An Analysis of Transcript Variation in Human Xp11.23

By

Tamsin Lee Eades

Thesis submitted for

the degree of Doctor of Philosophy

The Wellcome Trust Sanger Institute

and

Darwin College,

University of Cambridge

August 2005

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

This dissertation does not exceed the word limit set by the Biology Degree Committee.

Tamsin Eades

Abstract

Prior to its completion, it was conjectured that the human genome sequence contained in excess of 100,000 genes, whereas current estimates from the completed genome sequence have reduced this figure to between 20,000 and 25,000. This dramatic reduction could be partially attributed to the phenomenon of alternative splicing, where one gene has the ability to produce more than one functional mRNA transcript.

The aim of this thesis was to obtain a greater appreciation for the diversity of transcripts that can be generated from a single gene. Towards this end, the human genome sequence provided the definitive substrate for the study of transcript variation. A gene-rich region of the X chromosome was selected for analysis, where the frequency and the associated functional consequences of alternative splicing were assessed.

This thesis describes an analysis of the genome features and gene content in human Xp11.22-p11.3. Human Xp11.23 has the highest density of genes in the human X chromosome, and this analysis identified 77 known and 11 novel genes. The pseudogene content of this region was also high - it contained 59 processed and 7 non-processed pseudogenes. More detailed investigation of these revealed that the number and diversity of gene products generated from genes within Xp11.23 greatly exceeded the number of coding regions that it contained. In order to further study the impact of alternative splicing in Xp11.23 detailed analysis was completed on 18 genes using bioinformatic and comparative analysis together with a targeted RT-PCR sequencing strategy. This analysis identified more than 120 transcripts variants. Preliminary tissue profiling of these transcripts was completed using RT-PCR.

The functional consequences of alternative splicing were then investigated for one gene, polyglutamine binding protein 1, *PQBP1*. This ubiquitously expressed gene has been associated with various disease phenotypes including X-linked mental retardation, Renpenning syndrome and other neurodegenerative disorders. In concert with expression and evolutionary analysis, a cloned open reading frame

collection was generated, for 16 transcript variants. The relative abundance of minor transcript variants was determined in a panel of 20 human tissues where it was found that together the minor variants accounted for less than 10% of the all *PQBP1* transcripts.

Following *in silico* analysis of the predicted protein sequences, it was found that transcript variation was associated with variable inclusion of a nuclear localisation signal. Sub-cellular localisation analysis of the transcript variants in CHO-K1 and Cos7 mammalian cell lines showed that three isoforms were not redirected to the nucleus after translation. Further analysis revealed that 11 of the *PQBP1* transcripts harboured a premature termination codon. Preliminary mRNA kinetic assays, and nonsense mediated decay inhibition assays confirmed that at least five of these transcripts promoted rapid degradation via transcript surveillance pathways.

Acknowledgements

I would like to thank my supervisor Mark Ross, for his encouragement, support and advice throughout the entirety of this project. I feel incredibly fortunate to have had a supervisor who was has always made time for me and has taught me the value of meticulous experimental planning, execution and analysis. Special thanks are also extended to Alison Coffey and Colette Johnston who not only proof-read this thesis but also gave valuable advice on *everything*. I have benefited so much from your wisdom, and friendship – thank you.

I am also thankful for the computational support that I have received from numerous teams at the Sanger Institute throughout this project. Carol Scott and Sarah Hunt from human genetics informatics have provided perl scripts and advice on routine computational analysis of large amounts of sequence. Thanks to Darren Grafham for prioritising some human and mouse clones for finishing. James Gilbert and Stephen Keenan from the Institute's informatics team carried out the computational analysis of the genome sequence, while some of the genome annotation described in this thesis was completed by Jennifer Harrow and the human and vertebrate analysis (HAVANA) team at the Sanger Institute.

Thanks are also extended to people who helped me with my experiments. Firstly, to the current and past members of the X chromosome group and DNA collections team, Christine Burrows, Frances Lovell, Ruth Bennett, Eleanor Howard and Hazel Arbury thank you to you all for helping "find my feet" at the Sanger, for your practical advice and for lending solutions/equipment in times of need. I shared a panel of human total RNA samples with Jackie Bye and am extremely grateful and thankful for her advice and patience with all RNA work. Many resources used for the sub-cellular localisation experiments were kindly provided by team 62 (molecular genomics and proteomics). Special thanks are extended to its members especially Ian Dunham, John Collins, Charmain Wright and James Grinham. Philippe Couttet, gave me much needed advice on NMD and mRNA stability experiments and provided me with the pTRE-TIGHT (tet-off) plasmid. Thanks also to Ian Barrett and Gareth Howell for giving me such a friendly introduction to the Sanger Institute.

Finally, thank you to all of my friends and family who have been so patient and tolerant over the past 3 ½ years. It's now time to repay many favours and perhaps I will now get to spend a Christmas at home!!! I am especially indebted to my parents and to Graeme, I know that I wouldn't have made it to this stage without you, thank you.

