signals obtained at 260 and 295 nm were plotted against each other, apparent deviations from linearity were observed when temperatures reached ~ 70 °C, indicating presence of populated intermediates. Thus, in the case of the MEF2DG4, unfolding intermediates might be the dominant reason for the underestimated ΔH and $\Delta G_{(310K)}$ values. As compared to ΔH or $\Delta G_{(310K)}$, T_m is a better indicator for estimating the thermal stability of MEF2DG4, since it is the most precisely determined parameter even in the presence of multiple intermediates with different T_m values [65]. Due to its high T_m value, MEF2DG4 is among the most thermally stable G4s that have been reported so far.

Unfolding kinetics of MEF2DG4 was also investigated by using the 2-component ds dual-labeled systems (Figure 4.3 c). For the internal dual-labeled MEF2DG4 system (oligos **MEF-I** and **III**), its unfolding processes in 100 mM K^+ at 25 °C was too slow to be precisely measured and fitted, while for the end dual-labeled MEF2DG4 system

(oligos **MEF-II** and **III**), a detectable decrease of Cy5 fluorescence was observed after 6 hours. This relatively fast unfolding kinetics of end dual-labeled MEF2DG4 system is related to the labeling of Cy3 as discussed previously. Although the end dual-labeled MEF2DG4 might be less kinetically stable than the unlabeled MEF2DG4-forming sequence, its unfolding rate constant is still much smaller than other intramolecular G4s measured under similar conditions, such as the *MYC* promoter G4 [51]. Although an even smaller unfolding rate constant has been reported in a human telomeric G4 unfolding model (~ 6.9×10^{-6} s^{-1}, 100 mM K^+, 25 °C) [6], this might be related to the different models used. In the case of the MEF2DG4, the 30mer linker attached to the 5' of MEF2DG4-forming sequence (oligo **MEF-II**) provides an annealing site for the C-rich complimentary strand, which promotes the unfolding of adjacent G4 by increasing the local concentration of C-rich strand. In the human telomeric G4 case, no flanking sequence was attached, in which the complimentary strands (a PNA trap in [6]) are unable to promote the G4 unfolding. These complimentary strand-dependent G4 unfolding kinetics will be discussed later.

TnIc MNSG4 and -80G4

By carrying out CD melting experiments, thermodynamic parameters of TnIc MNSG4 and -80G4 unfolding were also determined. The TnIc MNSG4 and -80G4 were composed of anti-parallel and parallel G4 structures, respectively, in the presence of 100 mM K^+ (Figure 4.2 d, e, g, and h), although the possibility of parallel structures still cannot be fully ruled out in the MNSG4 case. The anti-parallel MNSG4 and parallel -80G4 exhibited similar thermal stability as indicated by their comparable T_m, ΔH, ΔS, and $\Delta G_{(310K)}$ values measured by CD melting experiments (Table 4.1). TnIc MNSG4 and -80G4 both have similar or higher thermal stability as compared to the human telomeric G4 depending on the sequence and methods used ($(TTAGGG)_4$, T_m, = 63 °C, $\Delta G_{(310K)}$ = 3.8 kcal mol^{-1} [67]; $AGGG(TTAGGG)_3$, T_m, = 63 °C, $\Delta G_{(310K)}$ = 4.6 kcal mol^{-1} [68]; $GGG(TTAGGG)_3$, T_m, = 65 °C, $\Delta G_{(310K)}$ = 5.0 kcal mol^{-1} [68]). As compared to other G4s with similar loop length, such as MYC, KIT, WNT1 and VEGF G4s (total loop length 10, 11, 12, and 13 respectively), both of the TnIc MNSG4 (total loop length 12 ~ 14 depending on folding pattern) and -80G4 (total loop length 11 ~ 13 depending on folding pattern) also show similar thermal stability [63]. This coherence indicates that intermediates in the processes of

thermal denaturation of MNSG4 and -80G4 are not as significant as those observed in the case of MEF2DG4. Furthermore, according to the already approved biological roles of human telomeric, MYC, VEGF, and KIT G4s, it is reasonable to assume that both TnIc MNSG4 and -80G4 are thermally stable enough to hold certain biological roles in the promoter of human *TnIc* in the context of adjacent duplex.

Kinetic stability of both TnIc MNSG4 and -80G4 were also investigated by studying their opening rates by hybridisation to the excess complimentary DNA strands (pseudo first-order reaction) at different temperatures (MNSG4: oligos **Tr_{MNS}-III** and **Comp-Cy5**, hybridised by **Tr_{MNS}-II**; -80G4: oligos **Tr_{-80}-III** and **Comp-Cy5**, hybridised by **Tr_{-80}-II**). As shown in Figure 4.12, the decaying curves of MNSG4 and -80G4 were best fitted by double exponential decays. Although weighted mean rates were analysed previously in similar unfolding studies [6], due to their distinct kinetic properties, the fast and slow unfolding components of both MNSG4 and -80G4 were regarded as different conformations and are discussed therefore separately. Generally, for the MNSG4, the temperature-independent fast unfolding component reflected hybridisation of partially folded or randomly coiled oligos, while the slow unfolding component represented unfolding of G4. For the -80G4, both slow and fast unfolding components were regarded as indicative of unfolding of different G4s. According to its lower activation energy and longer unfolding time constants especially at high temperature, the MNSG4 is less stable kinetically than the -80G4 (Table 4.2). Indeed, a small activation enthalpy and very negative entropy indicate that the unfolding of MNSG4 is entropically driven and that it can unfold relatively easily [3, 52]. On the other hand, the unfolding of the -80G4s is enthalpy-driven. These differences are believed to be related to the different loop arrangements in the MNSG4 and -80G4. According to limited kinetic stability data of other G4s, the MNSG4 ($t_{1/2\ (35\ °C)} \sim 2.4 \times 10^3$ s) and the fast unfolding -80G4 ($t_{1/2\ (35\ °C)} \sim 2.3 \times 10^3$ s) are slightly less kinetically stable than the others, while the slow unfolding -80G4 ($t_{1/2\ (35\ °C)} \sim 1.2 \times 10^5$ s) is much more stable than the others (*Oxytricha* telomeric G4: $t_{1/2\ (37\ °C)} \sim 6.5 \times 10^4$ s [69]; KIT1 G4: $t_{1/2\ (37\ °C)} \sim 5.7 \times 10^3$ s [13]; KIT2 G4: $t_{1/2\ (37\ °C)} \sim 2.8 \times 10^4$ s [13]; and human telomeric G4: $t_{1/2\ (37\ °C)} \sim 1.0 \times 10^4$ s [6]).

TnIc two-repeat MNSG4

As determined by CD melting experiments, two-repeat MNSG4 can adopt a mixture of parallel and anti-parallel G4s, which show a complicated unfolding pattern in their thermal denaturation processes (Figure 4.2 k). The reversely-correlated first derivatives of melting curves obtained at 263 and 290 nm clearly indicate the cooperative unfolding patterns of the parallel and anti-parallel G4s, which further supported the assumption that two G4s coexist on the same oligo (Figure 4.2 l). In the two parallel G4s that were adopted by the two-repeat MNSG4-forming sequence, one shows relatively high thermal stability (T_m, = 89.2 °C, $\Delta G_{(310K)}$ = 8.7 kcal mol^{-1}), while the anti-parallel G4 is less stable (T_m, = 68.3 °C, $\Delta G_{(310K)}$ = 3.3 kcal mol^{-1}) (Table 4.1).

Kinetic stability of TnIc two-repeat MNSG4 was also investigated at bulk and single molecule level by studying the hybridisation to its complimentary strand (pseudo first-order reaction) at 25 °C. In bulk measurement, fast and slow unfolding subpopulations were revealed, which correspond to hybridisation of partially folded structures and averaged unfolding behaviours of MNSG4, respectively. Interestingly, equal factions of the fast and slow unfolding subpopulations were revealed by both bulk and single molecule (**S1**, averaged unfolding kinetics) FRET experiments, indicating that the formation of these intramolecular G4s is concentration-independent (formation of intramolecular G4s). In single molecule kinetics measurements, two stable G4s were revealed with similar kinetic stability as KIT1 G4 [13], but with lower stability than the MEF2D, single repeat MNS and -80 G4s.

To comprehensively consider the thermal and kinetic stability of these G4s, G4/duplex competition assays were carried out. In G4/duplex competition assays, the dual-labeled G4-forming sequences were incubated at 95 °C for 15 min in the presence of 1.5-fold excess of C-rich complimentary strands, and then gradually cooled to room temperature to allow the formation of G4s. By following this protocol, G4s can survive only when they are thermally and kinetically stable. G4s with high T_m (thermally stable) can avoid trapping by duplex in prior to the their formation, while G4s with slow unfolding rate (kinetically stable) can result in substantial residual folded states in the presence of C-rich strands at lower temperature. Due to its highest thermal and kinetic stability, high FRET signals were only observed in the MEF2DG4 (Figure 4.9), indicating its remarkable overall stability.

4.9.4 Concerns in thermodynamic and kinetic characterisations of G4 unfolding

There are several concerns in obtaining and analysing thermodynamic and kinetic parameters in studies of G4 unfolding.

The first concern is the validity to compare the thermal stability among different G4s. It has been noticed that thermodynamic values for the unfolding of human telomeric G4s differ significantly in different literatures. For example, under similar conditions (100 mM K^+), the free energy changes of human telomeric G4s unfolding at 310 K ($\Delta G_{(310K)}$) ranged from the very stable (14.8 kcal mol^{-1} [70]) to moderately stable (3.0 kcal mol^{-1} [64]) (reviewed in [7, 65]). Although slightly different telomeric G4-forming sequences were used in these studies, it is not likely a reasonable explanation for such large discrepancies. Besides, the corresponding melting temperatures measured in these studies are quite coherent (T_m ~ 66 and 69 °C respectively). According to Lane *et al.*, this may attribute to the under- or overestimated ΔH with certain unknown reasons [65]. One possibility is the existence of populated intermediates in the process of thermal denaturation. It is commonly assumed that the thermal denaturation is a simple two-state process, with a folded state transiting into an unfolded state without any intermediates. However, the real scenario might be more complicated than this simplified two-state model. Although no such intermediates have been reported up to date in the G4 thermal denaturation, intermediates in the conformational switch of human telomeric G4 have been well studied already [71]. Multiple unfolding intermediates with low enthalpies would lead to a broadened melting curve without shifting the middle point, which then gives an underestimated ΔH value, and a still accurate T_m value in van't Hoff analysis, such as the MEF2DG4 discussed above. Fortunately, by utilizing a simple, dual-wavelength parametric test [66] as described above, the two-state denaturation process can be easily tested. Although in some cases T_m might be more accurate in assessing the thermal stability of a G4, T_m may also be altered in cases when heating or cooling is much faster than the transiting between folded and unfolded states of G4s leading to shifted melting curves (hysteresis effects) [62] or even inducing the formation of different G4 structures [72]. However, these phenomena only reported in solutions

with low cation concentrations. Thus, in the melting curves of MEF2D and TnIc G4s , hysteresis effects are not likely to happen. Another possibility inducing a biased ΔH may relate to the neglect of heat capacity changes (ΔC_p) when analysing the thermal denaturation curves. As pointed out by Lane *et al.* [65] and Chaires [7], ΔC_p should not be assume to be zero in analysing thermal denaturation curves of G4, since it originates from a series of intrinsic changes in the denaturation process of G4, such as exposure of hydrophobic surface of oligos [73], and elevated structural fluctuations of the partially unfolded intermediates [74]. Furthermore, ΔH and ΔC_p are closely correlated, and improper estimation of ΔC_p should lead to biased ΔH in van't Hoff analysis [75]. However, it is extremely difficult to fit the melting curve with accurate ΔC_p values. Thus, although little can be done to clarify this ambiguity, this problem must be acknowledged. Due to the above reasons, together with intrinsic errors of different equipments, it is not surprising to find that dramatically different thermodynamic parameters are reported by different laboratories for similar G4s. Thus, returning to the concern mentioned above, comparison of thermodynamic parameters among different G4s may not be suitable in the majority of cases.

Another concern is about taking the G4 unfolding as the rate-limiting step in studying G4 kinetic stability. The unfolding kinetics of G4 are usually measured by trapping the unfolded DNA with duplex in the presence of excess C-rich complimentary strand, and thus the whole reaction follows a pseudo first-order process. This is actually based on several preconditions. The first is that the rate-limiting step is the unfolding of G4, while the subsequent hybridisation to a C-rich strand is relatively fast as compared to the G4 unfolding. The unfolding rate constants of most G4s are between 10^{-6} s^{-1} and 10^{-3} s^{-1} [5, 6, 13, 51], while the hybridisation rate constants are normally around 10^5 M^{-1} s^{-1} [51]. Thus, in most bulk measurements, this precondition is tenable when the C-rich strands are present in micromolar range, which gives a pseudo first-order hybridisation rate constant of 0.1 s^{-1}. However, when the hybridisation was taking at nanomolar level, especially in the single molecule experiments, this precondition would be in jeopardy. In single molecule experiments, the pseudo first-order hybridisation rate constant would be in similar magnitude to the unfolding rate constant of G4s, and thus the overall unfolding reaction would be a second-order one. This would induce a smaller unfolding rate constant (prolonged unfolding time constant) at lower concentration, which explains the long unfolding

time constants of two-repeat MNSG4 and human telomeric G4 [4, 5] obtained in smFRET measurements. The other precondition is that the hybridisation happens after the complete unfolding of G4, or the "unfolding rate constant" of a G4 would be actually "transition rate constant" between the G4 and certain intermediates. Similar to protein unfolding, the unfolding of G4 is also composed of multiple partially unfolded intermediates. These intermediates, especially those from G4s with long loops, may provide the invasion sites for C-rich complimentary strands, which then further promote the unfolding process, such as a scheme proposed by Green *et al.* [6]. Then, the invasion of C-rich strand is usually followed by the further unfolding of the partially-folded structures, which gives a observed hybridisation rate constant smaller than that corresponding to a typical duplex hybridisation, such as the fast unfolding subpopulation identified in MNSG4 unfolding kinetics study (Table 4.2). In this case, the unfolding reaction is composed of a first-order transition process (G4 to intermediates) and a first-order invasion process (further resolution of intermediates by C-rich strand), and thus gives an overall second-order reaction. One method to rule out the possibility of C-rich strand invasion is to compare the observed unfolding rate constant in the presence of different concentration of C-rich strand [76][404].

Thus, although the unfolding thermodynamic and kinetic parameters of MEF2D and TnIc G4s are demonstrated in this chapter, further experiments would be necessary to clarify these ambiguities, which will be discussed in Section 6.4.

4.10 References

1. Parkinson, G. N., Lee, M. P., and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA, *Nature 417*, 876-880.
2. Li, J., Correia, J. J., Wang, L., Trent, J. O., and Chaires, J. B. (2005) Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal, *Nucleic Acids Res 33*, 4649-4659.
3. Lee, J. Y., Okumus, B., Kim, D. S., and Ha, T. (2005) Extreme conformational diversity in human telomeric DNA, *Proc Natl Acad Sci U S A 102*, 18938-18943.
4. Ying, L., Green, J. J., Li, H., Klenerman, D., and Balasubramanian, S. (2003) Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer, *Proc Natl Acad Sci U S A 100*, 14629-14634.
5. Green, J. J., Ladame, S., Ying, L., Klenerman, D., and Balasubramanian, S. (2006) Investigating a quadruplex-ligand interaction by unfolding kinetics, *J Am Chem Soc 128*, 9809-9812.
6. Green, J. J., Ying, L., Klenerman, D., and Balasubramanian, S. (2003) Kinetics of unfolding the human telomeric DNA quadruplex using a PNA trap, *J Am Chem Soc 125*, 3763-3767.
7. Chaires, J. B. (2010) Human telomeric G-quadruplex: thermodynamic and kinetic studies of telomeric quadruplex stability, *FEBS J 277*, 1098-1106.
8. Simonsson, T., Pecinka, P., and Kubista, M. (1998) DNA tetraplex formation in the control region of c-myc, *Nucleic Acids Res 26*, 1167-1172.
9. Ambrus, A., Chen, D., Dai, J., Jones, R. A., and Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization, *Biochemistry 44*, 2048-2058.
10. Phan, A. T., Kuryavyi, V., Burge, S., Neidle, S., and Patel, D. J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter, *J Am Chem Soc 129*, 4386-4392.
11. Rankin, S., Reszka, A. P., Huppert, J., Zloh, M., Parkinson, G. N., Todd, A. K., Ladame, S., Balasubramanian, S., and Neidle, S. (2005) Putative DNA quadruplex formation within the human c-kit oncogene, *J Am Chem Soc 127*, 10584-10589.
12. Kuryavyi, V., Phan, A. T., and Patel, D. J. (2010) Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter, *Nucleic Acids Res 38*, 6757-6773.
13. Fegan, A., Shirude, P. S., Ying, L., and Balasubramanian, S. (2010) Ensemble and single molecule FRET analysis of the structure and unfolding kinetics of the c-kit promoter quadruplexes, *Chem Commun (Camb) 46*, 946-948.
14. Huppert, J. L., and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome, *Nucleic Acids Res 35*, 406-413.
15. Olson, E. N. (2006) Gene regulatory networks in the evolution and development of the heart, *Science 313*, 1922-1927.
16. Potthoff, M. J., and Olson, E. N. (2007) MEF2: a central regulator of diverse developmental programs, *Development 134*, 4131-4140.
17. Alpert, J. S., Thygesen, K., Antman, E., and Bassand, J. P. (2000) Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction, *J Am Coll Cardiol 36*, 959-969.
18. Thygesen, K., Alpert, J. S., and White, H. D. (2007) Universal definition of myocardial infarction, *Eur Heart J 28*, 2525-2538.
19. Bhavsar, P. K., Dhoot, G. K., Cumming, D. V., Butler-Browne, G. S., Yacoub, M. H., and Barton, P. J. (1991) Developmental expression of troponin I isoforms in fetal human heart, *FEBS Lett 292*, 5-8.

