

**Studies of the effects of promoter sequence
variation on gene expression in human
chromosome 22**

Jamil Bacha



**Wolfson College
University of Cambridge**

This dissertation is submitted for the degree of Doctor of Philosophy



Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation does not exceed the page limit specified by the Biology Degree Committee.

Acknowledgements

This thesis would not have been possible without the support, both scientific and personal, of a great many people. First and foremost, my deepest personal thanks go to my supervisor Dr. Ian Dunham for his unstinting support and mentorship throughout my time in the group. I would also like to thank Dr. John Collins and Dave Beare for their accumulated lab and computer wisdom, and all the members of Team 62 for making this a fun and unforgettable experience. Special thanks also go to Dr. Nick Luscombe and (nearly-Dr.) Juanma Vaquerizas at the EMBL-European Bioinformatics Institute for their invaluable intellectual contribution to the array analysis work, and particularly to Juanma for carrying out the quality control and linear modelling analyses on the array data.

Thank you to my thesis committee Dr. Alex Bateman and Dr. Dave Vetrie (Sanger Institute) and Dr. James Ajioka (University of Cambridge) for their intellectual input and guidance. A big thank you to Dr. Steven Leonard and Dr. Sarah Hunt for helping me mercilessly flog the SNP database until it did what it was told (eventually!), and to Nick Matthews, Jonathan “Bill” Bailey and the Sanger Institute Sequencing Centre for their hard work on the promoter and clone re-sequencing. Thank you to Dr. Barbara Stranger for running the association analysis of promoter SNPs and for our fun talks on the Sanger bus, Dr. Robert Andrews and Dr. Gregory Lefebvre for contributing the clustering of co-expressed genes, and Dr. Thomas Down for his invaluable assistance with the motif generation and analysis. Thank you to Andy “Wilb” Dunham and Andy Bentley of the ExoSeq group for teaching me how to tame the wild lab robot! Thank you to Dr. Manolis Dermitzakis, Dr. Ewan Birney, Dr. Thomas Down and Dr. Vardhman Rakyan for their very interesting scientific discussions and to my fellow graduate students for their not-so-scientific ones. Good luck guys!

My love and gratitude go to my family, without whom I would not have had the opportunity to embark on this journey and fulfil my childhood ambitions of a scientific career. Finally, special thanks to Dr. Davina Stevenson for her love and companionship through the highs and the lows ... something I will always treasure.

Abstract

The molecular and physiological phenotype of a gene depends not only on the structure and properties of the protein it codes for, but on the regulation of the magnitude and timing of expression of that protein in the cell. The role of the promoter in gene regulation can be seen as an integrator of the numerous intra- and extra-cellular signals that influence the levels of transcription factors in the nucleus, with the output being the level of transcriptional initiation. The identification of transcription factor binding sites and promoter polymorphisms with real functional consequences continues to elude purely computational methods, and more experimental data is needed before this state of affairs is changed. In this project, I have re-sequenced the majority of promoters on human chromosome 22 from a panel of 48 unrelated individuals, generating a set of 807 promoter SNPs with associated genotype information. I then developed a novel high-throughput cloning strategy utilizing Gateway technology to produce a library of cloned promoter fragments, and applied this to generate a set of 293 promoter haplotypes from 84 different promoters. The functional significance of the promoter differences was assayed by luciferase reporter assays in HT1080, TE671, HEK293FT and HeLa cell lines. This revealed significant levels of sequence-dependent variation in promoter efficiency, with at least 22% of promoter SNPs having functional consequences. The performance of currently-known putative regulatory elements in retrospectively predicting functional variation was assessed, and found to be wanting. An expansion of upregulatory promoter mutations was noted in the population used, which has implications for the understanding of gene regulatory evolution. Analysis of the whole genome expression profiles of the four cell lines confirmed a qualitative correlation between promoter activity and *in vivo* gene expression, but also indicated that the presence of a known transcription factor binding site could often be ruled out as the mechanism for a functional promoter polymorphism. This study is the most detailed analysis to date of high throughput promoter assays, and is suitable for scaling up to genome-scale functional SNP discovery.

