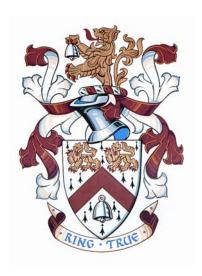
Studies of the effects of promoter sequence variation on gene expression in human chromosome 22

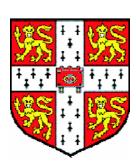
Jamil Bacha



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

A	BBREVI	[ATI	ONS AND SYMBOLS	IX
	ABBREV	IATIO	NS	IX
	IUPAC S	Symb	OLS FOR BASE POSITIONS	X
1	INTR	ODU	ICTION	1
	1.1 T	ΓRANS	SCRIPTIONAL REGULATION	3
	1.1.1	A be	estiary of genomic non-coding regulatory elements	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	Trai	nscription Initiation in Eukaryotes	10
	1.1.3	The	position of the promoter in the regulatory framework of the cell.	11
	1.2 T	Гне Е	UKARYOTIC PROMOTER	12
	1.2.1	The	Core Promoter	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	The	Proximal Promoter	18
	1.3 I	DENT	IFYING PROMOTERS	19
	1.3.1	Con	nputational approaches	19
	1.3.2	Exp	erimental approaches	24
	1.4 V	VARIA	TION IN PROMOTER SEQUENCES	27
	1.5 N	VATUI	RAL VARIATION IN GENE EXPRESSION LEVELS	29
	1.6 P	ROM	OTER POLYMORPHISMS IN DISEASE AND EVOLUTION	32
	1.7 A	AIMS (OF THIS THESIS	34
2	MAT	ERIA	ALS AND METHODS	37
	2.1	Сомм	ON BUFFER FORMULAE	38

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

3.2.4	PCR and sequencing of promoter fragments	64
3.2.5	ExoTrace pipeline for sequence analysis and SNP detection	65
3.2.6	Second round of primer design	67
3.2.7	PCR tests of the second batch of primers	68
3.2.8	Promoter sequencing results	69
3.2.9	Distribution of SNP types and allele frequencies	71
3.2.1	0 Comparison of polymorphic promoters with downstream gene	function
	73	
3.2.1	1 Analysis of the genomic context of promoter SNPs	75
3.2.1	2 Evolutionary analysis of the SNPs using the primate genomes	<i>7</i> 8
3.2.1	3 Association of promoter SNPs with gene expression levels	82
3.3	Conclusions	86
4 HIG	H-THROUGHPUT CLONING AND REPORTER ASSAYS ON	V A
PROMO'	TER HAPLOTYPE LIBRARY	90
4.1	Introduction	91
	RESULTS	
4.2.1		
4.2.2		
	nology	
4.2.3	•	
4.2.4		
4.2.5	• • • • • • • • • • • • • • • • • • • •	
4.2.6	J. J	
4.2.7		
4.2.8		
4.2.9		
4.2.1		
4.2.1		•
and a	annotation accuracy	v
4.2.1	·	
4.2.1	• • • • • • • • • • • • • • • • • • • •	
4.2.1		
4.2.1		130

	4.2.16	Evolutionary analysis of functional polymorphisms	132
4	4.3	Conclusion	136
5	MICI	ROARRAY ANALYSIS OF THE TRANSCRIPTION FACTOR	
		MENT OF TRANSFORMED CELL LINES	143
	5.1 I	NTRODUCTION	144
;		Results	
	5.2.1	Preparation and hybridisation of RNA samples from cell lines	148
	5.2.2	Normalisation of expression data	
	5.2.3	Quality control of scanned arrays	
	5.2.4	Comparison of endogenous gene expression with cloned promoter	
	activi	ty153	
	5.2.5	Correlation of binding sites at functional SNPs with transcription fac	ctor
	expre	ssion	160
	5.2.6	Classification of cell lines by promoter activity and gene expression.	164
	5.2.7	Search for regulatory elements active across the 4 cell lines	166
	5.3	CONCLUSION	172
6	DISC	USSION AND FUTURE WORK	177
(6.1 I	DISCUSSION	178
(FUTURE WORK	
7	REFI	ERENCES	188
8	APPE	NDICES	206
A	G	ENES TARGETED FOR PROMOTER RE-SEQUENCING	207
E	B SI	NPS DISCOVERED IN PROMOTER RE-SEQUENCING	210
(S_{i}	NPS AND INDELS IN CLONED PROMOTER FRAGMENTS	244
I) H	IAPLOTYPES CLONED INTO GATEWAY-MODIFIED PGL3 BASIC.	249
CC		UCIFERASE REPORTER ASSAY RESULTS AND SEQUENCE- IED HAPLOTYPES	258
I	7 D	F NOVO GENERATED MOTIES MATCHING IASPAR	280

Abbreviations and Symbols

Abbreviations

ANN	Artificial neural network	
ANOVA	ANOVA Analysis of variance	
BRE	TF _{II} B recognition element	
CAGE	Cap analysis of gene expression	
CAT	Chloramphenicol acetyltransferase	
СЕРН	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	

ТВР	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	ukey'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
С	С
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
Н	A or C or T
D	A or G or T
В	C or G or T
N	G or A or T or C