20. Bhavsar, P. K., Brand, N. J., Yacoub, M. H., and Barton, P. J. (1996) Isolation and characterization of the human cardiac troponin I gene (TNNI3), *Genomics 35*, 11-23.
21. Bhavsar, P. K., Dellow, K. A., Yacoub, M. H., Brand, N. J., and Barton, P. J. (2000) Identification of cis-acting DNA elements required for expression of the human cardiac troponin I gene promoter, *J Mol Cell Cardiol 32*, 95-108.
22. Kojima, T., Yoshikawa, Y., Takada, S., Sato, M., Nakamura, T., Takahashi, N., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., and Mori, N. (2001) Genomic organization of the Shc-related phosphotyrosine adapters and characterization of the full-length Sck/ShcB: specific association of p68-Sck/ShcB with pp135, *Biochem Biophys Res Commun 284*, 1039-1047.
23. Polushin, N., Malykh, A., Malykh, O., Zenkova, M., Chumakova, N., Vlassov, V., and Kozyavkin, S. (2001) 2'-modified oligonucleotides from methoxyoxalamido and succinimido precursors: synthesis, properties, and applications, *Nucleosides Nucleotides Nucleic Acids 20*, 507-514.
24. Das, H. K., Jackson, C. L., Miller, D. A., Leff, T., and Breslow, J. L. (1987) The human apolipoprotein C-II gene sequence contains a novel chromosome 19-specific minisatellite in its third intron, *J Biol Chem 262*, 4787-4793.
25. Barton, P. J., Cullen, M. E., Townsend, P. J., Brand, N. J., Mullen, A. J., Norman, D. A., Bhavsar, P. K., and Yacoub, M. H. (1999) Close physical linkage of human troponin genes: organization, sequence, and expression of the locus encoding cardiac troponin I and slow skeletal troponin T, *Genomics 57*, 102-109.
26. Hsu, S. T., Varnai, P., Bugaut, A., Reszka, A. P., Neidle, S., and Balasubramanian, S. (2009) A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics, *J Am Chem Soc 131*, 13399-13409.
27. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr., and Blackburn, E. H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs, *Cell 51*, 899-908.
28. Seenisamy, J., Rezler, E. M., Powell, T. J., Tye, D., Gokhale, V., Joshi, C. S., Siddiqui-Jain, A., and Hurley, L. H. (2004) The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4, *J Am Chem Soc 126*, 8702-8709.
29. Phan, A. T., Modi, Y. S., and Patel, D. J. (2004) Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter, *J Am Chem Soc 126*, 8710-8716.
30. Dexheimer, T. S., Sun, D., and Hurley, L. H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter, *J Am Chem Soc 128*, 5404-5415.
31. Dai, J., Dexheimer, T. S., Chen, D., Carver, M., Ambrus, A., Jones, R. A., and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution, *J Am Chem Soc 128*, 1096-1098.
32. Kelly, J. A., Feigon, J., and Yeates, T. O. (1996) Reconciliation of the X-ray and NMR structures of the thrombin-binding aptamer d(GGTTGGTGTGGTTGG), *J Mol Biol 256*, 417-422.
33. Sun, D., Guo, K., and Shin, Y. J. (2011) Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene, *Nucleic Acids Res 39*, 1256-1265.
34. De Armond, R., Wood, S., Sun, D., Hurley, L. H., and Ebbinghaus, S. W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter, *Biochemistry 44*, 16341-16350.
35. Cogoi, S., and Xodo, L. E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, *Nucleic Acids Res 34*, 2536-2549.

36. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, *Proc Natl Acad Sci U S A 99*, 11593-11598.
37. Mergny, J. L., and Lacroix, L. (2009) UV Melting of G-Quadruplexes, *Curr Protoc Nucleic Acid Chem Chapter 17*, Unit 17 11.
38. Greenfield, N. J. (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions, *Nat Protoc 1*, 2527-2535.
39. Yu, Z., Schonhoft, J. D., Dhakal, S., Bajracharya, R., Hegde, R., Basu, S., and Mao, H. (2009) ILPR G-quadruplexes formed in seconds demonstrate high mechanical stabilities, *J Am Chem Soc 131*, 1876-1882.
40. Qin, Y., Rezler, E. M., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4, *Nucleic Acids Res 35*, 7698-7713.
41. Viglasky, V. (2009) Platination of telomeric sequences and nuclease hypersensitive elements of human c-myc and PDGF-A promoters and their ability to form G-quadruplexes, *FEBS J 276*, 401-409.
42. Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R. A., and Yang, D. (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution, *Nucleic Acids Res 34*, 2723-2735.
43. Matsugami, A., Okuizumi, T., Uesugi, S., and Katahira, M. (2003) Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA, *J Biol Chem 278*, 28147-28153.
44. Dapic, V., Abdomerovic, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J. O., and Bates, P. J. (2003) Biophysical and biological properties of quadruplex oligodeoxyribonucleotides, *Nucleic Acids Res 31*, 2097-2107.
45. Padmanabhan, K., Padmanabhan, K. P., Ferrara, J. D., Sadler, J. E., and Tulinsky, A. (1993) The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer, *J Biol Chem 268*, 17651-17654.
46. Macaya, R. F., Schultze, P., Smith, F. W., Roe, J. A., and Feigon, J. (1993) Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution, *Proc Natl Acad Sci U S A 90*, 3745-3749.
47. John, D. M., and Weeks, K. M. (2000) van't Hoff enthalpies without baselines, *Protein Sci 9*, 1416-1419.
48. Zhou, W. H., and Ying, L. M. (2010) Single Molecule Fluorescence Resonance Energy Transfer and Ensemble Biophysical Characterization of a G-quadruplex Formed in the Promoter of Human Myocyte Enhancer Factor 2D, *Acta Phys-Chim Sin 26*, 1099-1106.
49. Bowen, M. E., Weninger, K., Ernst, J., Chu, S., and Brunger, A. T. (2005) Single-molecule studies of synaptotagmin and complexin binding to the SNARE complex, *Biophys J 89*, 690-702.
50. Mergny, J. L., De Cian, A., Ghelab, A., Sacca, B., and Lacroix, L. (2005) Kinetics of tetramolecular quadruplexes, *Nucleic Acids Res 33*, 81-94.
51. Halder, K., and Chowdhury, S. (2005) Kinetic resolution of bimolecular hybridization versus intramolecular folding in nucleic acids by surface plasmon resonance: application to G-quadruplex/duplex competition in human c-myc promoter, *Nucleic Acids Res 33*, 4466-4474.
52. Kumar, N., and Maiti, S. (2004) Quadruplex to Watson-Crick duplex transition of the thrombin binding aptamer: a fluorescence resonance energy transfer study, *Biochem Biophys Res Commun 319*, 759-767.

53. Suntharalingam, K., White, A. J., and Vilar, R. (2010) Two metals are better than one: investigations on the interactions between dinuclear metal complexes and quadruplex DNA, *Inorg Chem 49*, 8371-8380.
54. Suntharalingam, K., White, A. J., and Vilar, R. (2009) Synthesis, structural characterization, and quadruplex DNA binding studies of platinum(II)-terpyridine complexes, *Inorg Chem 48*, 9427-9435.
55. Fernando, H., Reszka, A. P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A. R., Neidle, S., and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene, *Biochemistry 45*, 7854-7860.
56. He, Y., Neumann, R. D., and Panyutin, I. G. (2004) Intramolecular quadruplex conformation of human telomeric DNA assessed with 125I-radioprobing, *Nucleic Acids Res 32*, 5359-5367.
57. Phan, A. T., Kuryavyi, V., Luu, K. N., and Patel, D. J. (2007) Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K+ solution, *Nucleic Acids Res 35*, 6517-6525.
58. Lim, K. W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D. J., Luu, K. N., and Phan, A. T. (2009) Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers, *J Am Chem Soc 131*, 4301-4309.
59. Qin, Y., and Hurley, L. H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, *Biochimie 90*, 1149-1171.
60. Shirude, P. S., Okumus, B., Ying, L., Ha, T., and Balasubramanian, S. (2007) Single-molecule conformational analysis of G-quadruplex formation in the promoter DNA duplex of the proto-oncogene c-kit, *J Am Chem Soc 129*, 7484-7485.
61. Shirude, P. S., and Balasubramanian, S. (2008) Single molecule conformational analysis of DNA G-quadruplexes, *Biochimie 90*, 1197-1206.
62. Risitano, A., and Fox, K. R. (2003) Stability of intramolecular DNA quadruplexes: comparison with DNA duplexes, *Biochemistry 42*, 6507-6513.
63. Kumar, N., and Maiti, S. (2008) A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes, *Nucleic Acids Res 36*, 5610-5622.
64. Olsen, C. M., Gmeiner, W. H., and Marky, L. A. (2006) Unfolding of G-quadruplexes: energetic, and ion and water contributions of G-quartet stacking, *J Phys Chem B 110*, 6962-6969.
65. Lane, A. N., Chaires, J. B., Gray, R. D., and Trent, J. O. (2008) Stability and kinetics of G-quadruplex structures, *Nucleic Acids Res 36*, 5482-5515.
66. Wallimann, P., Kennedy, R. J., Miller, J. S., Shalongo, W., and Kemp, D. S. (2003) Dual wavelength parametric test of two-state models for circular dichroism spectra of helical polypeptides: anomalous dichroic properties of alanine-rich peptides, *J Am Chem Soc 125*, 1203-1220.
67. Balagurumoorthy, P., and Brahmachari, S. K. (1994) Structure and stability of human telomeric sequence, *J Biol Chem 269*, 21858-21869.
68. Mergny, J. L., Phan, A. T., and Lacroix, L. (1998) Following G-quartet formation by UV-spectroscopy, *FEBS Lett 435*, 74-78.
69. Raghuraman, M. K., and Cech, T. R. (1990) Effect of monovalent cation-induced telomeric DNA structure on the binding of Oxytricha telomeric protein, *Nucleic Acids Res 18*, 4543-4552.
70. Li, W., Wu, P., Ohmichi, T., and Sugimoto, N. (2002) Characterization and thermodynamic properties of quadruplex/duplex competition, *FEBS Lett 526*, 77-81.
71. Gray, R. D., Li, J., and Chaires, J. B. (2009) Energetics and kinetics of a conformational switch in G-quadruplex DNA, *J Phys Chem B 113*, 2676-2683.
72. Rachwal, P. A., Brown, T., and Fox, K. R. (2007) Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes, *Biochemistry 46*, 3036-3044.

73. Prabhu, N. V., and Sharp, K. A. (2005) Heat capacity in proteins, *Annu Rev Phys Chem 56*, 521-548.
74. Lane, A. N., and Jenkins, T. C. (2000) Thermodynamics of nucleic acids and their interactions with ligands, *Q Rev Biophys 33*, 255-306.
75. Chaires, J. B. (1997) Possible origin of differences between van't Hoff and calorimetric enthalpy estimates, *Biophys Chem 64*, 15-23.
76. Oganesian, L., Moon, I. K., Bryan, T. M., and Jarstfer, M. B. (2006) Extension of G-quadruplex DNA by ciliate telomerase, *EMBO J 25*, 1148-1159.

Chapter 5 Molecular and Cell Biology Studies

5.1 Introduction

Since the discovery of the formation of human telomeric G4s in solution in the late 1980s [1], many biochemical and biophysical studies have investigated the spontaneous formation, thermodynamic and kinetic stability, and heterogeneous structures of G4s formed by DNA sequences containing runs of adjacent guanines in solution (reviewed in [2-4]). However, only recently has evidence started to accumulate indicating the presence and possible biological functions of G4s *in vitro* and *in vivo*. As compared to the outnumbered adjacent duplex DNA, G4s are much less stable and can exist only transiently in the human genome. Thus, most of the *in vitro* and *in vivo* evidence about the presence and biological functions of G4s was obtained by indirect approaches, which can be categorized into two groups.

The first approach is to directly probe the formation and location of G4s in living cells by using certain G4-specific antibodies via immunofluorescence staining. For example, by using an anti-parallel telomeric G4 specific antibody (scFv), Lipps *et al*. demonstrated that anti-parallel G4s are only absent in replication bands of ciliated protozoa by immunostaining, which suggested that these telomeric G4s are involved in macronuclear DNA association, and then are resolved during replication in ciliates [5]. Although this approach can provide the most direct evidence, unfortunately, its application is limited by the selectivity of the antibody and the particular properties of the G4 under study. In the case of ciliated protozoa, *Stylonychia* and *Oxytricha* contain up to 10^8 telomeres ($G_4(T_4G_4)_4$) at the end of gene-sized chromosomes (500 bp ~ 20 Kb) in each macronucleus [6], which makes them perfect model organisms to target and observe the formation of telomeric G4s by using the G4-specific antibody (scFv, K_d = 125 pM, recognize anti-parallel telomeric G4 at least 1000-fold more specific than parallel telomeric G4 [5]). Thus, when studying human telomeric G4s or promoter G4s, this approach cannot be used. On one hand, it is very difficult to find an antibody (or a ligand) as specific as the svFv to one type of G4, and on the other hand, a particular type of G4 (either human telomeric G4 or promoter G4) is normally

much outnumbered by adjacent duplex DNA in the human genome, which may also interact non-specifically with the antibody or ligand and give false positive results.

Thus, when studying the biological functions of human cancer-related promoter G4s *in vitro*, other approaches have been used to provide circumstantial evidence. For example, it is possible to stabilise artificially certain G4 structures by the use of specific G4-binding ligands, or one can destabilise them by mutating guanines involved in the formation of the G4, and then investigate subsequently the biological effects upon transcription activity by dual luciferase reporter assays.

The regulatory roles of G4s within the *MYC* promoter represent the most studied example using such approaches. After determining the formation and structures of the G4 in *MYC* promoter by biochemical methods, including DMS footprinting, and polymerase stop assays [7], Hurley's group found that the stabilisation of this promoter G4 by the small molecule TMPyP4 resulted in depressed downstream luciferase transcription, while a G to A mutation that disrupt the formation of this G4 resulted in a 3-fold increase in the basal *MYC* promoter transcription activity [7, 8]. These findings led to the discovery of a G4-mediated regulatory mechanism in the promoter of *MYC*, which provides an alternative target in cancer therapy as well as providing useful basic information on the biological function of G4s (reviewed in Section 1.2.4, *MYC*). By using similar approaches, inhibitory roles of G4s have been preliminarily proposed in the promoters of other cancer-related genes, such as *PDGF-A* [9], *VEGF-A* [10], *BCL-2* [11], and *KIT* [12, 13] (reviewed in Section 1.2.4). By using *in vitro* and *in vivo* approaches, together with characterisation of various G4-interacting proteins, accumulating evidence shows that the formation and stabilisation of these unique four-stranded structures are involved in diverse biological processes, and research into G4 biology is becoming more widespread. As mentioned in Section 1.7, there are emerging pieces of evidence indicating several interesting points about the potential relevance between G4s and cardiovascular gene regulation.

Interestingly, as shown in Chapter 3, bioinformatics studies revealed that the enrichment of G4s in the TRRs of genes active in the heart (80.0 ± 0.3 %, percentage of genes contain at least one G4-forming motifs in their TRRs) is much higher than those found in most of other tissues. Furthermore, G4s are also highly abundant in the core promoter regions of genes involved in almost all cardiac function-related

pathways reported in the KEGG PATHWAY database. Although all these data indicate that G4-mediated regulating mechanism may have also been adopted in cardiac function-related gene regulation, research about possible regulatory roles of cardiac G4s has not been reported yet.

The interaction between promoter G4s with possible TFs, especially those tissue-specific ones, is another important aspect for G4s to perform their biological functions (reviewed in detail in Section 1.7.3). Promoter G4s have been found to be overlap sometimes with binding sites of some ubiquitous TFs, such as Sp1 [14, 15]. Furthermore, the close juxtaposition of G4 with a TF binding site also raises the possibility that G4s act as micro-environments for the recruitment of tissue-restricted factors, such as the mechanism proposed in the MyoD example [16-18]. In the case of human *TnIc* promoter, Sp1 binding sites and potential E-boxes were identified within and around both of the consensus MNSG4 and -80G4-forming sequences. Particularly, the -80G4 was colonized with multiple cardiac-specific TF binding sites, such as the GATA4 [19]. As described in Chapter 4, as compared to G4s found in the promoter of proto-oncogenes, the TnIc MNSG4 and -80G4 were found to possess comparable stability, both of which can be further stabilised by the G4-binding ligand (the di-copper complex) in solution ($\Delta T_{m\ MNSG4} > 15$ °C, and $\Delta T_{m\ -80G4} > 20$ °C, Section 4.3 and 4.8.1).

Thus, according to its importance in the heart, high abundance of G4-forming motifs in its promoter, and comparable stability of these G4s, the promoter of human *TnIc* gene was subject to further investigation *in vitro* using a cell culture model system in order to understand the regulatory functions and biological importance of promoter G4s in cardiac function-related genes.

In this chapter, the regulatory functions of two TnIc G4-forming elements, the MNSG4 and -80G4, were first investigated separately by dual luciferase reporter assays in the absence of G4-binding ligands in HEK293 cells. Then, a group of synthetic G4-binding ligands were screened by measuring the transcription activity changes of the luciferase reporter vector driven by an engineered wild type *TnIc* promoter in HEK293 cells. Finally, the cooperative regulatory functions of human *TnIc* promoter G4s were examined, and possible regulatory mechanisms were proposed.