Contents

ABBREVIATIONS AND SYMBOLS.....	IX
ABBREVIATIONS	IX
IUPAC SYMBOLS FOR BASE POSITIONS	X
1 INTRODUCTION.....	1
1.1 TRANSCRIPTIONAL REGULATION.....	3
1.1.1 <i>A bestiary of genomic non-coding regulatory elements</i>	6
1.1.1.1 Promoters	6
1.1.1.2 Enhancers/Silencers	7
1.1.1.3 Insulators.....	8
1.1.1.4 Locus control regions.....	9
1.1.2 <i>Transcription Initiation in Eukaryotes</i>	10
1.1.3 <i>The position of the promoter in the regulatory framework of the cell</i>	11
1.2 THE EUKARYOTIC PROMOTER	12
1.2.1 <i>The Core Promoter</i>	14
1.2.1.1 TATA Box	15
1.2.1.2 Initiator.....	16
1.2.1.3 Downstream Promoter Element (DPE).....	17
1.2.1.4 TFIIB Recognition Element (BRE)	17
1.2.1.5 Motif ten element (MTE).....	18
1.2.2 <i>The Proximal Promoter</i>	18
1.3 IDENTIFYING PROMOTERS	19
1.3.1 <i>Computational approaches</i>	19
1.3.2 <i>Experimental approaches</i>	24
1.4 VARIATION IN PROMOTER SEQUENCES	27
1.5 NATURAL VARIATION IN GENE EXPRESSION LEVELS.....	29
1.6 PROMOTER POLYMORPHISMS IN DISEASE AND EVOLUTION	32
1.7 AIMS OF THIS THESIS.....	34
2 MATERIALS AND METHODS	37
2.1 COMMON BUFFER FORMULAE	38

2.2	CELL CULTURE PROTOCOLS & MEDIA.....	39
2.2.1	<i>Media for HeLa and HT1080 cell lines</i>	39
2.2.2	<i>Media for TE671 and HEK293FT cell lines</i>	39
2.2.3	<i>Passaging Cells</i>	39
2.3	CHAPTER 3 PROTOCOLS	40
2.3.1	<i>Selection of promoters for re-sequencing</i>	41
2.3.2	<i>Primer design</i>	42
2.3.3	<i>Optimisation of genomic PCR</i>	42
2.3.4	<i>High-throughput PCR of promoter fragments</i>	44
2.3.5	<i>Cleanup of PCR products</i>	45
2.3.6	<i>Sequencing of PCR products</i>	45
2.4	CHAPTER 4 PROTOCOLS	45
2.4.1	<i>Creation of pools and design of oligos</i>	45
2.4.2	<i>PCR of promoters from pool templates</i>	46
2.4.3	<i>Gateway cloning into pDONR223</i>	47
2.4.4	<i>Transformation and preparation of pDONR223 haplotype libraries</i>	48
2.4.5	<i>Gateway cloning into pGL3 Basic GW</i>	49
2.4.6	<i>Colony PCR of clones from pGL3 Basic GW haplotype libraries</i>	49
2.4.7	<i>Sequencing of colony PCR products</i>	50
2.4.8	<i>Preparation of plasmids for high-throughput transfection</i>	51
2.4.9	<i>Co-transfection of cell lines with reporter plasmids</i>	52
2.4.10	<i>Assay of firefly and renilla luciferase levels</i>	53
2.5	CHAPTER 5 PROTOCOLS	53
2.5.1	<i>Preparation of total RNA from cell lines</i>	53
2.5.2	<i>Sample preparation and hybridisation on whole genome expression arrays</i> 54	
3	SNP-MINING OF CHROMOSOME 22 PROMOTERS BY RE-SEQUENCING.....	55
3.1	INTRODUCTION	56
3.2	RESULTS	61
3.2.1	<i>Selection of promoters for SNP-mining</i>	61
3.2.2	<i>Primer design</i>	61
3.2.3	<i>Primer tests and PCR optimisation</i>	63

3.2.4	<i>PCR and sequencing of promoter fragments</i>	64
3.2.5	<i>ExoTrace pipeline for sequence analysis and SNP detection</i>	65
3.2.6	<i>Second round of primer design</i>	67
3.2.7	<i>PCR tests of the second batch of primers</i>	68
3.2.8	<i>Promoter sequencing results</i>	69
3.2.9	<i>Distribution of SNP types and allele frequencies</i>	71
3.2.10	<i>Comparison of polymorphic promoters with downstream gene function</i> 73	
3.2.11	<i>Analysis of the genomic context of promoter SNPs</i>	75
3.2.12	<i>Evolutionary analysis of the SNPs using the primate genomes</i>	78
3.2.13	<i>Association of promoter SNPs with gene expression levels</i>	82
3.3	CONCLUSIONS.....	86
4	HIGH-THROUGHPUT CLONING AND REPORTER ASSAYS ON A PROMOTER HAPLOTYPE LIBRARY	90
4.1	INTRODUCTION	91
4.2	RESULTS	97
4.2.1	<i>Experimental strategy</i>	97
4.2.2	<i>Modification of pGL3-Basic to confer compatibility with Gateway technology</i>	99
4.2.3	<i>Selection of target fragments for cloning and functional testing</i>	101
4.2.4	<i>Prediction of promoter haplotypes</i>	103
4.2.5	<i>Construction of DNA pools and PCR of promoter fragments</i>	105
4.2.6	<i>Creation of haplotype libraries</i>	107
4.2.7	<i>Screening haplotype libraries by sequencing</i>	107
4.2.8	<i>Reasons for attrition at each cloning step</i>	108
4.2.9	<i>Successfully cloned promoter SNPs</i>	109
4.2.10	<i>Functional testing of promoter haplotypes with luciferase assays</i>	112
4.2.11	<i>Comparison of promoter activities to transcription start site profile and annotation accuracy</i>	116
4.2.12	<i>Analysis and visualization of haplotype differences</i>	118
4.2.13	<i>Analysis of individual functional SNPs</i>	124
4.2.14	<i>Synergistic effects between functional SNPs</i>	129
4.2.15	<i>Context analysis of functional SNPs</i>	130

