Gene identification and characterisation on the human X chromosome

Ian Barrett

Darwin College, University of Cambridge

January 2004

This dissertation is submitted for the degree of Doctor of Philosophy.

This dissertation 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 size limit for the Biology Degree Committee.

Abstract

This thesis investigates the gene composition and evolution of regions of the human X chromosome, including data from comparative genome analysis of other organisms.

Chapter Three presents studies undertaken to annotate genes within the Xq22-q23 region of the human X chromosome. Selected features of the region are discussed, including investigation of alternative polyadenylation site usage, an insertion of the mitochondrial genome into the nuclear genome, and an inverted duplication and potential gene fusion event involving the NXF2 and TCP11-like genes.

As a result of the annotation described in Chapter Three, extensive paralogy within the Xq22 region was discovered, along with additional examples of paralogy between Xp and Xq22-q23. Work in subsequent Chapters attempted to characterise these aspects further and provide information on the evolution of the regions.

Chapter Four describes work undertaken to map and sequence the region of the mouse genome corresponding to human Xq22-q23, in order to investigate the evolution of the Xq22 paralogues. Annotation of genes within the region of the mouse X chromosome and the orthology of the human and mouse regions is described. Features of the mouse region, such as the presence of two large repeat families, are also discussed.

Chapter Five presents phylogenetic and expression profile analysis of the Xq22 paralogues, and examines orthology in the corresponding region of the mouse genome.

Chapter Six includes a discussion of Xp/Xq paralogy, and presents studies providing evidence for a segmental duplication leading to this paralogy. In addition, orthologues of the genes involved in the paralogy are identified in the marsupial mouse, *Sminthopsis macroura,* and their genomic localisations determined. Evidence suggesting a minimum age of the duplication is presented. Comparative analysis of human, mouse, *Fugu rubripes* and *Sminthopsis macroura* genomic sequence is described.

Acknowledgements

I don't really know how to begin thanking everyone. I've been extremely fortunate to have benefited from the kindness, help, advice and expertise of so many people during these studies, and have had great times along the way (as well as numerous genome T-shirts). The selfless help I've received from so many people really has been very touching and humbling.

Firstly, thanks to Tim, David and Andy for your support and understanding over the last months. Thank you also to the BGS and Education Committee for their support. Financial support from the MRC and The Wellcome Trust is gratefully acknowledged.

A huge thanks to all the Sanger Institute cloning, sequencing and finishing teams for all their work spanning sequence from human, mouse and that furry little fella, the marsupial mouse – without you none of this work would have been possible. Thanks especially to Kirsten McLay, Darren Grafham, Christine Bird, Siobhan Whitehead and Lucy Matthews for answering so many of my queries. Thanks as well to Sean Humphray and Glen Threadgold for all the help with the mouse mapping. A big thanks to Paul Hunt for putting up with me and for all your gridding and being generally very helpful, and to Carol Carder for showing me the ropes, and letting me upset the carefully balanced workings of your lab! Thanks as well to the now HAVANA group, for the sequence analysis, especially Stephen Keenan, whom I badgered on many occasions.

I'd also like to thank Jess Tyson, Karen Woodward and Willem Rens for our collaborations, which I enjoyed very much, and from which I learned a great deal. A particular thanks to Willem for the gifts of resources and advice, and your infectious enthusiasm!

Thank you Jackie Bye and Susan Rhodes for all the work you put into the cDNA resources, and especially to you Jackie for all the advice and encouragement, and our discussions! I'm watching out for that next invention… Thank you to John Collins and David Beare for all the advice on the transcript mapping, and for paving the way.

A very big thanks to all the FISH gang and Nigel Carter. I had a great time as an "imposter" in the lab. Thanks very much to you all for your help in times of need in getting things done. I won't forget my last day at the Sanger, checking out some FISH images (nor my special FISH-cake Louisa….!). Thanks particularly to Ruby Banerjee, Susan Gribble and especially to Deb Burford for really coming through when things were tight (hope I can repay you in some way in your PhD studies).

Thanks to Christina Hedberg-Delouka and Christine Rees for keeping me in check (as much as you could…) and also to Joan Green and Andrew King for help and advice on all things library-related (and sorry for the often late returns..!). Thanks to Richard Summers as well for all the reprographics support and photos for my leaving gifts.

Thanks to Kate Rice and David Bentley for the chats and advice, I hope HumGen continues to go from strength to strength. I feel very lucky to have been part of it. Thank you to David Vetrie, Stephan Beck and Nabeel Affara for helpful advice. A big thanks to Cords, Matt, Rob and Em for being good mates and also to Paw and Vicki for a lot of laughs and chats. Thanks to "Tammar" Tamsin as well for all the encouragement and chats, and congratulations on your rowing achievements. Can ya tell wot it is yit? "Frantastic" Fran Lovell and Christine Burrows, I especially owe you guys a great deal in all the help you gave me when times were tight (and even when they weren't). Thanks also to Adam Whittaker for all the hyb jokes, and Jen Conquer, Helen Steingruber and Barbara Gorick for all the office banter! Thanks very much

to “Pod’s Angels” Hazel Arbery and Ruth Bennett as well, for your real efforts in the last push. You’re real stars.