5.2 Alignment of the human *TnIc* -600/+67 bp promoter region

The human *TnIc* -600/+67 bp promoter region containing both MNSG4 and -80G4 elements was aligned with comparable regions of several mammalian (orangutan, mouse and rat) *TnIc* gene promoters using the Geneious Pro 4.8.5 program (Figure 5.1). This revealed that both MNSG4 and -80G4 elements were conserved in human and orangutan promoters (pairwise identity: MNSG4: 74.2%, and -80G4: 100%). Multiple putative Sp1 binding sites were also observed in both G4 elements. Besides the MNSG4 and -80G4, a conserved -270/-144 bp G/C-rich region was identified in human and orangutan, which may adopt multiple two/three-tetrad G4s. Potential E-boxes in the human promoter, hE1 and hE2 [20], both locating very close to the G4-forming sequences, were also found to be conserved between human and orangutan. Although E-boxes and Sp1 binding sites are only conserved in human and orangutan, multiple other TF binding sites were identified in all four species, such as the CACC-box, GATA binding sites, and A/T-rich region [20]. Although proximal promoter regions (-110 bp ~ TSS) from all four species are relatively conserved, the G4-forming motifs centred at -80 bp are only found in human and orangutan. Furthermore, the MNS G4-forming motifs are also missing in mouse and rat *TnIc* promoters. This may indicate that G4-mediated *TnIc* transcription regulation is specific in primate animals, or that G4 elements outside the promoter region may have regulatory function in rodents.

```
human      -600  TGAGAGAGGAAGTGTGCTGGGGGCC---TGCCTGGACTCCCTTGTCTG-------AGGGA
orangutan  -598  TG---------GTGTGCTGAGTGCCGGTCCCCTGGGCCCACT--GTTGTTTCTCTAAATT
mouse      -583  TGACA-TGCTGCTTTTGGGAAGGACAAGGGCATAGCCTTCCTTGAATGTGATTCAAATAA
rat        -593  TGACA-TGCTACTGTTG-GAGGGACAAGGACATATCCTCCCTTGGATGCGGTTCAAATAA

                                                  MNSG4-1
human      -550  GGGGCTGGGGGCCTGGACCCCTGGGTCTGAGGGAGGAGGGGCTGGGGGCCTGGACT-CCT
orangutan  -549  TTGTCTCTGTGTCTT-ATTTCTTT-TCTCAGTCTCTCGTCCCACCCGACTAGAAATACCC
mouse      -524  AAAGCGCAGGGTC---ACCTCTGTGTGTC---CTGTC--CTCAGTCGGTCTCCAGTGCCA
rat        -535  AGAGCGCAGGGTC---ACCTCTGTGTGTC---CTGTC--CTCTGTTGGTCTCCAGTGCCA

                            MNSG4-2                                 MNSG4-3
human      -491  ----GGGTCTGAGGGAGGAGGGGC-----TGGGGCCTGGACTCTTGGGTCTGAGGGAGGA
orangutan  -497  ACA--GGTGTGGAGGGGCAGGCCACCCCTTCAGTCCTGGACTCCTGGGTCTGAGGGAGAA
mouse      -472  AGAAGGGTG-GGAGCAGCAGGGA------TGGGTCCCG-ATTG-TGGGTTTG-----CAA
rat        -483  AGAAGGGGGTGGGACAAGGGGGGA-----TGGGCCCCGGA-TGGTGGGTTTC-----CAA

                                                  MNSG4-4
human      -440  G-GGGCTG-GGGCCTGGACT-CTTGGGTCTGAGGGAAGAGGAGCTGGGGGCCTG-----G
orangutan  -439  G-GGGCTGAGGGCCTGGACT-CCTGGGTCTGAGGGAGGAGGGGCTGGGG-CCTG-----G
mouse      -426  GTGGGCTTGG---CTGAATTTCACAGGAGTGAAGGATAA-------------------G
rat        -434  GTGGACTTGG---CTGAATTTCACATCAGTGTAGGCTGAGGACTTAGGGTCCTGCACATG
```

(Figure 5.1 to be continued on the next page)

(continued Figure 5.1)

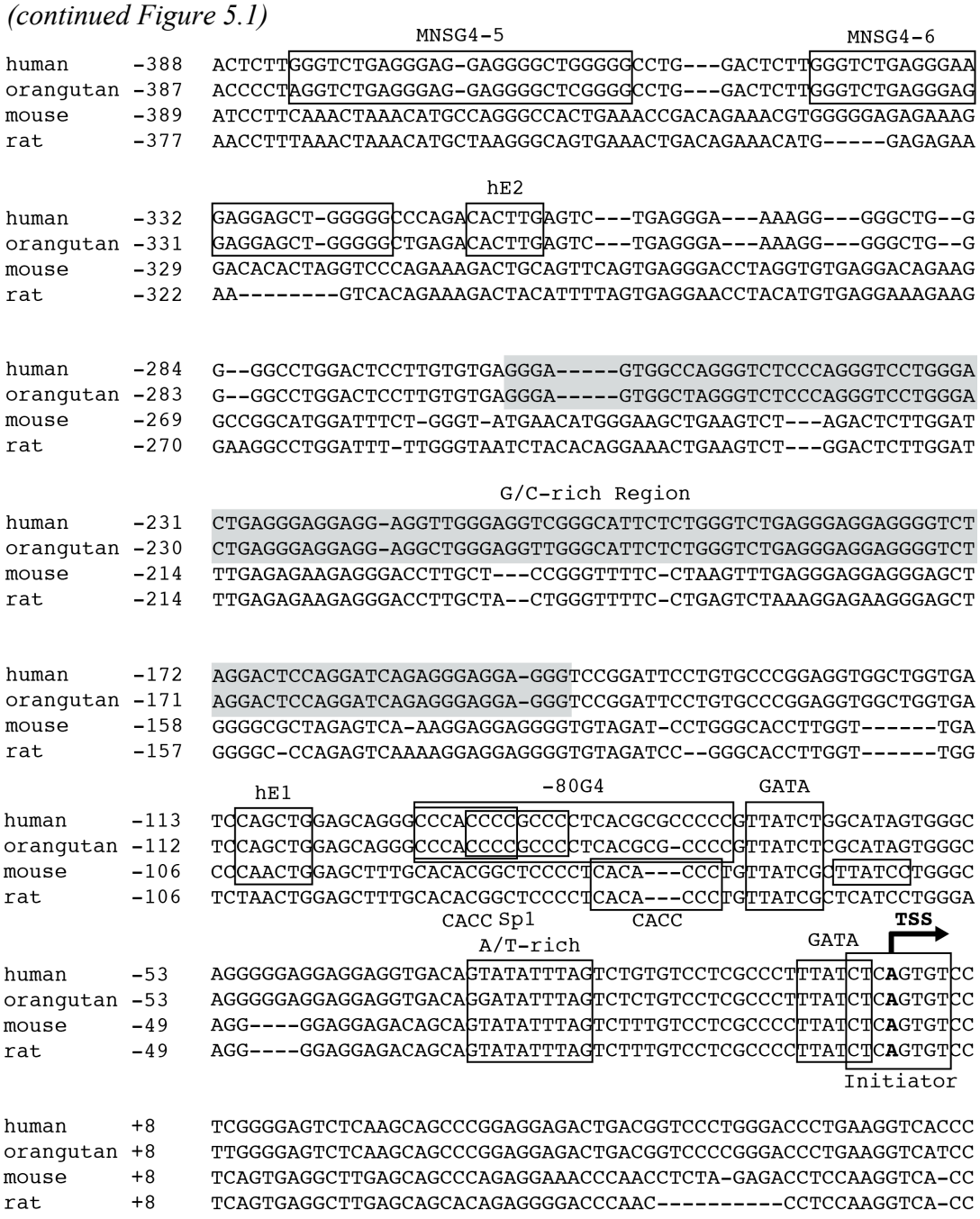


Figure 5.1: Alignment of proximal promoters of human, orangutan, mouse and rat *TnIc* genes.

The transcription start site (TSS, bold **A**) lies within an Initiator element that is conserved across the four species. Conserved GATA and A/T-rich elements are found within 100 bp upstream of the TSS. The -80G4-forming sequences are only conserved in human and orangutan, in which conserved CACC box and Sp1 binding sites are also found. 8 bp upstream to the -80G4-forming sequence, a conserved E-box (hE1: "CAGCTG") is found in human, orangutan, and mouse. Between -531 and -321, 6 repeats of conserved G4-forming minisatellite (MNSG4) are found in human and orangutan. 6 bp downstream to the MNSG4-forming sequence, another conserved E-box (hE2: "CACCTG") is found in human and orangutan. A G/C-rich region is also found between -263 and -146 bp region in human *TnIc* proximal promoter. This conserved G/C-rich region in human and orangutan may adopt noncanonical G4 structures *in vivo*. TFs binding sites were reported by Bhavsar *et al.* [20].

5.3 Alignment of human *TnIc* wild type promoter and constructed promoter

In the pGL3-Basic derived expression vector hTnIc-6MNS(E)-WT (E indicates "engineered" wild type), the recombinated human *TnIc* promote region (-531 ~ +51 bp) driving luciferase expression was generated by three rounds of ligation of MNSG4 consensus sequences, and two rounds of recombination (described in Section 2.3.4). As compared to the wild type human *TnIc* promoter sequence, nucleotide sequence differences may exist, and may result in different cellular activities in subsequent experiments. Thus, the proximal promoter and 5' coding sequences of wild type human *TnIc* (-550 ~ +100 bp) were firstly aligned with the sequences in the promoter region of the luciferase expression vector hTnIc-6MNS(E)-WT. As shown in Figure 5.2, nucleotide differences between the wild type and engineered wild type *TnIc* promoters were mainly found in two regions. The first region containing several nucleotide differences is the MNSG4 element locating between -528 and -319 bp, in which seven nucleotide mismatches were found. This is mainly due to that in the process of constructing the engineered wild type human *TnIc* promoter, I generated a concatemer based on the MNSG4 consensus sequence. On the other hand, since the G-rich hypervariable minisatellite sequences can enhance recombination in the human genome [21] and are highly polymorphic and hypermutable [22], in the MNSG4 region of wild type human *TnIc* promoter, nucleotide polymorphisms among individuals may exist. The second region with nucleotide differences is the region around the E-box, hE2 (between -312 and -307 bp), in which *attB2* and *Pst*I RE sites were introduced in the processes of plasmid construction. Except for potential Sp1 binding sites in the MNSG4 elements, all the aforementioned nucleotide differences will not introduce or eliminate any known TF binding sites in the engineered human *TnIc* promoter as compared to the wild type sequence.

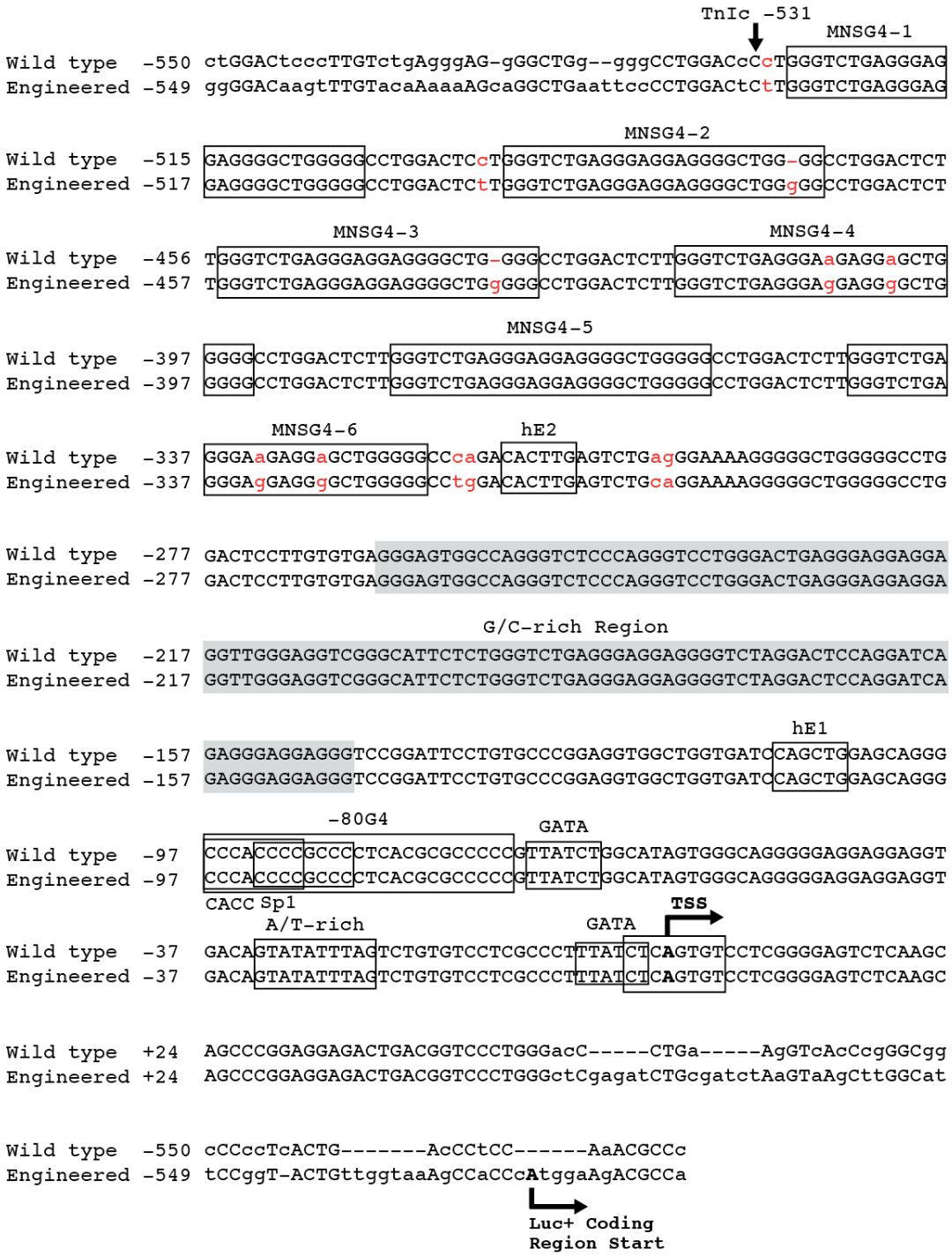


Figure 5.2: Alignment of the proximal promoter of wild type human *TnIc* genes and the engineered *TnIc* promoter.

All possible transcription regulation related elements are highlighted in boxes, and oligonucleotide differences between the wild type (Wild type) and engineered wild type (Engineered) human *TnIc* promoters are shown in red. Generally, in the process of plasmid construction, nucleotides differences were introduced into the engineered wild type promoter in the MNSG4 elements (between -528 and -319 bp, 7 nucleotide mismatches) and around the potential E-box (hE2, between -312 and -307 bp, 4 nucleotide mismatches). TFs binding sites were reported by Bhavsar *et al.* [20].

5.4 Comparison of transcription activities of wild type (WT), engineered wild type (E) and MNSG4-mutated (M) human *TnIc* promoters

The transcription activities of human *TnIc* promoters containing wild type, engineered wild type, or MNSG4-mutated MNS element are determined by dual luciferase reporter assays in the HEK293 cell line. Human *TnIc* promoters containing different MNS elements were inserted into the pGL3-Basic-GW vector as described in Section 2.3.4, giving the plasmids hTnIc-6MNS(E)-WT (E denotes Engineered), and hTnIc-6MNS-M (M denotes Mutated). Transcription activity of these constructed vectors were compared with the vector hTnIc-6MNS-WT, in which the luciferase was driven by -531 ~ +67 bp of wild type *TnIc* promoter region. Sequence differences of the MNSG4s in are shown in Figure 5.3 a. As shown in Figure 5.3 b, the human *TnIc* promoter containing engineered wild type MNS element (consensus MNS) exhibited the highest transcription activity. The transcription activities of both of the *TnIc* promoters containing wild type and mutated MNS elements were ~ 20 % lower than the engineered one. Since the sequences of MNS element in the hTnIc-6MNS-WT and hTnIc-6MNS-M constructs are dramatically different from each other, and the only similarity between them is that the formation of several MNSG4s were disrupted due to either nucleotide polymorphism (MNS4 and MNS6 in hTnIc-6MNS-WT, Figure 5.3 a) or G to A mutations (hTnIc-6MNS-M, Figure 5.3 a), it is very likely that the depressed transcription activities are related to the disruption of MNSG4 formation. Indeed, the transcription activity of in the hTnIc-6MNS(E)-WT construct is the highest, in which all 6 consensus MNSG4s are able to adopt stable G4s (Figure 5.3 a). Another possibility is that in both wild type (hTnIc-6MNS-WT) and mutated (hTnIc-6MNS-M) MNS constructs, multiple less-well conserved Sp1 binding sites (GGGGCTGGGG) were missing as compared to the engineered wild type MNS constructs (hTnIc-6MNS(E)-WT), which results in depressed transcription activity. To further investigate the regulatory roles of the MNSG4, constructs containing different numbers of consensus MNSG4 repeat were generated and corresponding transcription activities were examined by dual luciferase reporter assays in HEK293 cells.

a

Wild type human TnIc MNSG4 element in hTnIc-6MNS-WT:

```
MNS1    GGGTCTGAGGGAGGAGGGGCTGGGGGCCTGGACTCCT
MNS2    GGGTCTGAGGGAGGAGGGGCT-GGGGCCTGGACTCTT
MNS3    GGGTCTGAGGGAGGAGGGGCT-GGGGCCTGGACTCTT
MNS4    GGGTCTGAGGGAAGAGGAGCTGGGGGCCTGGACTCTT
MNS5    GGGTCTGAGGGAGGAGGGGCTGGGGGCCTGGACTCTT
MNS6    GGGTCTGAGGGAAGAGGAGCTGGGGGCCCAGACACTT
```

Engineered wild type human TnIc MNSG4 element in hTnIc-6MNS(E)-WT:

```
Consensus GGGTCTGAGGGAGGAGGGGCTGGGGGCCTGGACTCTT
```

Mutated human TnIc MNSG4 element in hTnIc-6MNS-M:

```
Consensus GAATCTGAGAGAGGAGAGACTGAGAGCCTGGACTCTT
```

b

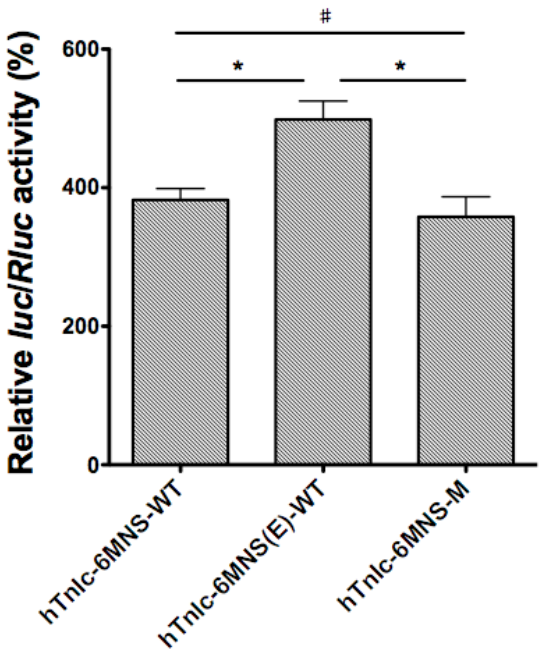


Figure 5.3: Comparison of the sequences of MNSG4 elements in different constructs and their different transcription activities.

(a) Comparisons of the sequences of MNS elements in wild type, engineered wild type and MNSG4-mutated human *TnIc* promoters. Nucleotide differences from the MNSG4 consensus sequences are highlighted by red color. In the MNSG4-mutated promoter, guanines are replaced by adenines to disrupt formation of G4s. **(b)** Transcription activities of human *TnIc* promoters containing wild type MNS element (hTnIc-6MNS-WT), engineered wild type MNS element (hTnIc-6MNS(E)-WT), and mutated MNS element (hTnIc-6MNS-M). The promoter containing 6 repeat of engineered wild type MNSG4 (consensus) exhibited highest transcription activity (~ 20% higher) as compared to the promoters containing 6 repeats of wild type and mutated MNSG4. Transcription activities were indirectly determined by dual luciferase reporter assays, and were presented as mean ± SD of triplicate samples from three individual experiments. Statistically significance difference was assessed using a two-tailed unpaired Student's *t*-test (* $P < 0.05$ and ♯ $P \geq 0.05$, no significant differences).