4.2.16	<i>Evolutionary analysis of functional polymorphisms.....</i>	132
4.3	CONCLUSION.....	136
5	MICROARRAY ANALYSIS OF THE TRANSCRIPTION FACTOR COMPLEMENT OF TRANSFORMED CELL LINES	143
5.1	INTRODUCTION	144
5.2	RESULTS	148
5.2.1	<i>Preparation and hybridisation of RNA samples from cell lines</i>	<i>148</i>
5.2.2	<i>Normalisation of expression data</i>	<i>149</i>
5.2.3	<i>Quality control of scanned arrays</i>	<i>149</i>
5.2.4	<i>Comparison of endogenous gene expression with cloned promoter activity</i>	<i>153</i>
5.2.5	<i>Correlation of binding sites at functional SNPs with transcription factor expression</i>	<i>160</i>
5.2.6	<i>Classification of cell lines by promoter activity and gene expression... </i>	<i>164</i>
5.2.7	<i>Search for regulatory elements active across the 4 cell lines.....</i>	<i>166</i>
5.3	CONCLUSION.....	172
6	DISCUSSION AND FUTURE WORK.....	177
6.1	DISCUSSION	178
6.2	FUTURE WORK	184
7	REFERENCES.....	188
8	APPENDICES	206
A	<i>GENES TARGETED FOR PROMOTER RE-SEQUENCING.....</i>	<i>207</i>
B	<i>SNPS DISCOVERED IN PROMOTER RE-SEQUENCING.....</i>	<i>210</i>
C	<i>SNPS AND INDELS IN CLONED PROMOTER FRAGMENTS</i>	<i>244</i>
D	<i>HAPLOTYPES CLONED INTO GATEWAY-MODIFIED PGL3 BASIC... </i>	<i>249</i>
E	<i>LUCIFERASE REPORTER ASSAY RESULTS AND SEQUENCE- CONFIRMED HAPLOTYPES</i>	<i>258</i>
F	<i>DE NOVO GENERATED MOTIFS MATCHING JASPAR</i>	<i>289</i>

Abbreviations and Symbols

Abbreviations

ANN	Artificial neural network
ANOVA	Analysis of variance
BRE	TF _{II} B recognition element
CAGE	Cap analysis of gene expression
CAT	Chloramphenicol acetyltransferase
CEPH	Centre d'Etude du Polymorphisme Humain
ChIP	Chromatin immunoprecipitation
DNA	Deoxyribose nucleic acid
DPE	Downstream promoter element
EMSA	Electrophoretic mobility shift assay
ENCODE	ENCyclopaedia Of DNA Elements
EST	Expressed sequence tag
GO	Gene ontology
Indel	Insertion/deletion polymorphism
LCR	Locus control region
LD	Linkage disequilibrium
MTE	Motif ten element
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
Pol II	RNA polymerase II
RLU	Relative light units
RNA	Ribose nucleic acid
mRNA	Messenger RNA
RT-PCR	Reverse transcriptase PCR
SAGE	Serial analysis of gene expression
SELEX	Systematic Evolution of Ligands by EXponential enrichment
SNP	Single nucleotide polymorphism
TAF	TATA-associated factor

TBP	TATA-binding protein
TF	Transcription factor
TFBS	Transcription factor binding site
TF_{II}D	TBP-associated factor II D
TSS	Transcription start site
Tukey's HSD	Tukey's Honestly Significantly Different test
UTR	Un-translated region
VeGA	Vertebrate Gene Annotation

IUPAC Symbols for base positions

IUPAC Code	Meaning
A	A
C	C
G	G
T/U	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	G or A or T or C