This project was funded by the Wellcome Trust and the Sanger Institute.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abbreviations

1st EF	First exon finder
ABI	Applied Biosystems
ACeDB	A *C. elegans* database
ACTB	beta actin
AEDB	alternative exon database
ANS	anisomycin
ASAP	the Alternative Splicing Annotation Project
ASD	alternative splicing database project
ASDB	alternative splicing database
ASG	alternative splicing gallery
AT	annealing temperature
ATP	adenosine 5'- triphosphate
BAC	bacterial artificial chromosome
bis-acrylamide	(N, N'-methylene-)bis-acrylamide
BLAST (-n -p)	basic local alignment search tool (-nucleotide -protein)
BLAT	basic local alignment tool
Blixem	BLast matches In an X-windows Embedded Multiple alignment
BMD	Becker muscular dystrophy
ß-ME	ß-mercaptoethanol
bp	base pair
BSA	bovine serum albumin
ºC	degrees Celsius
cDNA	complementary DNA
CDS (c-)	coding sequence (consensus)
CEN	centromere
CHX	cycloheximide
CNS	central nervous system
CpG	cytidyl phosphoguanosine dinucleotide
Cps	counts per second
C_T	cycle threshold
CT- antigen	cancer testis antigen
CTP (d-)	cytidine 5'-triphosphate (deoxy-)
DAPI	4',6-Diamidino-2-phenylindole
dbEST	database of expressed sequence tags

DDBJ	DNA DataBase of Japan
DEPC	diethyl pyrocarbonate
DKFZ	Deutsches Krebs Forschung Zentrum (German cancer research centre)
DMD	Duchenne muscular dystrophy
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
Dnase I	deoxyribonuclease A
dNTP	deoxyribonucleotide 5' triphosphate
dsDNA	double-stranded deoxyribonucleic acid
dsRNA	double-stranded ribonucleic acid
EBI	European Bioinformatic Institute
ECL	enhanced chemiluminescense
ECR	evolutionary conserved region
EDTA	ethylenediamine tetra-acetic acid
EJC	exon junction complex
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Open Software Suite
ePCR	electronic polymerase chain reaction
ES cell	embryonic stem cell
ESE	exon splicing enhancer
ESS	exon splicing silencer
EST	expressed sequence tag
FITC	fluorescein isothiocyanate
g	gram
g	force of gravity (relative centrifugal force)
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
G banding	Geisma banding
GTP (d-)	guanosine 5'-triphosphate (deoxy-)
HASDB	human alternative splicing database
HAVANA	human and vertebrate genome annotation and analysis
HEPES	N-[2-hydoxyethyl]piperazin-N'-[2-ethansulphonic acid]
HGMP	Human Genome Mapping Project
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
hnRNP	heterogenous nuclear ribonucleoprotein
HRP	horseradish peroxidase

HMM	hidden Markov model
HS	hierarchical shotgun
HUGO	Human genome organisation
ICI	Imperial Chemical Industries
INDELs	insertion or deletion of deoxyribonucleic acid
IHGSC	International Human Genome Sequencing Consortium
IPTG	isopropyl ß-D-thiogalactoside
kb	kilobase pairs
kDa	kilodalton
l	litre
LB	Luria-Bertani
LINE	long interspersed nuclear element
LTR	long terminal repeat
M	molar
Mb	megabase pairs
MER	medium reiterative repeat
MGC	Mammalian Gene Collection
µg	microgram
µl	microlitre
µM	micromolar
min(s)	minute(s)
mg	milligram
ml	millilitre
mM	millimolar
mm	millimetre
mya	million years ago
NCBI	National Centre for Biotechnology Information
NEDO	New Energy and Industrial Technology Development Organisation
ng	nanogram
nm	nanometre
NMD	nonsense mediated decay
O/N	overnight
oligo-dT	deoxyribothymidyl oligonucleotide
OMIM	On-line Mendelian Ineritance in Man
ORF	open reading frame
PAC	P1-derived artificial chromosome

PAGE	polyacrylamide gel electrophoresis
PBS (-T)	phosphate buffered saline (-Tween 20)
PCR	polymerase chain reaction
RASL	RNA annealing selection and ligation
pg	picogram
PALSdb	putative alternative splicing database
PE	Perkin Elmer
poly(A)	polyadenylation
PTC	premature termination codon
PUR	puromycin
RNA	ribonucleic acid
RNAi	ribonucleic acid interference
mRNA	messenger ribonucleic acid
rRNA	ribosomal ribonucleic acid
snRNA	small nucleolar ribonucleic acid
hnRNA	heteronuclear ribonucleic acid
hnRNP	heteronuclear ribonucleoprotein
snRNP	small nucleolar ribonucleoprotein
dsRNA	double-stranded ribonucleic acid
siRNA	short-interfering ribonucleic acid
tRNA	transfer ribonucleic acid
miRNA	micro-ribonucleic acid
RISC	ribonucleic acid induced silencing complex
RNase A	ribonuclease A
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
SAGE (Long-)	serial analysis of gene expression
SDS	sodium dodecyl sulphate
s	seconds
siRNA	short interfering RNA
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
SSAHA	sequence Search and Alignment by Hashing Algorithm
ssDNA	single-stranded ribonucleic acid
SSH	suppressive subtractive hybridisation

SSX	synovial sarcoma X-linked
STS	sequence tagged site
TAP	transcript assembly programme
tBLAST -n, -x	translated blast local alignment sequence tool -vs nucleotide database -vs protein database
TLC	thin layer chromatography
TEL	telomere
TEMED	N,N,N',N'-tetramethylethylenediamine
TFB (-I, -II)	Transformation buffer -I, -II
TIGR	the Institute for Genomic Research
Tris	tris(hydroxymethyl)aminoethane
TSS	transcription start site
TTP	thymidine 5'-triphosphate
TTS	transcription termination site
U	unit
UCSC	University of California, Santa Cruz
UTR	untranslated region
uv	ultraviolet
V	volt
VEGA	vertebrate genome annotation
v/v	volume/volume
W	watt
w/v	weight/volume
WAS	Wiskott Aldrich syndrome
Wash U.	Washington University
WGS	whole genome shotgun
WTSI	Wellcome Trust Sanger Institute
Xace	X chromosome version of ACeDB
X-gal	5-bromo-4-chloro-3-indolyl-ß-D-galactoside
XIC	X-inactivation centre
Xist	X inactive specific transcript
YAC	yeast artificial chromosome