5.5 Transcription activities of human *TnIc* promoters containing variable number of MNSG4

Since it has been revealed by solution studies that multiple MNSG4s can fold/unfold cooperatively (Figure 4.2 k, l, and Figure 4.8, discussed in Section 4.8.3, TnIc two-repeat MNSG4 part), it would be interesting to study if multiple MNSG4s can also act cooperatively in regulating downstream gene expression. Possible regulatory roles of MNSG4 in the promoter of human *TnIc* were investigated by using dual luciferase reporter assays in the HEK293 cell line. As described in Section 2.3.4, using self-ligation and recombination processes, pGL3-derived luciferase vectors driven by human *TnIc* promoters containing variable repeat of MNSG4 were constructed (Figure 2.4 ~ 2.8). Vectors driven by 6 repeats of MNSG4 (hTnIc-6MNS(E)-WT), 3 repeats of MNSG4 (hTnIc-3MNS(E)-WT), 1 repeat of MNSG4 (hTnIc-1MNS(E)-WT), and no MNSG4 (hTnIc-299-WT) were transfected into HEK293 cells without G4-binding ligands as described in Section 2.3.12. Transcription activities of these human *TnIc* promoters were indirectly measured by the ratios between firefly luciferase luminescence and *Renilla* luciferase luminescence. As shown in Figure 5.4, as compared to the engineered wild type containing six repeats of consensus MNSG4, human *TnIc* promoters containing three repeats, one repeat, or no repeat exhibited similar activities in driving downstream luciferase transcription ($P \geq 0.05$). This indicates that in the absence of G4-binding ligands, both of the MNSG4 and Sp1 binding sites in the MNSG4 element do not appear to be involved in regulating downstream gene transcription, at least in HEK293 cells. However, this is in contradiction to the previous findings that the disruption of MNSG4 formation may depress the transcription activity of human *TnIc* promoter (Figure 5.3, Section 5.4). This is likely due to that the polymorphic nucleotides in the MNSG4 of the wild type vector (hTnIc-6MNS-WT) and the multiple mutation sites in the MNSG4-mutated vector (hTnIc-6MNS-M) introduced or disrupted binding sites of certain unknown TF(s) that functions in HEK293 cells.

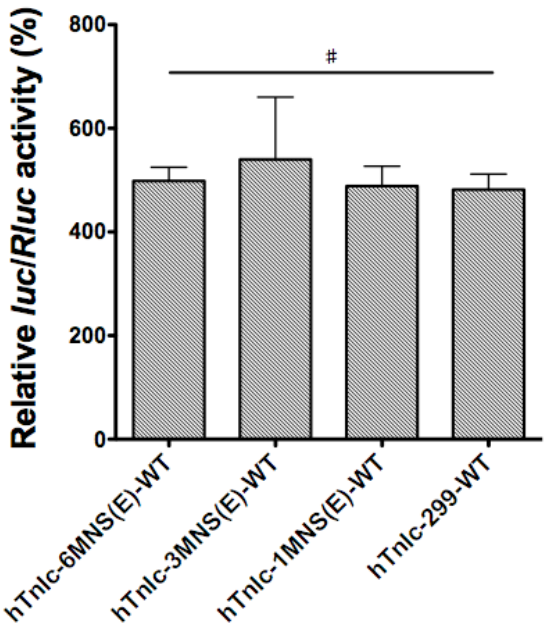


Figure 5.4: Comparison of the transcription activities of human *TnIc* promoters containing variable number of MNSG4.

Transcription activities of human *TnIc* promoters containing 6 repeats of consensus MNSG4 (hTnIc-6MNS(E)-WT), 3 repeats of consensus MNSG4 (hTnIc-3MNS(E)-WT), 1 repeat of consensus MNSG4 (hTnIc-1MNS(E)-WT), and no repeat of MNSG4 (hTnIc-299-WT). No transcription activity difference was found among these human *TnIc* promoters. Transcription activities were indirectly determined by dual luciferase reporter assays (ratio between firefly luciferase and *Renilla* luciferase), and were presented as mean ± SD of triplicate samples from three individual experiments. Statistically significance difference was assessed using a two-tailed unpaired Student's *t*-test (# $P \geq 0.05$, no significant differences).

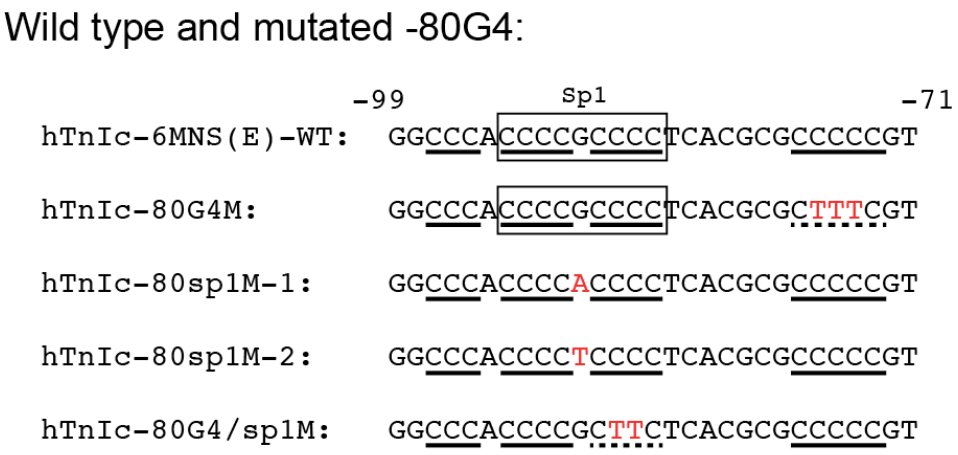
5.6 Transcription activities of human *TnIc* promoters with wild type and mutated -80G4

Regulatory roles of the -80G4 together with a conserved Sp1 binding site in the promoter of human *TnIc* were also investigated by using dual luciferase reporter assays in the HEK293 cell line. As described in Section 2.3.4, site-directed mutations were introduced into the luciferase expression vector driven by the engineered wild type human *TnIc* promoter (hTnIc-6MNS(E)-WT) via *Dpn*I mediated site-directed mutagenesis. This gave vectors containing mutated -80G4-forming sequence but with an intact Sp1 binding site (hTnIc-80G4M), intact -80G4-forming sequence but mutated Sp1 binding site (hTnIc-80sp1M-1/2), or mutated -80G4-forming sequence and Sp1 binding site (hTnIc-80G4/sp1M) (Figure 5.5 a). Vectors were transfected into HEK293 cells in the absence of G4-binding ligands as described in Section 2.3.12,

and transcription activities of these human *TnIc* promoters were indirectly measured by the ratios between firefly luciferase luminescence and *Renilla* luciferase luminescence. As shown in Figure 5.5 b, human *TnIc* promoter containing -80G4 with disrupted G4-forming sequence but intact Sp1 binding site (hTnIc-80G4M) had similar transcription activity as the engineered wild type promoter (hTnIc-6MNS(E)-WT). However, when the Sp1 binding site was mutated (hTnIc-80sp1M-1/2), transcription activity of the human *TnIc* promoter was decreased by ~ 25 %. Furthermore, when both -80G4-forming sequence and Sp1 binding site were mutated (hTnIc-80G4/sp1M), transcription activity was further decreased to 50% as compared to the engineered wild type promoter. This may relate to the further mutated Sp1 binding site in vector hTnIc-80G4/sp1M as compared to that in vectors hTnIc-80sp1M-1/2. These results indicate that similar to what was observed in the MNSG4 experiments, the -80G4 is not directly involved in regulating downstream luciferase transcription in HEK293 cells. However, the Sp1 binding site seems to be critical in determining downstream gene expression. This is similar to the situation where an Sp1 binding site found in the G4-forming sequence (KIT1) in the promoter of human *KIT* gene appears to be critical in determining the maximal activity of the human *KIT* promoter [23].

The above results implied that both MNSG4 and -80G4 are not directly involved in regulating downstream gene transcription in HEK293 cells. One possible reason might be that functions of these G4-forming motifs in human cardiomyocytes are dependent on the stabilising effects by certain cardiac-specific TFs or proteins, which are absent in the HEK293 cells. To further investigate this possibility, a group of G4-binding ligands were screened, and their cellular effects were determined by dual luciferase reporter assays.

a



b

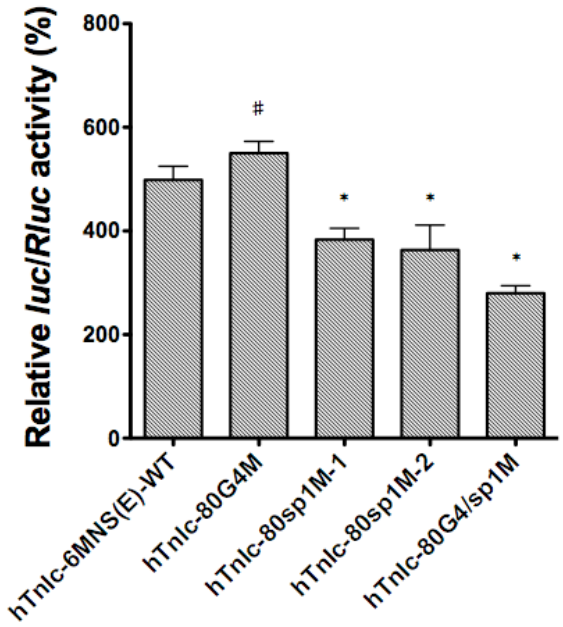


Figure 5.5: Comparison of the sequences and transcription activities of human *TnIc* promoters with wild type and mutated -80G4.

(a) Sequences of wild type and mutated -80G4 (between -99 and -71 bp to the TSS of human *TnIc*) in different luciferase expression vectors. Sp1 biding sites are shown in black boxes. Cytosines involved in formation of -80G4 on the template strand are highlighted by underlines. Mutilation sites are shown in red, in which those disrupt the formation of -80G4 are highlighted by dashed underlines. **(b)** Transcription activities of human *TnIc* promoters containing different -80G4s. Mutation of -80G4 alone (hTnIc-80G4M) did not change the transcription activity, while the mutation of Sp1 binding site (hTnIc-80sp1M-1/2) led to ~ 25% depression in the transcription activity. When both G4 formation and Sp1 binding sites were mutated (hTnIc-80G4/sp1M), 50% decrease of the transcription activity was found. Transcription activities were indirectly determined by dual luciferase reporter assays (ratio between firefly luciferase and *Renilla* luciferase), and were presented as mean ± SD of triplicate samples from three individual experiments. Statistically significance differences between the artificial wild type vector and others were assessed using a two-tailed unpaired Student's *t*-test (* $P < 0.05$ and # $P \geq 0.05$ no significant differences).

5.7 G4-binding ligands screened in the HEK293 cell line

In order to investigate if stabilisation of G4-forming motifs in the human *TnIc* promoter could change the transcription activities of the promoter, a group of synthesised G4-binding ligands were screened *in vitro*. The ligands were designed and synthesised by Prof Ramon Vilar's group, among which the metal-containing compounds were most promising (compound 1 - 3 as illustrated in Figure 5.6 a). Dose-response curves of compound 1 - 3 were generated by measuring the firefly (vector hTnIc-6MNS(E)-WT) and *Renilla* luciferase ratios of lysates of transfected HEK293 cells incubated in the presence of ligands at different concentrations. Vector hTnIc-6MNS(E)-WT was used because it contains the greater number of G4 repeats, which was predicted to give the most significant response if the ligand-stabilisation of G4s is involved in determine the promoter activity. As shown in Figure 5.6 b, compound 1, which has a single platinum-conjugated moiety, shown a weak activating effect towards downstream gene transcription that the promoter activity was increased by 2 times at 100 μM concentration (EC_{50} = 812.1 μM). Compound 2 contains two linked platinum-conjugated moieties, which are able to interact with the G4 via a combination of π-π end-stacking interactions (with the terminal guanine tetrad) and electrostatic interactions (with the loops and grooves of quadruplexes) [24]. This compound exhibited modest activating effects towards downstream luciferase transcription that it increased the promoter activity by more than 3 times at 100 μM concentration (EC_{50} = 8.6 μM). By replacing platinum with copper, the third compound (compound 3) demonstrated most significant activation effect, which elevated the promoter activity by approximately 5-fold at 100 μM concentration (EC_{50} = 0.6 μM). These results agreed well with what have been observed in biophysical characterisations in solution. For example, FID (Fluorescent Intercalator Displacement) assays performed by Prof Ramon Vilar's group have revealed that compound 1 has the lowest affinity and conferred the weakest G4-stabilising ability to both MNSG4 and -80G4 (compound 1: $^{MNS}DC_{50}$ = 0.76, $^{-80G4}DC_{50}$ = 0.55, $\Delta T_{m\ MNSG4}$ = 6.9 °C), while compounds 2 and 3 have modest (compound 2: $^{MNS}DC_{50}$ = 0.39, $^{-80G4}DC_{50}$ = 2.22, $\Delta T_{m\ MNSG4}$ = 11.0 °C) and highest (compound 3: $^{MNS}DC_{50}$ = 0.31, $^{-80G4}DC_{50}$ = 0.19, $\Delta T_{m\ MNSG4}$ = 21.7 °C) affinity and G4-stabilising ability. Together, these observations suggest that the compound 3 (the di-copper complex) gave the best performance in living cells.

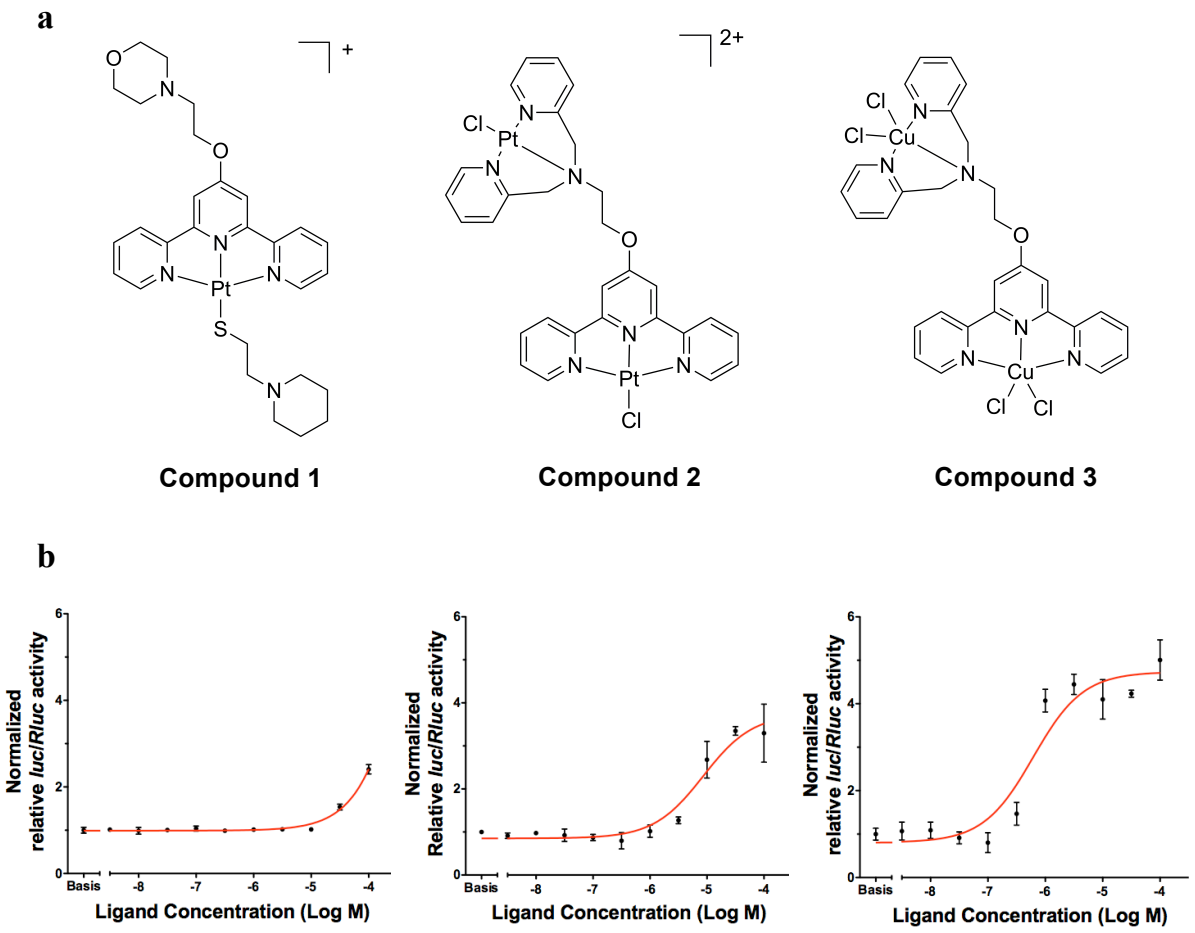


Figure 5.6: Structures of G4-binding ligands, and corresponding dose-response curves.

(a) Structures of compound 1 (Pt complex), 2 (Pt-Pt complex), and 3 (Cu-Cu complex) that were used in subsequent dose-response investigations. **(b)** Dose-response curves of compound 1, 2, and 3. The compound 1 only slightly activated downstream luciferase expression (EC_{50} = 812.1 μM), the compounds 2 moderately activated downstream luciferase expression (EC_{50} = 1.6 μM), and the compound 3 significantly activated downstream gene expression (EC_{50} = 0.6 μM). Experiments were carried out by measuring the firefly (hTnIc-6MNS(E)-WT) and *Renilla* luciferase ratio changes in the presence of ligand with different concentrations. Detailed protocol was described in Section 2.3.13. The firefly and *Renilla* luciferase ratios measured in the presence of ligand were presented as mean ± SD of triplicate samples from three individual experiments, and were normalized to the blank control (the firefly and *Renilla* luciferase ratio measured without ligand, shown as "Basis" point). Significant difference was observed between individual dose-response curves ($P < 0.05$, as assessed using a two-tailed unpaired Student's *t*-test).

5.8 Other possible mechanisms that "activate" the transcription activity of human *TnIc* promoter via G4-binding ligands

Although cellular observations were in agreement with the biophysical characterisations suggesting that the compound 3 (the di-copper complex) has the best performance in interacting with human TnIc G4s, the actual activation mechanism *in vitro* may not result from direct interactions between the compound and promoter G4s. When measuring the dose-response curves of compounds 2 and 3, the increased firefly and *Renilla* luciferase ratios were actually caused by a combination of increased firefly luciferase luminescence, and slightly decreased *Renilla* luciferase luminescence. To understand the actual reason for this phenomenon, several possibilities were postulated and investigated.

All three compounds contain three linked hexatomic rings, which also exist in the substrate of *Renilla* luciferase, the coelenterate-luciferin (coelenterazine) (compare Figure 5.6 a with Figure 2.9). Thus, one possibility is that these compounds may interact with *Renilla* luciferase as antagonistic blockers, and hence exaggerate the activation effects of these compounds. To investigate this possibility, cell lysates containing firefly and *Renilla* luciferases were incubated at 37 °C with high concentration (10 μM) of compounds 1 - 3 respectively, and luciferase activity ratios were measured. As shown in Figure 5.7 a, when the cell lysates were incubated with compounds at 37 °C for 1 hour, no changes were observed as compared to that measured in the absence of compounds. Furthermore, when incubated overnight, decreased firefly and *Renilla* luciferases ratios were observed due to the fast degradation of the firefly luciferases. Thus, the antagonistic effects of compounds towards *Renilla* luciferase were eliminated.

