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

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