Carol Scott, I bet you’ll be glad I’ve finished! I can’t thank you enough for all the remote-control help you’ve given me, and thanks also to you and Sarah Hunt for all the help while I was actually a genuine member of the Sanger Institute.

Gareth “Gazza” Howell, thanks for all you’ve taught me and for all the clowning around, although I’m glad to be able to walk the corridors without apprehension…. On a more serious note, thanks for all the work you put into Xq22, which formed the basis for this work. Thanks as well to Simon Gregory for the help, advice and laughs, just a shame you weren’t around for the rugby this year…! Thanks both for all your advice and friendship, and for all the good times (plenty more to come).

Pod (a.k.a. Alison Coffey), you really have been a star. I’ve learned so much from you and you really have been an inspiration in more ways than one. Thanks for everything, and especially for all the help in the final push – definitely couldn’t have done it without you.

Mark Ross, what can I say? You’ve taught me a lot over the last years, more than you’ll probably realise and not just good science. You’re an excellent scientist to learn from and a great mentor. Thanks very much from both of us for all you’ve done.

People often say they couldn’t have done it without you etc. but in this case especially I really couldn’t. I’ve been helped by, and learned from, more people than I’ve been able to mention here. I’ll always be grateful to everyone (you and I know who you are) for all your help, and for all the fantastic memories I’ve got from my time at the Sanger, a very special place.

And now thanks to all our family and friends for all their support over the years and getting me this far. Thanks to Andy for the chats and to Mark for being a good mate.

Thank you especially to my long-suffering wife Gill (well, not long yet but pretty suffering) for all your support and love over good times and bad and for all your help, especially your help with the final printing and formatting. You’ve really held things together over the last year, which has truly tested us both. Can’t wait to get on with the rest of our married life, and see what it holds for us.

Thank you to Dad for being understanding when I didn’t feel like giving an update on progress and for all your love, support, interest and encouragement. You really are someone to look up to.

And finally, a very special thanks to Mum for all that you’ve given and taught me. There just isn’t enough space to put it all here. You’re an incredible inspiration and very much loved. I’d like to dedicate this work to you and Dad. Sleep well.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Chapter 6

List of Tables

Chapter 5

Chapter 6

Abbreviations

µg	microgram
µl	microlitre
µM	micromolar
µm	micrometre
^{0}C	degrees celsius
ACeDB	A *C. elegans* Database
AT	annealing temperature
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
Bp	base pair
cDNA	complementary deoxyribonucleic acid
CEN	centromere
cm	centimetre
cM	centimorgan
CpG	cytidyl phosphoguanosine dinucleotide
Ctg	contig
DDC	duplication degeneration complementation
DDW	double distilled water
DMD	duchenne muscular dystrophy
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridisation
g	gram or force of gravity
HAVANA	Human And Vertebrate ANalysis and Annotation
HTGS	high throughput genomic sequence
HUGO	Human Genome Organisation
kb	kilobase pairs
LCR	low copy repeat
LD	linkage disequilibrium
LINE	long interspersed nuclear element
LTR	long terminal repeat
M	molar
Mb	megabase pairs
MHC	major histocompatability complex
MIR	medium interspersed repeat
mL	millilitre
mM	millimolar
mm	millimetre
mRNA	messenger ribonucleic acid
Mya	million years ago
NCBI	National Centre for Biotechnology Information
ng	nanogram
nm	nanometre
numts	nuclear mitochondrial insertions
OMIM	Online Mendelian Inheritance in Man
ORF	open reading frame
PAC	P1-derived artificial chromosome

PAR	pseudoautosomal region
PBS	phosphate-buffered saline
PC	personal computer
PCR	polymerase chain reaction
RACE	rapid amplification of cDNA ends
RFLP	restriction fragment length polymorphisms
RH	radiation hybrid
RNA	ribonucleic acid
rpm	revolutions per minute
RT	reverse transcriptase
RT-PCR	reverse transcriptase-polymerase chain reaction
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
SSPCR	single-sided specificity polymerase chain reaction
STS	sequence tagged site
TEL	telomere
U	units
UTR	untranslated region
UV	ultra-violet
V	volts
VEGA	VErtebrate Genome Annotation
WGS	whole-genome shotgun
XAR	X added region
XCR	X conserved region
Xic	X inactivation centre
XIST	X inactivation specific transcript
YAC	yeast artificial chromosome