The other possibility is the cytotoxicity of these compounds. It is possible that the overall protein production decreased when compounds concentration increased, or that at high concentration, the compounds could induce cell breakage, and lead to release of protein products into the medium. In both cases, the *Renilla* luciferase in the cells would decrease as the cytotoxicity increase. Cytotoxicity of these compounds were measured by CytoTox96® non-radioactive cytotoxicity LDH (lactate dehydragenase) assays, which indicates the integrity of cell membranes. As shown in Figure 5.7 b, compound 1 was not cytotoxic even at 10 μM, while for compounds 2

and 3, significant cell leakage was observed (increased LDH in the medium as compared to the overall LDH). Thus, it is very likely that the decreased *Renilla* luciferase luminescence at high concentration of ligands was related to the releasing of *Renilla* luciferase into the medium. In the future work, mRNA levels of firefly and *Renilla* luciferase need to be measured to give direct evidence of the activation effects induced by these compounds.

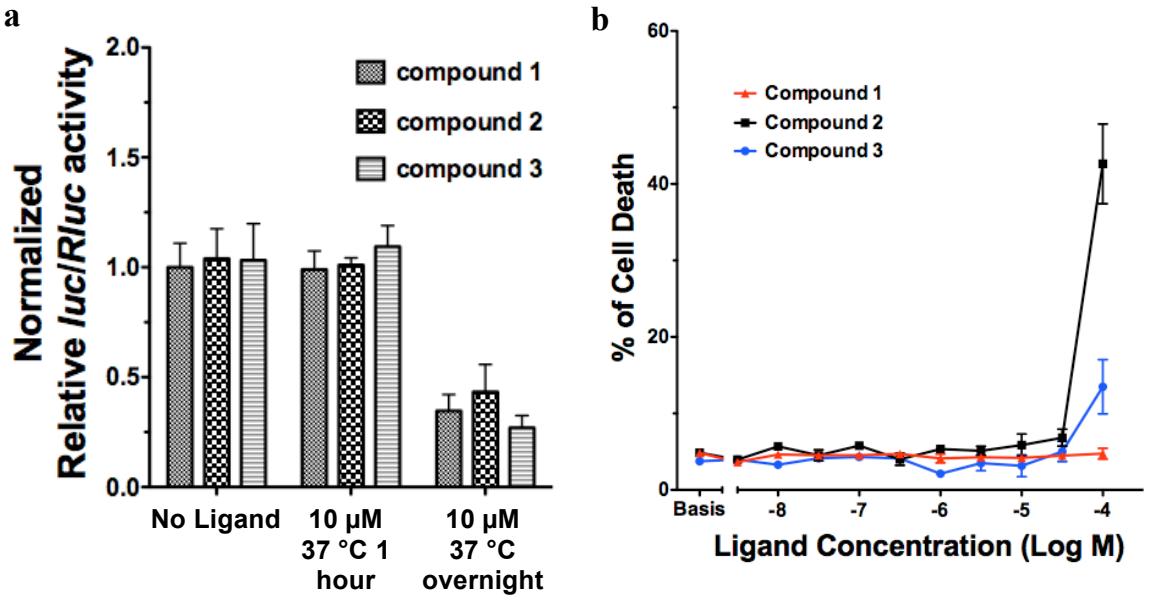


Figure 5.7: Antagonistic effects and cytotoxicity of compounds 1 - 3.
(a) Cell lysates containing firefly and *Renilla* luciferases were incubated with 10 μM of compounds 1 - 3 for 1 hour and overnight at 37 °C. No antagonistic effect was found in all three compounds. All firefly and *Renilla* luciferase ratios were presented as mean ± SD of triplicate samples from three individual experiments, and were normalized to the ratios measured in the absence of ligand. **(b)** Cytotoxicity of compounds 1 - 3 measured by CytoTox96® non-radioactive cytotoxicity LDH assays according to the manufacturer's manual. In each measurement, HEK293 cells were incubated with compounds with different concentration for 24 hours before carrying out the LDH assays. All data points were presented as mean ± SD of triplicate samples from three individual experiments.

5.9 Positive regulatory functions of human *TnIc* promoter G4s

To investigate the regulatory functions of *TnIc* promoter G4s, the firefly and *Renilla* luciferase ratios of different vectors were measured by incubating with and without the G4-binding ligand compound 3 (the di-copper complex). An overall G4 mutant, hTnIc-6MNS/-80G4-M, was constructed from the hTnIc-6MNS-M by site-directed mutagenesis (the primer pair -80G4-M-F/R was used, Table 2.2) to mutate the -80G4-

forming sequence as well as the Sp1 binding site within. These luciferase vectors were first transfected to HEK293 cells in the absence of Cu-Cu complex for 24 hours prior to cell lysis. As shown in Figure 5.8 a, all vectors with MNSG4 mutations shown similar transcription activities as compared to the engineered wild type one (hTnIc-6MNS(E)-WT). The -80G4 mutation (with an intact Sp1 binding site) disrupting the G4 formation slightly elevated the transcription activity of *TnIc* promoter. The overall G4 mutant, hTnIc-6MNS/-80G4-M, exhibited lowest transcription activity. The decreased luciferase expression from this vector may result from a combination of multiple mutation sites in MNSG4, -80G4, and the Sp1, which might have introduced certain inhibitory factor binding sites.

As mentioned previously that since *TnIc* is exclusively expressed in cardiomyocytes, the formation and stabilisation of these promoter G4s may be associated with certain cardiac-specific proteins or TFs, which are absent in HEK293 cells. Since it has been reported that short proximal promoter of human *TnIc* is enough to confer cardiac-specific expression in human cells [25], it is possible that the luciferase expression level of these vectors without ligands is just the basal expression level in HEK293 cells. To test this speculation, luciferase assays were repeated in the presence of 100 μM compound 3. As shown in Figure 5.8 b, when incubated with compound 3 for 24 hours before cell lysis, the expression activity of luciferase from all vectors increased differently. Generally, the transcription activity of human *TnIc* promoter gradually decreased when the number of MNSG4 repeat decreased, while the lowest transcription activities were observed in vectors with mutated -80G4. When Pt-complex was used as a control, no significant change of luciferase expression was observed.

To further discriminate activation effects of MNSG4 and -80G4, dose-response curves of different luciferase vectors treated with compound 3 were generated. As shown in Figure 5.8 c, the strongest activation effect was observed in the vector with wild type MNSG4 and -80G4 (hTnIc-6MNS(E)-WT), in which downstream luciferase expression was increased for 6.5 times in the presence of 3.2 μM Cu-Cu complex. When the -80G4 is intact, the activation effects of downstream luciferase decreased proportionally with the decrease of repeat number of MNSG4 (hTnIc-6MNS(E)-WT > hTnIc-3MNS(E)-WT > hTnIc-1MNS(E)-WT > hTnIc-299-WT). However, when the -80G4 is mutated (hTnIc-80G4M, hTnIc-6MNS/-80G4-M),

lowest activation effects were observed again. These results suggest that the -80G4, locating in a region with high location significance as mentioned in Chapter 3, may operate as a gatekeeper in regulating downstream luciferase expression in HEK293 cells; the MNSG4, on the other hand, acts as a fine-tuner in the regulation process when the -80G4 is intact. It is likely that both types of G4s orchestrate the gene expression of cardiac troponin I in the heart.

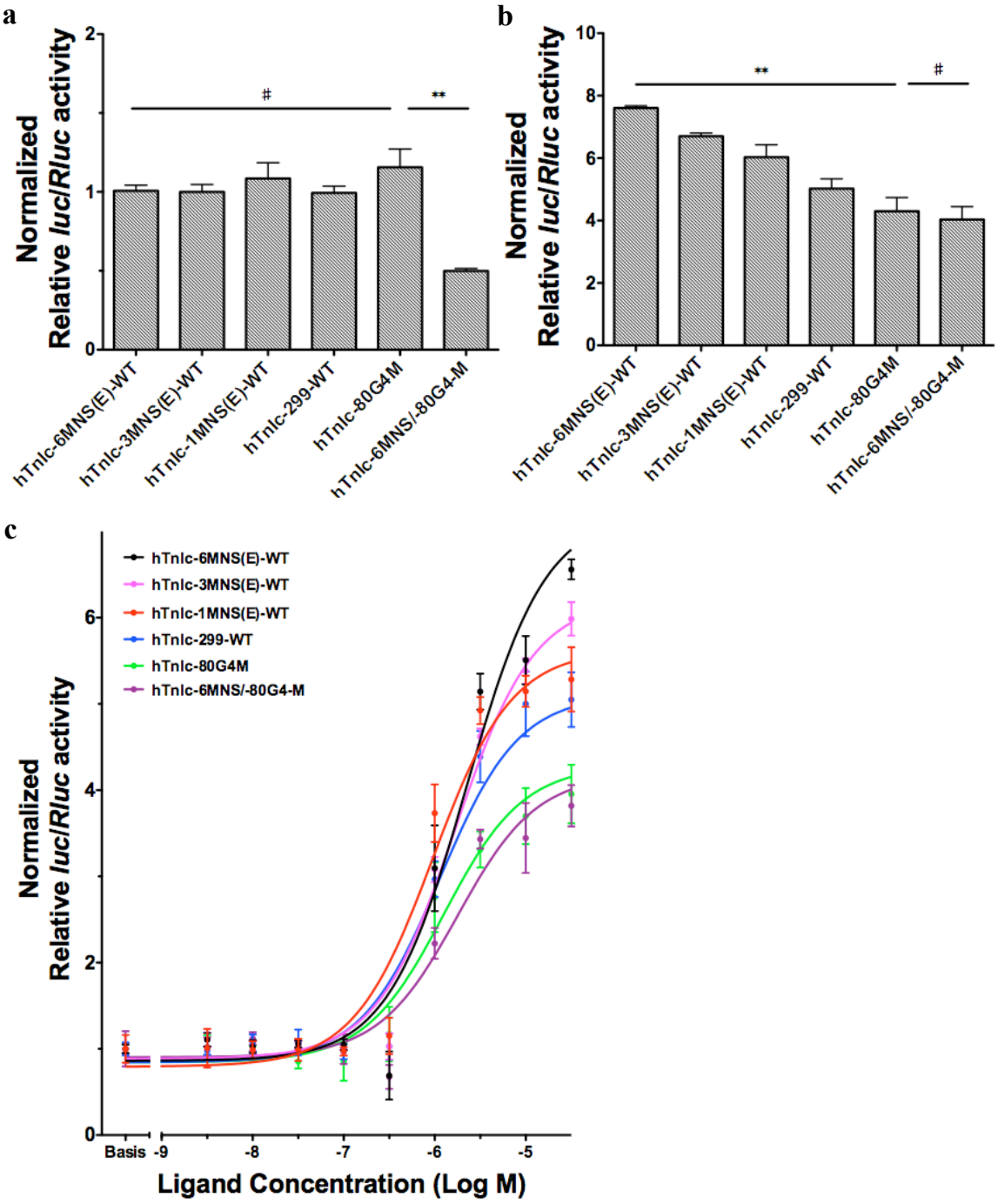


Figure 5.8: Transcription activity changes of human *TnIc* promoters in the presence of compound 3 (the di-copper complex).

(a) Transcription activity of constructed human *TnIc* promoters in the absence of compound 3. **(b)** Transcription activity of constructed human *TnIc* promoters in the presence of 100 μM compound 3. **(c)** Dose-response curves of compound 3 with different luciferase vectors. Experiments were carried out by measuring the firefly and *Renilla* luciferase ratio changes in the presence of ligand with different concentrations. Detailed protocol was described in Section 2.3.13. The firefly and *Renilla* luciferase ratios measured in the presence of ligand were presented as mean ± SD of triplicate samples from three individual experiments, and were normalized to the blank control (the firefly and *Renilla* luciferase ratio measured without ligand, shown as "Basis" point). Statistically significance differences were assessed using a two-tailed unpaired Student's *t*-test (** $P < 0.01$, * $P < 0.05$ and # $P \geq 0.05$, no significant differences).

5.10 Discussion

5.10.1 MNSG4 and -80G4 elements are only conserved in the promoters of *TnIc* in the genomes of primate animals

As shown in Section 5.2, both TnIc MNSG4 and -80G4 were only found to be conserved in primate animals. Particularly, the conserved MNSG4 element with varied number of repeat is identified in the promoter region, and the fifth intron of *TnIc* in genomes of human, chimpanzee, and rhesus (in orangutan, sequences of the fifth intron of *TnIc* are not available). However, both of the G4 elements are not conserved in the rodent genomes. The weak evolutionary conservation of both MNSG4 and -80G4 may impede the original hypothesis that they are involved in regulating downstream gene transcription. Indeed, most of promoter G4s with potential regulatory roles are expected to be conserved in different species [14, 26, 27], such G4s indentified in the promoter of *KIT* [28], and in the 5' UTR of *NRAS* [29], where the G4-forming motifs are highly conserved in the human and rodent genomes. However, in some cases, conservation among species is not a necessary precondition to evaluate the regulatory roles of a G4-forming motif. For example, the *MYC* promoter G4, which has been widely accepted as a promising target in cancer therapy, is not conserved in rodent. Besides the *MYC* example, circumstantial evidence also supports the original hypothesis.

The first evidence comes from the stability of the MNSG4 in the human genome. Minisatellites, particularly for GC-rich ones, have been associated with genome fragile sites and are close to many recurrent translocation breakpoints (reviewed in [30, 31]). For example, G4-forming minisatellite CEB1 are able to induce genomic instability in the absence of a DNA helicase Pif1 in the budding yeast [32]. Furthermore, hypervariable and hypermutatable minisatellites correlate with various diseases. For example, allelic variation of minisatellite in IDDM2 locus may play an important role in determining the susceptibility to type I diabetes [33]. A linkage between variability of a minisatellite found at the 3' of the apolipoprotein B gene and coronary heart disease has also been demonstrated [34]. However, although multiple mutation sites within exons and introns of human *TnIc* have been revealed to associate with hypertrophic cardiomyopathy (HCM), and thus human *TnIc* has been

regarded as a HCM gene [35, 36], surprisingly, no mutation has been reported in the minisatellite regions yet. This genomic stability of TnIc MNSG4 possibly indicates that the heart has evolved certain mechanisms to maintain its stability or utilize even it as regulatory elements by certain human cardiac-restricted proteins.

The other evidence is the previous finding showing that the full length of -80G4 forming sequence is required for maximal transcription activity of human *TnIc* minimal promoter [19]. As shown in Figure 5.1, the -80G4 is surrounded by multiple TF binding sites, such as E-box, GATA-box, and A/T-rich element. All these TF binding sites are conserved in both human and rodents. Indeed, previous studies revealed that TFs such as GATA-4, MEF-2, and Oct-1 are able to bind to the *TnIc* promoter and hence regulate its transcription activity in human [19], rat [37], and mouse [38]. However, the -80G4, which is only found in human *TnIc* promoter, was found to be able to interact with Sp1, Sp3 (similar as the NHEIII region in promoter of *MYC*), as well as other unidentified factors, possibly cardiac-restricted ones [19]. In the three runs of guanines in 3' of -80G4 (C-rich sequences on the coding sequence: -97 bp CCCACCCCGCCC -86 bp), the CACC/Sp1 binding site is critical in determining the maximal transcription activity of human *TnIc* promoter. Bhavasa *et al.* found that the mutation of CACC/Sp1 binding site located in this region results in 85 % reduction of transcription activity of human *TnIc* promoter [19]. For the first run of three guanines at 5' of the -80G4, it is conserved across all four species (-76 ~ -74 bp). These data suggest that all four runs of guanines of this unique GC-rich region are important with regard to its regulatory roles and evolutionary conservation.

Thus, the genomic stability of MNSG4 element and the critical regulatory functions of -80G4 element in the promoter of human *TnIc* suggest that the transcription of *TnIc* might be regulated in a species-restricted way by these G4-forming elements in human.

5.10.2 HEK293 cell line vs. rat/mouse cardiomyocytes

If the transcription of *TnIc* was regulated differently in human and rodent heart, one inevitable question is that what is the appropriate cell model to study the regulatory roles of these G4 elements. Unlike cancerous cells, adult cardiomyocytes are highly

differentiated and are unable to proliferate. Thus, although the neonatal cardiomyocytes prepared from small animals such as rat [20, 39] and mouse [40] are commonly used in researches about cardiovascular diseases, in this project, besides the practical and ethical restrictions, the HEK293 cell line was chosen due to several reasons. The first reason is based on the assumption that G4-mediated transcriptional regulation of human *TnIc* is tissue- and species-restricted (exclusively in human cardiomyocytes). Thus, in studying the regulatory roles of these G4 elements, the HEK293 cell line is not worse than the neonatal rat/mouse cardiomyocytes, since both of them do not possess the mechanism to utilize these G4 elements. The second reason is the simplicity of the HEK293 cell line. As shown in previous studies, although it is very likely that the maximal transcription activity of *TnIc* promoter is regulated in a species-restricted manner [19], the basal transcription involves several common cardiac-specific TFs, such as GATA-4, MEF-2, and Oct-1 in human and rodents [5, 37, 38]. If the formation of G4 elements is truly involved in determining the tissue- and species-restricted transcription of human *TnIc*, in neonatal rat/mouse cardiomyocytes, the transcription activity changes of human *TnIc* promoter induced by the ligand-stabilised G4s wouldn't be as apparent as those in HEK293 cells, in which only the lowest *TnIc* promoter activity can be found without G4-binding ligands. Thus, to answer the question that whether or not the formation of G4 elements can change the transcription activity of human *TnIc* promoter, HEK293 cell line was firstly used in this study, while the cardiomyocytes would be used in future experiments.

5.10.3 Interaction between compound 3 and TnIc G4 elements in living cells

In this study, the regulatory roles of G4 elements in human *TnIc* promoter were mainly measured by comparing the transcription activities of engineered human *TnIc* promoter in HEK293 cells with and without the G4-binding ligand by dual luciferase reporter assays. This is based on the assumption that the G4-binding ligand, compound 3, is able to directly interact with these TnIc G4 elements in living cells, and subsequently induce transcription activity changes. However, similar to most studies published so far, direct evidence indicating the interaction between G4-binding ligands and G4s in living cells are extremely difficult to obtain. In this study,

such direct evidence has not yet been obtained, and thus causing certain ambiguities in interpreting the ligand-induced cellular responses. However, several pieces of indirect evidence support the conclusion that the positive regulatory roles of these *TnIc* promoter G4s relate to the stabilisation of these elements by the ligand. The first evidence is the selectivity of the ligand to the TnIc G4 elements as compared to the ds DNA. Previous study showed that the compound 3 exhibits a relatively strong selectivity to G4s as compared to ds DNA (~ 100-fold [24]), although its selectivity among different G4s is relatively low (FID results (μM): $^{htelo}DC_{50}$ = 0.28, $^{cmyc}DC_{50}$ = 0.17 $^{MNS}DC_{50}$ = 0.31, $^{-80G4}DC_{50}$ = 0.19). These features of compound 3 removed the doubts that this ligand may interact with dsDNA, and change the transcription activity of *TnIc* promoter in HEK293 cells indirectly (such as the inhibitory effects of Quarfloxin towards MYC [41, 42]). The second evidence is the preference of the ligand between the TnIc MNSG4 and -80G4. In Chapter 4, biophysical studies of G4-ligand interactions indicated that the Cu-Cu complex has better performance when interacting with the -80G4 (Section 4.7, and FID results mentioned in Section 5.7). Similarly, in cell studies, activation effects from the ligand-stabilised -80G4 are more prominent. Finally, the prevalent activating effect from -80G4 in regulating downstream gene expression is in good agreement with our hypothesis in the bioinformatics studies about location significance of promoter G4s (G4s found in the region with high distribution probability are more important). Thus, although additional experiments are necessary in the future to fully clarify the interaction between compound 3 and TnIc G4 elements in living cells, it is very likely that the activation effects of these G4 elements are directly induced by interacting with the G4-binding ligand.

5.10.4 Possible mechanisms of G4-mediated human *TnIc* promoter activation

G4/Sp1-mediated mechanism

Belonging to Sp family, Sp1 and Sp3 are ubiquitously expressed C_2H_2 zinc finger factors and are involved in regulating large number of both housekeeping and tissue-specific genes [43, 44]. Sp1 and Sp3 are also involved in regulating many cardiac genes such as α-cardiac actin [45], and cardiac troponin T [46]. Although Sp1 and Sp3 both contain 3 conserved Cys2His2 zinc fingers in their DNA-binding domain,

and are highly homological to each other (> 90 % sequence homology), their actual functions vary depending on the particular promoter involved [46, 47]. In most cases, Sp1 acts as a transactivator, while Sp3 functions as a repressor or weak activator. For example, in the case of *MYC*, whose promoter contains five Sp1 binding sites, binding of Sp1 to the promoter is involved in serum-induced transcription activation of *MYC* [48], and cotransfection of Sp1 transactivates the *MYC* promoter in both mammalian and insect cells [49]. On the contrary, overexpression of Sp3 in these cells inhibits the Sp1-mediated transactivation by competing with Sp1 binding [49]. More interestingly, in promoters containing multiple Sp1 binding sites, such as the *MYC* promoter, Sp1 firstly forms tetramers at the binding sites, assembles higher-order complexes, and exerts transcriptional synergism through direct protein-protein interactions, while the Sp3 mainly act as monomers [50, 51]. In the field of G4 studies, the Sp1 binding site closely relates to G4-forming motifs. Several independent bioinformatics studies have revealed that the colonization of G4-forming motifs and this zinc finger TF bindings sites are highly conserved [14, 15, 52]. Furthermore, Sp1 is able to bind both canonical duplex DNA and G4s in solution with comparable affinity [53]. All these data suggest that a co-regulatory role between G4 and Sp1.

Similar to *MYC* promoter, the human *TnIc* promoter also contains multiple Sp1 binding sites overlapped with both MNSG4 and -80G4 elements. Both Sp1 and Sp3 can bind to the CACC/Sp1 binding site (overlapped with the -80G4) in human *TnIc* promoter, and the binding of Sp1 to this region can strongly transactivate the promoter in NIH3T3 cells [54]. Furthermore, the CACC/Sp1 binding site is critical in determining the maximal transcription activity of minimal human *TnIc* promoter in neonatal rat cardiomyocytes [19]. In this study, it is also found that in the absence of G4-binding ligand, the basal transcription activity of human *TnIc* promoter is mainly determined by this Sp1 binding site in the -80G4 element. As shown in Figure 5.4, 5.5 b and 5.8 a, the disruptions of both MNSG4 and -80G4 formation do not change the basal transcription activity of human *TnIc* promoter, as long as CACC/Sp1 binding site within -80G4 element is intact, while the mutation of this CACC/Sp1 binding sites results in more 20 % repression (Figure 5.5 b). However, in the presence of G4-binding ligand, the transcription activation is dependent on both intact -80G4 and Sp1

binding site in it (Figure 5.8 b). Thus, by putting all these findings together, a G4/Sp1-mediated transcription activation mechanism of human *TnIc* promoter was proposed.

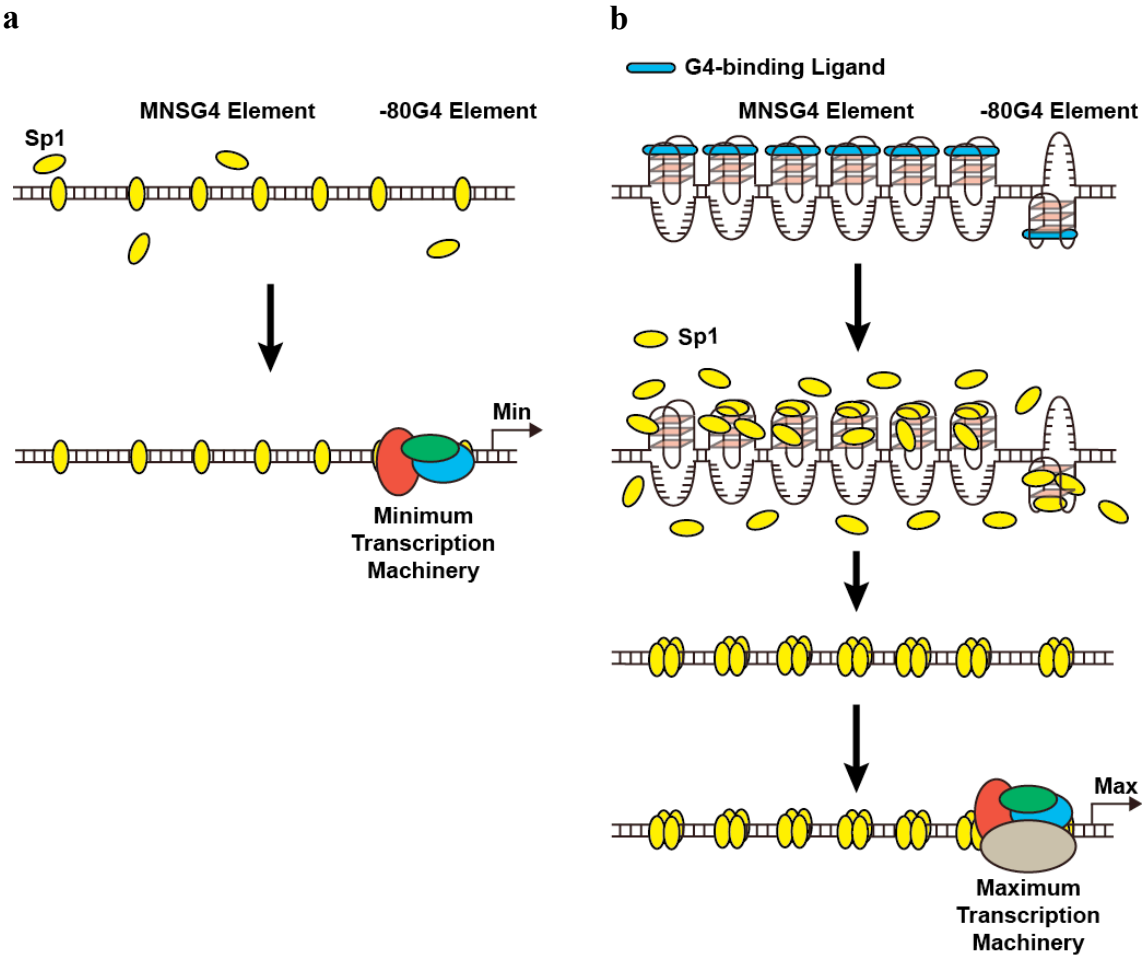


Figure 5.9: Schematic illustration of the G4/Sp1-mediated mechanism of human *TnIc* promoter activation.

(a) In the absence of G4-binding ligand or certain cardiac-restricted G4-binding protein, G4s cannot form in the promoter of human *TnIc*. Monomeric Sp1 interacts with the Sp1 binding sites in the promoter and recruits minimum transcription machinery, which then initiates the basal (minimum) transcription of human *TnIc*. **(b)** In the presence of G4-binding ligand (or certain cardiac-restricted G4-binding protein in human cardiomyocytes), formation of both MNSG4 and -80G4 recruits multiple Sp1. This will increase the local concentration of Sp1, and hence promote the formation of tetrameric Sp1. Sp1 tetramers also interact with Sp1 binding sites in the promoter and recruit maximum transcription machinery, or induce DNA looping to bring distal regulatory element [50, 51], such as enhancers, which induce maximum transcription activity.

In this scheme (Figure 5.9), the folding of both MNSG4s and -80G4 is firstly promoted by certain cardiac-restricted protein, similar to the G4-binding ligand, compound 3, in HEK293 cells. Then, the Sp1 tightly binds to the folded MNSG4s and -80G4 resulting in increased local concentration of Sp1, which subsequently induces the formation of Sp1 tetramers associated with the CACC/Sp1 box in the -80G4 element. Next, the Sp1 tetramers form higher-order protein complex, or induce DNA looping [50, 51], and initiate the maximal transcription activity. However, in the absence of such G4-binding protein or ligand, G4s cannot form in the human *TnIc* promoter, and only low concentration of Sp1 can be found around the CACC/Sp1 binding site, which results in only basal transcription activity. Thus, as shown in Figure 5.8 a, in the absence of G4-binding ligand, construct with mutated -80G4, but intact Sp1 binding site (hTnIc-80G4M) had similar transcription activity as the wild type one (hTnIc-6MNS(E)-WT), because only few Sp1 can bind to the canonical duplex DNA. Accordingly, mutation of this Sp1 binding site lead to decreased transcription activity of human *TnIc* promoter (hTnIc-6MNS/-80G4-M, Figure 5.8 a; hTnIc-80sp1M-1/2, and hTnIc-80G4/sp1M, Figure 5.5 b) as compared to the wide type one. In the presence of G4-binding ligand, the disruption of -80G4 formation (hTnIc-80G4M, Figure 5.8 b) cannot recruit tetramer Sp1 and thus give lowest transcription activity, similar as the overall mutant (hTnIc-6MNS/-80G4-M, Figure 5.8 b). However, to further investigate the validity of this G4/Sp1-mediated mechanism, *in vitro* studies about the binding of Sp1 to the -80G4, and the formation of Sp1 tetramers would be necessary in the future.

Superhelicity-induced mechanism

It has been know that the formation of G4 structures within the DNase I or S1 nuclease hypersensitive pPu/pPy tracts in gene promoters are often associated with negative supercoiling stress, which facilitates the local unwind or melting of duplex DNA [55-57]. For instance, formation of G4 structures have been revealed in negatively supercoiled plasmid containing the ILPR [58-60] and the promoter region of *VEGF* [61, 62], *MYC* [63], and so on. Meanwhile, transcription process by the RNA polymerase (RNApol) is another origin of the negative supercoiling. Hindering and rotation of the template DNA and RNApol generate negative supercoiling in the upstream region, and positive supercoiling in the downstream region [64]. This supercoiling stress may accumulate and inhibit further DNA transactions [65], unless

dissipated by certain mechanism. For example, topoisomerases catalyse and guide the unknotting or unkinking of these DNA supercoils by transiently breaking and then resealing DNA strands during transcription [66]. But in an *in vivo* study about the *MYC* promoter, this transcription-induced supercoiling force was found to enable DNA transitions to conformations other than B-DNA [67]. Thus, it is possible that in the case of human *TnIc* promoter, the stabilisation of multiple G4s by the ligand facilitates the local unwinding of the negatively supercoiled plasmid, and resolves the transcription-induced supercoiling forces, which hence intensify transcription rate (Figure 5.10). The -80G4-forming motif locates very close to the TSS, and hence the formation of this G4 should have a strong effect in respect to release the transcription-induced supercoiling forces. Furthermore, the formation of the -80G4 may also result in the exposure of multiple TF binding sites around it. This hypothetic scheme may explain the "gatekeeper" property the -80G4 element in the human *TnIc* promoter. The MNSG4 element, which is relatively distant from the TSS, and with less TF binding sites around, has weaker effects to the transcription activity as compared to the -80G4 element, and thus behaves as a "fine-tuner". However, to fully characterise this mechanism, *in vitro* and *in vivo* DMS footprinting to probe the formation of these G4 elements would be necessary.

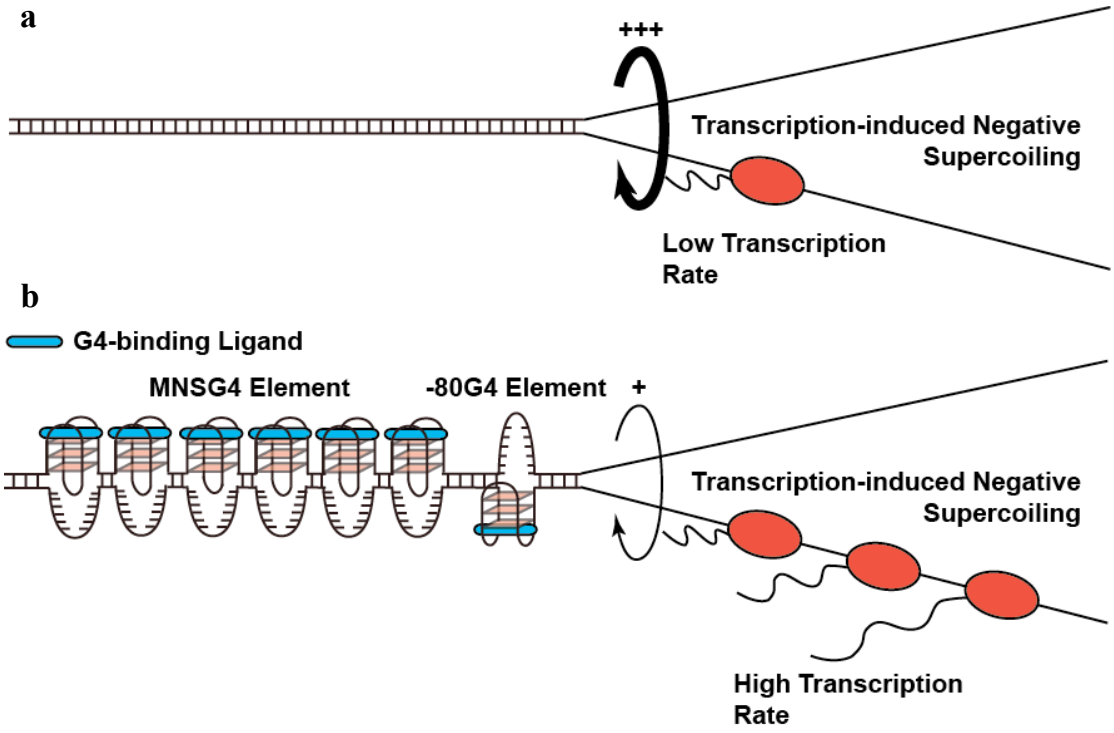


Figure 5.10: Schematic illustration of the superhelicity-induced mechanism of human *TnIc* promoter activation.

(a) In the absence of G4-binding ligand or certain cardiac-restricted G4-binding protein, G4s cannot form in the promoter of human *TnIc*. The transcription-induced negative supercoiling accumulates in the upstream region, which generates hindrance and inhibits further DNA transactions. This results in a low transcription rate. **(b)** Once stabilised by the G4-binding ligand or certain cardiac-restricted G4-binding protein, the formation of G4s in the human *TnIc* promoter partially dissipates the transcription-induced negative supercoiling, and thus results in a high transcription rate of the gene.

G4/E-box-mediated mechanism

Recently, a G4-mediated regulatory mechanism involving adjacent E-box has been proposed. In the MyoD example, Fry *et al.* proposed that the promoter G4s formed in muscle-specific genes may function as a binding sites for homodimeric MyoD to increase concentration of MyoD in the micro-environment around the promoters. The accumulated homodimeric MyoD then associate with constitutive E-proteins to form heterodimers, which bind to adjacent E-box (E47) and activate downstream gene expression [16-18]. This mechanism may also exist in the human cardiomyocyte,

since several cardiac-restricted bHLH proteins have been found to be important in heart development, such as HAND1 and HAND2 [68-70]. Interestingly, in the promoter of human *TnIc*, potential E-boxes are also identified around both MNSG4 and -80G4 [19]. These co-localized G4-forming motifs and adjacent E-boxes may suggest that G4/E-box-mediated mechanism similar as those proposed in MyoD example might be utilized in regulating *TnIc* transcription in human cardiomyocytes.

In summary, although still very preliminary, these results could provide insights into the biological roles and potentials of cardiac promoter G4s. In general, the position of promoter G4 is important in estimating its related biological functions. Meanwhile, actual mechanisms of promoter G4s in living cells might be much more complicated than a single molecular switch, and several G4s may act cooperatively. As shown in the *TnIc* example, although the mechanism is still yet to be further clarified, it is apparent that multiple G4s in the *TnIc* promoter act together in determining downstream gene expression. This study shows the feasibility of multi-targeting strategy via promoter G4s in manipulating certain gene expression or even modulating pathway performance. It is necessary that all the G4-forming motifs in a given promoter are carefully investigated to understand their integral regulatory functions.

5.11 References

1. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr., and Blackburn, E. H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs, *Cell 51*, 899-908.
2. Qin, Y., and Hurley, L. H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, *Biochimie 90*, 1149-1171.
3. Chaires, J. B. (2010) Human telomeric G-quadruplex: thermodynamic and kinetic studies of telomeric quadruplex stability, *FEBS J 277*, 1098-1106.
4. Dai, J., Carver, M., and Yang, D. (2008) Polymorphism of human telomeric quadruplex structures, *Biochimie 90*, 1172-1183.
5. Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H. J., and Pluckthun, A. (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei, *Proc Natl Acad Sci U S A 98*, 8572-8577.
6. Prescott, D. M. (1994) The DNA of ciliated protozoa, *Microbiol Rev 58*, 233-267.
7. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, *Proc Natl Acad Sci U S A 99*, 11593-11598.
8. Grand, C. L., Han, H., Munoz, R. M., Weitman, S., Von Hoff, D. D., Hurley, L. H., and Bearss, D. J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo, *Mol Cancer Ther 1*, 565-573.
9. Qin, Y., Rezler, E. M., Gokhale, V., Sun, D., and Hurley, L. H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4, *Nucleic Acids Res 35*, 7698-7713.
10. Sun, D., Liu, W. J., Guo, K., Rusche, J. J., Ebbinghaus, S., Gokhale, V., and Hurley, L. H. (2008) The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex-interactive agents, *Mol Cancer Ther 7*, 880-889.
11. Douarre, C., Gomez, D., Morjani, H., Zahm, J. M., O'Donohue M, F., Eddabra, L., Mailliet, P., Riou, J. F., and Trentesaux, C. (2005) Overexpression of Bcl-2 is associated with apoptotic resistance to the G-quadruplex ligand 12459 but is not sufficient to confer resistance to long-term senescence, *Nucleic Acids Res 33*, 2192-2203.
12. Bejugam, M., Sewitz, S., Shirude, P. S., Rodriguez, R., Shahid, R., and Balasubramanian, S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression, *J Am Chem Soc 129*, 12926-12927.
13. McLuckie, K. I., Waller, Z. A., Sanders, D. A., Alves, D., Rodriguez, R., Dash, J., McKenzie, G. J., Venkitaraman, A. R., and Balasubramanian, S. (2011) G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells, *J Am Chem Soc 133*, 2658-2663.
14. Eddy, J., and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, *Nucleic Acids Res 36*, 1321-1333.
15. Todd, A. K., and Neidle, S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements, *Nucleic Acids Res 36*, 2700-2704.
16. Walsh, K., and Gualberto, A. (1992) MyoD binds to the guanine tetrad nucleic acid structure, *J Biol Chem 267*, 13714-13718.

17. Yafe, A., Etzioni, S., Weisman-Shomer, P., and Fry, M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes, *Nucleic Acids Res 33*, 2887-2900.
18. Shklover, J., Weisman-Shomer, P., Yafe, A., and Fry, M. (2010) Quadruplex structures of muscle gene promoter sequences enhance in vivo MyoD-dependent gene expression, *Nucleic Acids Res 38*, 2369-2377.
19. Bhavsar, P. K., Dellow, K. A., Yacoub, M. H., Brand, N. J., and Barton, P. J. (2000) Identification of cis-acting DNA elements required for expression of the human cardiac troponin I gene promoter, *J Mol Cell Cardiol 32*, 95-108.
20. Bhavsar, P. K., Brand, N. J., Yacoub, M. H., and Barton, P. J. (1996) Isolation and characterization of the human cardiac troponin I gene (TNNI3), *Genomics 35*, 11-23.
21. Wahls, W. P., Wallace, L. J., and Moore, P. D. (1990) Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells, *Cell 60*, 95-103.
22. Denoeud, F., Vergnaud, G., and Benson, G. (2003) Predicting human minisatellite polymorphism, *Genome Res 13*, 856-867.
23. Park, G. H., Plummer, H. K., 3rd, and Krystal, G. W. (1998) Selective Sp1 binding is critical for maximal activity of the human c-kit promoter, *Blood 92*, 4138-4149.
24. Suntharalingam, K., White, A. J., and Vilar, R. (2010) Two metals are better than one: investigations on the interactions between dinuclear metal complexes and quadruplex DNA, *Inorg Chem 49*, 8371-8380.
25. Gallo, P., Grimaldi, S., Latronico, M. V., Bonci, D., Pagliuca, A., Ausoni, S., Peschle, C., and Condorelli, G. (2008) A lentiviral vector with a short troponin-I promoter for tracking cardiomyocyte differentiation of human embryonic stem cells, *Gene Ther 15*, 161-170.
26. Verma, A., Halder, K., Halder, R., Yadav, V. K., Rawal, P., Thakur, R. K., Mohd, F., Sharma, A., and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species, *J Med Chem 51*, 5641-5649.
27. Du, Z., Kong, P., Gao, Y., and Li, N. (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome, *Biochem Biophys Res Commun 354*, 1067-1070.
28. Fernando, H., Reszka, A. P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A. R., Neidle, S., and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene, *Biochemistry 45*, 7854-7860.
29. Kumari, S., Bugaut, A., Huppert, J. L., and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation, *Nat Chem Biol 3*, 218-221.
30. Vergnaud, G., and Denoeud, F. (2000) Minisatellites: mutability and genome architecture, *Genome Res 10*, 899-907.
31. Bois, P., and Jeffreys, A. J. (1999) Minisatellite instability and germline mutation, *Cell Mol Life Sci 55*, 1636-1648.
32. Ribeyre, C., Lopes, J., Boule, J. B., Piazza, A., Guedin, A., Zakian, V. A., Mergny, J. L., and Nicolas, A. (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo, *PLoS Genet 5*, e1000475.
33. Bennett, S. T., Lucassen, A. M., Gough, S. C., Powell, E. E., Undlien, D. E., Pritchard, L. E., Merriman, M. E., Kawaguchi, Y., Dronsfield, M. J., Pociot, F., and et al. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus, *Nat Genet 9*, 284-292.
34. Friedl, W., Ludwig, E. H., Paulweber, B., Sandhofer, F., and McCarthy, B. J. (1990) Hypervariability in a minisatellite 3' of the apolipoprotein B gene in patients with coronary heart disease compared with normal controls, *J Lipid Res 31*, 659-665.
35. Kimura, A., Harada, H., Park, J. E., Nishi, H., Satoh, M., Takahashi, M., Hiroi, S., Sasaoka, T., Ohbuchi, N., Nakamura, T., Koyanagi, T., Hwang, T. H., Choo, J. A., Chung, K. S., Hasegawa, A., Nagai, R., Okazaki, O., Nakamura, H., Matsuzaki, M.,

Sakamoto, T., Toshima, H., Koga, Y., Imaizumi, T., and Sasazuki, T. (1997) Mutations in the cardiac troponin I gene associated with hypertrophic cardiomyopathy, *Nat Genet 16*, 379-382.
36. Murphy, R. T., Mogensen, J., Shaw, A., Kubo, T., Hughes, S., and McKenna, W. J. (2004) Novel mutation in cardiac troponin I in recessive idiopathic dilated cardiomyopathy, *Lancet 363*, 371-372.
37. Murphy, A. M., Thompson, W. R., Peng, L. F., and Jones, L., 2nd. (1997) Regulation of the rat cardiac troponin I gene by the transcription factor GATA-4, *Biochem J 322 (Pt 2)*, 393-401.
38. Di Lisi, R., Millino, C., Calabria, E., Altruda, F., Schiaffino, S., and Ausoni, S. (1998) Combinatorial cis-acting elements control tissue-specific activation of the cardiac troponin I gene in vitro and in vivo, *J Biol Chem 273*, 25371-25380.
39. Iwaki, K., Sukhatme, V. P., Shubeita, H. E., and Chien, K. R. (1990) Alpha- and beta-adrenergic stimulation induces distinct patterns of immediate early gene expression in neonatal rat myocardial cells. fos/jun expression is associated with sarcomere assembly; Egr-1 induction is primarily an alpha 1-mediated response, *J Biol Chem 265*, 13809-13817.
40. Wang, G. W., Schuschke, D. A., and Kang, Y. J. (1999) Metallothionein-overexpressing neonatal mouse cardiomyocytes are resistant to H2O2 toxicity, *Am J Physiol 276*, H167-175.
41. Gonzalez, V., Guo, K., Hurley, L., and Sun, D. (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein, *J Biol Chem 284*, 23622-23635.
42. Brooks, T. A., and Hurley, L. H. (2010) Targeting MYC Expression through G-Quadruplexes, *Genes Cancer 1*, 641-649.
43. Kaczynski, J., Cook, T., and Urrutia, R. (2003) Sp1- and Kruppel-like transcription factors, *Genome Biol 4*, 206.
44. Suske, G., Bruford, E., and Philipsen, S. (2005) Mammalian SP/KLF transcription factors: bring in the family, *Genomics 85*, 551-556.
45. Sartorelli, V., Hong, N. A., Bishopric, N. H., and Kedes, L. (1992) Myocardial activation of the human cardiac alpha-actin promoter by helix-loop-helix proteins, *Proc Natl Acad Sci U S A 89*, 4047-4051.
46. Azakie, A., Fineman, J. R., and He, Y. (2006) Sp3 inhibits Sp1-mediated activation of the cardiac troponin T promoter and is downregulated during pathological cardiac hypertrophy in vivo, *Am J Physiol Heart Circ Physiol 291*, H600-611.
47. Yu, B., Datta, P. K., and Bagchi, S. (2003) Stability of the Sp3-DNA complex is promoter-specific: Sp3 efficiently competes with Sp1 for binding to promoters containing multiple Sp-sites, *Nucleic Acids Res 31*, 5368-5376.
48. Liu, J., and Levens, D. (2006) Making myc, *Curr Top Microbiol Immunol 302*, 1-32.
49. Majello, B., De Luca, P., Suske, G., and Lania, L. (1995) Differential transcriptional regulation of c-myc promoter through the same DNA binding sites targeted by Sp1-like proteins, *Oncogene 10*, 1841-1848.
50. Mastrangelo, I. A., Courey, A. J., Wall, J. S., Jackson, S. P., and Hough, P. V. (1991) DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement, *Proc Natl Acad Sci U S A 88*, 5670-5674.
51. Su, W., Jackson, S., Tjian, R., and Echols, H. (1991) DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1, *Genes Dev 5*, 820-826.
52. Kumar, P., Yadav, V. K., Baral, A., Saha, D., and Chowdhury, S. (2011) Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide, *Nucleic Acids Res 39*, 8005-8016.
53. Raiber, E. A., Kranaster, R., Lam, E., Nikan, M., and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro, *Nucleic Acids Res 40*, 1499-1508.

54. Dellow, K. A., Bhavsar, P. K., Brand, N. J., and Barton, P. J. (2001) Identification of novel, cardiac-restricted transcription factors binding to a CACC-box within the human cardiac troponin I promoter, *Cardiovasc Res 50*, 24-33.
55. McCarthy, J. G., and Heywood, S. M. (1987) A long polypyrimidine/polypurine tract induces an altered DNA conformation on the 3' coding region of the adjacent myosin heavy chain gene, *Nucleic Acids Res 15*, 8069-8085.
56. Michelotti, G. A., Michelotti, E. F., Pullner, A., Duncan, R. C., Eick, D., and Levens, D. (1996) Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo, *Mol Cell Biol 16*, 2656-2669.
57. Sun, D., Guo, K., Rusche, J. J., and Hurley, L. H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents, *Nucleic Acids Res 33*, 6070-6080.
58. Hammond-Kosack, M. C., and Docherty, K. (1992) A consensus repeat sequence from the human insulin gene linked polymorphic region adopts multiple quadriplex DNA structures in vitro, *FEBS Lett 301*, 79-82.
59. Hammond-Kosack, M. C., Dobrinski, B., Lurz, R., Docherty, K., and Kilpatrick, M. W. (1992) The human insulin gene linked polymorphic region exhibits an altered DNA structure, *Nucleic Acids Res 20*, 231-236.
60. Hammond-Kosack, M. C., Kilpatrick, M. W., and Docherty, K. (1993) The human insulin gene-linked polymorphic region adopts a G-quartet structure in chromatin assembled in vitro, *J Mol Endocrinol 10*, 121-126.
61. Sun, D. (2010) In vitro footprinting of promoter regions within supercoiled plasmid DNA, *Methods Mol Biol 613*, 223-233.
62. Sun, D., Guo, K., and Shin, Y. J. (2011) Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene, *Nucleic Acids Res 39*, 1256-1265.
63. Sun, D., and Hurley, L. H. (2009) The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression, *J Med Chem 52*, 2863-2874.
64. Droge, P. (1994) Protein tracking-induced supercoiling of DNA: a tool to regulate DNA transactions in vivo?, *Bioessays 16*, 91-99.
65. Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J., and Block, S. M. (1998) Force and velocity measured for single molecules of RNA polymerase, *Science 282*, 902-907.
66. Kouzine, F., Sanford, S., Elisha-Feil, Z., and Levens, D. (2008) The functional response of upstream DNA to dynamic supercoiling in vivo, *Nat Struct Mol Biol 15*, 146-154.
67. Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells, *Nat Struct Mol Biol 13*, 1055-1059.
68. Srivastava, D. (2006) Making or breaking the heart: from lineage determination to morphogenesis, *Cell 126*, 1037-1048.
69. Risebro, C. A., Smart, N., Dupays, L., Breckenridge, R., Mohun, T. J., and Riley, P. R. (2006) Hand1 regulates cardiomyocyte proliferation versus differentiation in the developing heart, *Development 133*, 4595-4606.
70. Scott, I. C., Anson-Cartwright, L., Riley, P., Reda, D., and Cross, J. C. (2000) The HAND1 basic helix-loop-helix transcription factor regulates trophoblast differentiation via multiple mechanisms, *Mol Cell Biol 20*, 530-541.

Chapter 6 Summary and Future Work

Since the discovery of G4 structures formed by human telomeric oligonucleotides in solution [1], this four-stranded DNA structure has been found to be associated with diverse biological processes, such as regulating downstream gene transcription [2, 3], mediating mRNA splicing [4, 5], maintaining genome stability [6], and interfering telomere maintenance (reviewed in [7], and Section 1.2.3). Furthermore, bioinformatics researches have shown that these G4-forming sequences are particularly enriched in the telomere region of eukaryotic chromosomes and within the promoter and 5'UTR regions of human genes [8, 9], particularly those involved in maintaining the immortality of cancerous cells [10]. Thus, the cancer-related enrichment and distinctive structures of G4s have inspired research interest of using G4 DNA as novel therapeutic targets in cancer treatment, for example, the selective inhibition of transcription of a given oncogene using designed small molecules to stabilise a particular G4 (reviewed in [11]). Although there are some successful examples in designing G4-binding ligands to target a particular G4 *in vitro* and *in vivo*, specifically discriminating and targeting G4s via small synthetic G4-binding ligands remains very challenging at present due to several intrinsic hurdles as mentioned in Section 3.7. It seems to be inevitable that when applied *in vivo*, G4-binding ligands will interact with multiple G4s other than their original, specific targets. Interestingly, it has been demonstrated that in certain cases, the off-target or multi-targeting effects of G4-binding ligands could be used to achieve higher anti-cancer efficiency, such as the examples of naphthalene diimide derivative 1 [12], ligand 12459 [4, 13], and Quarfloxin [14] as mentioned in Section 1.6.

On the other hand, although anti-cancer activities of G4s (both of human telomeric G4 and oncogene promoter G4s) have been intensely investigated, the therapeutic potential of G4s in other diseases is still largely unexplored. According to previous bioinformatics studies, G4s are widely distributed in the human genome, particularly in the TRRs, which organise downstream gene transcription [8, 9, 15-17]. Furthermore, individual examples have also highlighted potential biological roles of G4s in promoters of several genes relevant to aspects of cardiovascular function, including *PDGF-A*, *VEGF-A* and *HIF-1α* (reviewed in Section 1.7.1). Thus, it is

plausible to assume that promoter G4s are also involved in regulating cardiac function-related genes in both physiological and pathological processes, and these G4s might be used as potential targets in the treatment of certain cardiomyopathies.

The main aims of this thesis are: (i) to evaluate the importance of G4s in the TRRs of genes related to cardiac function by bioinformatics approaches and choose model genes according to the bioinformatics results (Chapter 3), (ii) to characterise the biophysical properties of G4s in TRRs of the model genes in solution (Chapter 4), and (iii) to investigate the regulatory roles of these G4s and to test multi-targeting strategy at the single gene level (Chapter 5).

6.1 Importance of G4s in cardiac function-related genes

The importance of G4s was firstly evaluated at the tissue level by counting the percentage of genes containing at least one potential G4-forming motif ($G_{\geq 3}(N_{1\sim 7}G_{\geq 3})_{\geq 3}$) in their TRRs (-2,000 ~ +1,000 bp). High enrichment of G4s in the TRRs of genes active in a particular tissue may reflect that G4s are favoured in this tissue in the process of evolution, and thus increase the possibility to reveal G4-mediated regulatory mechanisms in that tissue. By analysing 50 tissue-specific protein-encoding transcriptomes datasets, tissues with high and low G4 enrichment in the TRRs have been revealed (tissue-specific G4 enrichment, Figure 3.1, Table 3.1). Interestingly, although for most tissues, G4 enrichment is consistent when different regions of TRR were included in analysis, there are some exceptions. For example, in smooth muscle, adipose tissue, olfactory bulb, and fetal thyroid, although the G4 enrichment is relatively low when the whole length of TRR was included, the G4 enrichment in the distal or/and proximal promoter region is among the highest (Figure 3.1, Table 3.1). This tissue- and location-specific enrichment of G4s in TRRs suggests that the occurrence or reserve of G4s in TRRs is not a random process in different tissues, which may result from the natural selection in order to evolve tissue-specific gene regulation. In this section, it was found that the enrichment of G4s in TRRs of genes active in the heart is among the highest, no matter whether the whole or partial region of TRR was analysed. Although the actual reason of the high G4 enrichment in the heart needs to be further investigated in the future, this preliminary result gives the

first evidence indicating that G4-forming sequences are favoured or at least less disfavoured in the TRRs of cardiac function-related genes.

The importance of G4s was also evaluated at a pathway level by considering the abundance and location significance of G4s found in TRRs. G4 abundance score (*F* score) of each transcript was calculated as the percentage of nucleotide involved in G4 formation in its TRR (Eq. 2.26). G4 location significance score (*Q* score) was calculated to take into account the biased distribution of G4-forming sequences in TRRs (Eq. 2.27 and 2.28), which reflects evolutionary pressure. The positional distribution of G4s (Figure 3.3 a) and the histograms of *F* (Figure 3.2 a ~ c) and *Q* (Figure 3.3 c ~ d) scores demonstrated that in the proximal promoter (-500 ~ -1 bp) and downstream region relative to TSS (TSS ~ +1,000 bp) G4s are more favoured on the coding strand as compared to the template strand. The abundance of G4s on the coding strand downstream to the TSS has been observed in genomes of human [9] and other warm-blooded animals [18, 19], which are involved in maintaining separation of coding and template strand during transcription [20]. However, the reason for the slightly asymmetric distribution of G4s in the proximal promoter region is still unclear, which may be worthy of further investigation. By calculating the cumulative frequencies of the G4 abundance (CF_o) and location significance (CQ_o) scores of both strands, G4-important pathways were determined as described in Section 2.4.4. Pathways with high G4-importance showed a strong connection with their functions, and can be categorized into six distinct groups (Section 3.6, Table 3.5). Interestingly, as compared to the pathways involved in cancers, six out of seven cardiac function-related pathways were found to be equally G4-important (Table 3.5). The high abundance and location significance of G4s in cancer-related pathways might partially contribute to the successes of targeting G4s in anti-cancer treatment even in the presence of off-target and multi-targeting effects. Although very basic, these findings clearly supported the original hypothesis that G4s are important in regulating cardiac function-related genes and, potentially, multi-targeting strategy may have similar therapeutic potential as those observed in anti-cancer research.

To validate the bioinformatics results (i.e. that G4s are involved in regulating cardiac function-related genes) and to initially test the multi-targeting strategy within a single gene, two genes with high G4-importance in their TRRs, human *MEF2D* and *TnIc*, were chosen as models. The human MEF2D belongs to the MEF2 family, the

members of which are all G4-important in the TRR according to the bioinformatics results (CF_{max} and CQ_{max} values are: MEF2A, 79.8 % and 85.1 %, MEF2B, 62.8 % and 74.0 %, MEF2C, 51.7 % and 61.5 %, and MEF2D, 69.0 % and 93.8 %). The human TnIc, known as the only sarcomeric protein found to be exclusively expressed in adult cardiac myocytes, is also highly G4-important in respect of its G4 abundance and location significance in the TRR (CF_o 92.6%, and CQ_o 95.2%).

6.2 Biophysical characterisations of human *MEF2D* and *TnIc* promoter G4s

In Chapter 4, biophysical properties of G4-forming sequences from the promoter of human *MEF2D* (MEF2DG4: between -169 and -140 bp), and *TnIc* (MNSG4: consensus sequence between -528 and -319 bp, and -80G4: centred at -80 bp) were thoroughly investigated in solution.

6.2.1 Formation and conformation

Formation of G4 structures from these sequences was determined by EMSA and CD spectroscopy. In CD spectroscopy, MEF2DG4-forming sequence was found to adopt a hybrid of parallel/anti-parallel G4(s) in solutions containing K^+ and Li^+ (Figure 4.2 a), while the MNSG4 and -80G4 from human *TnIc* promoter adopted anti-parallel (possibly mixed with a small amount of parallel G4) and parallel structures, respectively, in solution containing K^+ (MNSG4: Figure 4.2 d, and -80G4: Figure 4.2 g). Due to its high spatial-temporal resolution and sensitivity, the smFRET technique was also used to reveal the conformational heterogeneity of these G4s. In smFRET conformational analysis, formation of G4s from all these sequences was also confirmed by the existence of high FRET subpopulations (Figure 4.4, 4.6 ~ 4.9). In particular, the two-repeat MNSG4 was found to be able to adopt parallel and anti-parallel G4s on the same strand due to the its high FRET subpopulations (Figure 4.8, **S4**s), although its CD spectrum is complicated (Figure 4.2 j). Furthermore, results from the CD spectroscopy correlated well with those from smFRET conformational analysis in respect to the conformation and stability of these G4s (discussed in detail

in Section 4.9.2). The folding pattern of MNSG4 in the presence of K^+ was determined as $G_3N_5G_3N_5G_3N_2G_3$ by DMS footprinting (Figure 4.16 a). Although the folding pattern of the -80G4 was difficult to resolve, the partially protected guanines also clearly indicated the formation of different G4s (Figure 4.16 b). The different folding pattern of these cardiac promoter G4s as compared to the proto-oncogene promoter G4s ($G_3N_{1\text{-}3}G_3N_{2\text{-}9}G_3N_1G_3$ [21]) may relate to their distinct biological functions *in vivo*.

6.2.2 Unfolding thermodynamics

Unfolding thermodynamic parameters of these G4s were determined mainly by CD thermal denaturation experiments (CD melting). Although the thermal stability of MEF2DG4 ($\Delta G_{(310K)} \sim 5.1$ kcal mol^{-1}) was found to be lower than that of human telomeric G4s [22, 23] due to the populated intermediates revealed by the dual-wavelength parametric test [24], MEF2DG4 is among the most thermally stable G4s that have been reported so far according to its very high melting temperature ($T_m \sim 93$ °C) in the presence of K^+ (Figure 4.2 a ~ c, Table 4.1). For the TnIc G4s, the anti-parallel MNSG4 and parallel -80G4 exhibited similar thermal stability in the presence of K^+ (Figure 4.2 d ~ i, Table 4.1) as compared to the human telomeric G4 [22, 23], and G4s in the promoters of *MYC*, *KIT*, *WNT1* and *VEGF* with similar loop length [23, 25]. Given the proved biological roles of these G4s, both of MNSG4 and -80G4 should be thermally stable enough to possess similar biological functions. Furthermore, thermal denaturation of two-repeat MNSG4 clearly demonstrated cooperative unfolding patterns of the parallel and anti-parallel G4s (Figure 4.2 k, l), which suggests that multiple MNSG4s in the promoter of human TnIc may also act cooperatively to influence downstream gene transcription.

6.2.3 Unfolding kinetics

The unfolding rate constant of the end dual-labeled MEF2DG4 was measured to be $8.3 \pm 0.4 \times 10^{-6}$ s^{-1} (Figure 4.11), which is much smaller than most of other intramolecular G4s measured under similar conditions, such as the *MYC* promoter G4 [26]. The MNSG4 ($t_{1/2\ (35\ °C)} \sim 2.4 \times 10^3$ s) was found to be less kinetically stable than

the slow unfolding species of -80G4 ($t_{1/2\ (35\ °C)}$ ~ 1.2 × 10^5 s) (Table 4.2). Furthermore, the kinetic stability of the two-repeat MNSG4 was even lower ($t_{1/2\ (25\ °C)}$ ~ 1.1 × 10^3 s, Table 4.3), and this may be caused by the single-stranded linker between individual MNSG4. Due to the weak kinetic stability of MNSG4, it is possible that in the presence of C-rich strand, such as the situation in the cells, MNSG4s may be less biologically important as compared to the -80G4.

According to the biophysical characterisations, although it is clear that the MEF2D and TnIc G4s were able to adopt relatively stable G4 structures in single-stranded forms in solution, it is unable to prove the existence of these G4s *in vitro* and *in vivo* due to several reasons. Firstly, as discussed in Section 4.9.4, due to several extrinsic and intrinsic limitations, it is inappropriate to conclude the relative stability between MEF2D/TnIc G4s and other G4s by simply comparing the thermodynamic and kinetic parameters. Secondly, similar as the human ILPR G4s [27], multiple MNS G4s may interact with each other and give dramatically different folding/unfolding behaviour. This G4-G4 interaction further illustrated the biological or even therapeutic values of multi-targeting strategy of G4-binding ligands. Thirdly, besides their intrinsic stability, the formation of G4s *in vitro* and *in vivo* is determined by a variety of other environmental factors, such as molecular crowding [28], transcription induced supercoiling [29], and the presence of different G4-interacting proteins (reviewed in Section 1.2.5). Thus, the formation and biological functions of G4s have to be investigated case by case.

6.3 Regulatory roles of G4s in human *TnIc* promoter

In Chapter 5, the regulatory roles of human *TnIc* promoter G4s were preliminarily investigated mainly by the dual luciferase reporter assays in the HEK293 cell line. Briefly, it was found that without the stabilising effects from the G4-binding ligands, disruption of the G4-forming sequences in both MNSG4 and -80G4 elements did not change the downstream luciferase transcription level (Figure 5.4 and 5.5). However, the mutation of Sp1 binding site within the -80G4 (-93 ~ -85 bp) inhibited downstream luciferase expression by round 50 % (Figure 5.5). By screening a group of conjugated metallo-organic G4-binding ligands, the di-copper complex (compound

3, Figure 5.6 a) was chosen as the best ligand according to its affinity and stabilising ability (Section 4.8, Figure 4.17 and 4.18) to both MNSG4 and -80G4 in solution, weak cytotoxicity (Figure 5.7 b), and the best cellular responses (Figure 5.6 b, Section 5.7). By incubating this ligand with HEK293 cells transfected with a variety of luciferase vectors driven by constructed human *TnIc* promoters containing different G4 elements, the cooperative regulatory roles of MNSG4 and -80G4 have been revealed. Generally, the -80G4 mainly acts as a gatekeeper in regulating downstream gene transcription, while the MNSG4s are fine-tuners, whose repeats number positively relate to the activity of downstream gene transcription when the -80G4 is intact (Figure 5.8, Section 5.9). These *in vitro* results agreed well with what have been found by bioinformatics studies and biophysical characterisations in solution. For example, according to the hypothesis in Chapter 3, -80G4 should be more important than the MNSG4 in regulating downstream gene transcription due to its higher location significance. In Chapter 4, the -80G4 was found to be more thermally and kinetically stable than the MNSG4 in the absence and presence of the G4-binding ligand (compound 3, di-copper complex). Thus, although evidence showing the direct interactions between the ligand and G4s in cells is still missing at this stage, it is plausible to assume that the differential activating effects of the luciferase expression vectors in the presence of G4-binding ligand were induced by the stabilisation of different G4s in the promoter of human *TnIc*. Thus, by comparing these results with other proposed G4-mediated mechanisms, three possible G4-mediated *TnIc* promoter activation mechanisms have been proposed, which are G4/Sp1-mediated, superhelicity-induced, and G4/E-box-mediated mechanisms (discussed in Section 5.10.4).

However, as discussed in Section 5.10.2, the *in vitro* studies were carried out in HEK293 cells rather than cardiac myocytes at this stage, and thus it is reasonable to argue whether the proposed mechanisms exist in cardiac myocytes even if one of them do account for the activation of humane *TnIc* promoter observed in HEK293 cell model. There are several lines of evidence indirectly indicating the potential biological relevance of these findings. Firstly, the bioinformatics results showed that TRRs in genes active in the heart and involved in cardiac function-related pathways are highly enriched by potential G4-forming motifs (Chapter 2). From the viewpoint of evolution, it is very likely that cardiac myocytes have evolved certain mechanisms

to manipulate the genes with high G4 contents in their TRRs to tolerate or even to utilize G4s in the processes of transcription. Secondly, the cooperative regulatory behaviours of MNSG4 of -80G4 reflected that these G4 elements are not isolated to each other during the process of evolution. Given the high abundance of G4s in TRRs of cardiac function-related genes, this further enhances the feasibility to manipulate transcription of these genes via multi-targeting strategy at the single gene level, or even at pathway level. Finally, although Sp1 and E proteins are ubiquitous, and MyoD is restricted in skeletal muscle, TFs with similar G4-interacting ability may be involved in targeting G4s in cardiac function-related genes and regulating their transcription in a tissue-specific manner. For example, zinc finger protein GATA4, whose binding site was found only 1 nucleotide from the -80G4, can activate various cardiac genes, such as human *TnIc* [30], *α*- and *β-MHC* genes [31]. The C2H2 zinc finger factor KLF13, which like several KLFs that are expressed in heart, was shown to interact with GATA-4 and modulate its function the heart development [32]. Furthermore, similar to the example of MyoD, the homodimer of a cardiac-restricted bHLH protein, HAND1, binds to canonical duplex DNA at least an order of magnitude lower than that of HAND1/E-factor heterodimers [33]. Other TFs active in the heart with similar DNA-interacting motifs as compared to the Sp1 (zinc finger) and MyoD (bHLH), may therefore also possibly act as G4-interacting proteins.

In summary, these results could provide the first evidence about the biological roles and potentials of cardiac promoter G4s. In general, actual mechanisms of promoter G4s in living cells might be much more complicated than isolated single molecular switch. In the example of the human *TnIc* promoter, several G4s may act cooperatively to determine downstream gene transcription. This study shows the feasibility of multi-targeting strategy via promoter G4s in manipulating certain gene expression or even modulating pathway performance.

6.4 Future work

In Chapter 3, the enrichment of G4s in cardiac function-related genes was simply determined by the high percentage of TRRs containing at least one G4-forming motif in genes active in the heart. As has been pointed out in Section 3.7.1, it is possible that

there are a group of ubiquitous genes with high G4 contents in their TRRs involved in different tissues. Thus, to further reveal the genes contributing to the high G4 enrichment in the heart, the overlap of genes containing G4s in their TRRs will be analysed and compared among different tissues. In addition, it would be interesting to compare the tissue-specific G4 enrichment in genomes from different species in order to determine if tissue- and/or location-specific G4 enrichment is conserved among species. On the other hand, although cardiac function-related pathways were found to be highly G4-important, little is known about the actual genes with high G4-importance. Thus, genes with high G4-importance found in those cardiac function-related pathways in human could be compared with those found in rat and mouse to reveal conserved G4-forming motifs. By analysing sequences similarities among these conserved cardiac G4s, grouping genes containing conserved G4s with similar sequence composition in the TRRs, and locating their position in certain pathways, it is possible to reveal potential regulatory roles of G4s at pathway level, and propose therapies for certain cardiovascular diseases by targeting these conserved G4s with similar G4-binding ligands (multi-targeting strategy).

Although unfolding thermodynamic and kinetic properties of MEF2D and TnIc G4s have been thoroughly investigated in Chapter 4, it is difficult to accurately compare the relative stability of them with other G4s involved in regulating oncogene transcription. To answer this question, by following the same labeling strategy and experimental approach, the MYC G4 could be used as a prototype. Furthermore, the folding patterns and possible conformations of both MNSG4 and -80G4 are still unclear due to multiple guanines (> 3) found in the G-runs. This could be addressed by specifically mutating those ambiguous guanines found in the DMS footprinting experiments, and determining their folding patterns and possible conformations via CD spectroscopy, smFRET analysis, and DMS footprinting. Meanwhile, to further evaluate their biological importance *in vitro*, probing the formation of these TnIc G4s in supercoiled plasmids via DMS footprinting would also be necessary in the future.

In respect to the regulatory roles of TnIc G4s, several questions need to be answered to further confirm the actual mechanisms. The first question is whether the activating effect is caused by direct interaction between the TnIc G4 elements and the ligand *in vitro*. Given the strong quenching ability of the di-copper complex, this could be addressed by comparing the quenching of fluorophores labeled plasmid containing

TnIc G4s in the absence of presence of the ligands. The second question is which TFs are involved G4-mediated *TnIc* transcription regulation. By incubating TnIc G4s with nuclear extracts from HEK293 cells and neonatal rat cardiomyocytes, TF(s) or other protein(s) that are able to interact with these G4s could then be revealed by EMSA experiments. Supershift EMSA with specific antibodies that recognize candidate proteins could be used to confirm involvement of those TF(s) or protein(s) [30]. Meanwhile, all experiments performed in HEK293 cells will be repeated in neonatal rat/mouse cardiomyocytes to directly evaluate the cardiac-specific regulatory roles of these G4s in the future.

6.5 References

1. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr., and Blackburn, E. H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs, *Cell 51*, 899-908.
2. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, *Proc Natl Acad Sci U S A 99*, 11593-11598.
3. Gonzalez, V., and Hurley, L. H. (2010) The c-MYC NHE III(1): function and regulation, *Annu Rev Pharmacol Toxicol 50*, 111-129.
4. Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J. L., and Riou, J. F. (2004) Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing, *Nucleic Acids Res 32*, 371-379.
5. Marcel, V., Tran, P. L., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M. P., Hall, J., Mergny, J. L., Hainaut, P., and Van Dyck, E. (2011) G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms, *Carcinogenesis 32*, 271-278.
6. Sanders, C. M. (2010) Human Pif1 helicase is a G-quadruplex DNA-binding protein with G-quadruplex DNA-unwinding activity, *Biochem J 430*, 119-128.
7. Neidle, S. (2010) Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer, *FEBS J 277*, 1118-1125.
8. Huppert, J. L., and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome, *Nucleic Acids Res 35*, 406-413.
9. Huppert, J. L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res 36*, 6260-6268.
10. Eddy, J., and Maizels, N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome, *Nucleic Acids Res 34*, 3887-3896.
11. Balasubramanian, S., Hurley, L. H., and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?, *Nat Rev Drug Discov 10*, 261-275.
12. Gunaratnam, M., Swank, S., Haider, S. M., Galesa, K., Reszka, A. P., Beltran, M., Cuenca, F., Fletcher, J. A., and Neidle, S. (2009) Targeting human gastrointestinal stromal tumor cells with a quadruplex-binding small molecule, *J Med Chem 52*, 3774-3783.
13. Riou, J. F., Guittat, L., Mailliet, P., Laoui, A., Renou, E., Petitgenet, O., Megnin-Chanet, F., Helene, C., and Mergny, J. L. (2002) Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands, *Proc Natl Acad Sci U S A 99*, 2672-2677.
14. Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C. B., Proffitt, C., Trent, K., Whitten, J. P., Lim, J. K., Von Hoff, D., Anderes, K., and Rice, W. G. (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis, *Cancer Res 69*, 7653-7661.
15. Todd, A. K., Johnston, M., and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA, *Nucleic Acids Res 33*, 2901-2907.
16. Huppert, J. L., and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome, *Nucleic Acids Res 33*, 2908-2916.
17. Eddy, J., and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, *Nucleic Acids Res 36*, 1321-1333.
18. Verma, A., Halder, K., Halder, R., Yadav, V. K., Rawal, P., Thakur, R. K., Mohd, F., Sharma, A., and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species, *J Med Chem 51*, 5641-5649.

19. Du, Z., Kong, P., Gao, Y., and Li, N. (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome, *Biochem Biophys Res Commun 354*, 1067-1070.
20. Du, Z., Zhao, Y., and Li, N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription, *Genome Res 18*, 233-241.
21. Qin, Y., and Hurley, L. H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, *Biochimie 90*, 1149-1171.
22. Balagurumoorthy, P., and Brahmachari, S. K. (1994) Structure and stability of human telomeric sequence, *J Biol Chem 269*, 21858-21869.
23. Kumar, N., and Maiti, S. (2008) A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes, *Nucleic Acids Res 36*, 5610-5622.
24. Wallimann, P., Kennedy, R. J., Miller, J. S., Shalongo, W., and Kemp, D. S. (2003) Dual wavelength parametric test of two-state models for circular dichroism spectra of helical polypeptides: anomalous dichroic properties of alanine-rich peptides, *J Am Chem Soc 125*, 1203-1220.
25. Olsen, C. M., Gmeiner, W. H., and Marky, L. A. (2006) Unfolding of G-quadruplexes: energetic, and ion and water contributions of G-quartet stacking, *J Phys Chem B 110*, 6962-6969.
26. Halder, K., and Chowdhury, S. (2005) Kinetic resolution of bimolecular hybridization versus intramolecular folding in nucleic acids by surface plasmon resonance: application to G-quadruplex/duplex competition in human c-myc promoter, *Nucleic Acids Res 33*, 4466-4474.
27. Schonhoft, J. D., Bajracharya, R., Dhakal, S., Yu, Z., Mao, H., and Basu, S. (2009) Direct experimental evidence for quadruplex-quadruplex interaction within the human ILPR, *Nucleic Acids Res 37*, 3310-3320.
28. Xue, Y., Kan, Z. Y., Wang, Q., Yao, Y., Liu, J., Hao, Y. H., and Tan, Z. (2007) Human telomeric DNA forms parallel-stranded intramolecular G-quadruplex in K+ solution under molecular crowding condition, *J Am Chem Soc 129*, 11185-11191.
29. Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells, *Nat Struct Mol Biol 13*, 1055-1059.
30. Dellow, K. A., Bhavsar, P. K., Brand, N. J., and Barton, P. J. (2001) Identification of novel, cardiac-restricted transcription factors binding to a CACC-box within the human cardiac troponin I promoter, *Cardiovasc Res 50*, 24-33.
31. Charron, F., Paradis, P., Bronchain, O., Nemer, G., and Nemer, M. (1999) Cooperative interaction between GATA-4 and GATA-6 regulates myocardial gene expression, *Mol Cell Biol 19*, 4355-4365.
32. Nemer, M., and Horb, M. E. (2007) The KLF family of transcriptional regulators in cardiomyocyte proliferation and differentiation, *Cell Cycle 6*, 117-121.
33. Scott, I. C., Anson-Cartwright, L., Riley, P., Reda, D., and Cross, J. C. (2000) The HAND1 basic helix-loop-helix transcription factor regulates trophoblast differentiation via multiple mechanisms, *Mol Cell Biol 20*, 530-541.