Physical, Transcriptional and Comparative Mapping on the Human X Chromosome

by

Gareth Rhys Howell

Thesis submitted for the

degree of Doctor of Philosophy

The Open University

1st March 2002

The Wellcome Trust Sanger Institute

Wellcome Trust Genome Campus

Hinxton

Cambridge, UK

This thesis is dedicated to my wife, Andrea

and to my daughters, Megan and Cadi.

Abstract

Progress in the study of the human genome has resulted in the production of a complete clone map and associated 'working draft' sequence. This will underpin the completion of the sequence itself, the annotation of genes and other features, and application of this new found knowledge. This thesis focuses on the evolving methods to determine the map, and to use the emerging sequence for the study of genes, incorporating new studies of other genomes to enhance progress in understanding and interpretation to biology and medicine. The success of the endeavours is necessarily accompanied by the development and evolution of new technologies and by critical assessment of the progress in acquiring knowledge of the genomic information.

Evolution of mapping technologies included the development of the larger insert bacterial cloning systems (PACs) and (BACs), and an increase in available landmarks both from YAC maps and RH maps. The work described in chapter 3, followed this evolution and was applied to construct a 6 Mb sequence-ready bacterial clone contig map in Xq22. A minimum set of clones was chosen for genomic sequencing. The resulting sequence map was compared to previously published maps and analysed both for common repeats, and previously unidentified low copy repeats.

The availability of the emerging sequence of the human genome provided a resource for identification of the features encoded within. In chapter 4, the sequence of a 7 Mb region in Xq23-24 was analysed for the presence of genes using a combination of sequence similarity searches against both protein and DNA databases, and *ab initio* gene prediction. Predicted genes were confirmed where possible, by generating novel

cDNA sequence. The region contained 33 confirmed genes (of which 14 were confirmed during this study), 11 predicted genes and 20 pseudogenes.

Comparative genome sequence analysis is a powerful method both for aiding human gene identification and identifying other features encoded within the human genome such as regulatory elements. Comparing the genomes of two or more species also provides insights into the evolution of the species since the divergence from a common ancestor. Sequence from a 1 Mb region in human Xq24 was compared in two other species, mouse and zebrafish. In chapter 5, bacterial clone contigs for sequencing were constructed in the mouse by designing mouse-specific STSs orthologous to human sequence for clone isolation. In chapter 6, bacterial clone contigs for sequencing were constructed in zebrafish using STS from exons of human genes to identify zebrafish BAC clones by reduced stringency hybridisation. Comparative analysis of the region showed that humans and mice are more highly conserved than humans and zebrafish, in terms of gene content and organisation. A combination of comparative sequence analysis tools identified 14 novel potential conserved sequences between human and mouse, one of which was also conserved in zebrafish.

Acknowledgements

I would like to thank my supervisors David Bentley and Mark Ross for all the help, advice and guidance provided throughout the course of this thesis. On a number of occasions career development 'chats' were required to ensure it kept on track and for that I am most grateful. You have both ensured I have stayed focussed whilst doing a part-time PhD.

A number of people have made invaluable contributions, without which this thesis would not have been possible. Particular thanks go to The Sanger Institute sequencing teams: Darren Graffham, Christine Bird and other members of the X chromosome finishing team, particularly because Xq22 proved a horrid region to finish. Christine spent many hours toying with cosmids, solving finishing problems caused by duplications and deletions – it wasn't my fault! Adrienne Hunt also deserves a special mention for work on both human, but also zebrafish clones, rushing finishing through to provide me with some sequence to work with. All FISH analyses were carried out by the FISH group and in particular, Pawandeep Dhami, whose life would have been much easier if I had not required 'just a few' fibrefish experiments! Jackie Bye for cDNA resources, general advice and proof reading. Sarah Hunt and Carol Scott for informatics support. The Chromosome 22 group, including John Collins, Dave Beare and Ian Dunham - "Much of the sequence analysis in this thesis would not have been possible without the advice and pearl scripts from these three wise men" (is that what you told me to put?).

Thanks go to the members of the Experimental Gene Annotation Group; Graeme, Jackie, Kevin, Liz and Kate, for allowing me the time to finish this thesis. You have all kept the group going. Maybe you'll see me in the lab soon?

Personal thanks go to my friends. Pod, I promised you a special mention and hopefully this is good enough. You have taught me so much over the years about all sorts of things and supported me all the way. Your skills relating to lab work and thesis writing, reading and checking were invaluable. You certainly know how to proof read a thesis! Also, your unprintable one-liners have kept the amusement levels high! Ian and Tamsin, fellow PhD students, for your energies, discussions and proof reading and also for cheering me up when on the odd occasion I have felt a little bit grumpy. Apologies to Ian for deleting his gene structures and changing the Xq22 map at will. Cords, without the tea and bacon sandwich interruptions, thesis writing would have been much duller. Dave, you were there at the start of the thesis (1400 cosmids – thanks!) and at the end (proof reading), thanks for all your help. And Simon – you have been there as a friend and confidant throughout my time at the Sanger Institute, without your straight talking and competitiveness who knows where I would be?!

Final thanks go to my family: My wife Andrea, who has been there for me throughout and believed in me all the way. You have given encouragement whenever needed, even when not always understanding a word of what I was saying! And Dad and Mum, for your will, passion and enthusiasm.

To every one mentioned, and no doubt many I've omitted, I owe you a pint (or a half – I am Welsh after all!)

Table of Contents page

List of Figures:

Chapter 4

Chapter 5:

List of Tables:

Glossary of Abbreviations

ACeDB	*A C. elegans* database
ANT2	adenosine nucleotide transporter 2
Alu-PCR	*Alu*-element-mediated polymerase chain reaction
ATP (dATP, ddATP)	adenosine 5'-triphosphate (deoxy-, dideoxy-)
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
BLIXEM	BLAST In an X-windows Embedded Multiple Alignment
β-ME	β-mercaptoethanol
bp	base pair
BSA	bovine serum albumin
BTK	Bruton's tyrosine kinase
^{o}C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
chr	chromosome
cM	centiMorgan
cm	centimetre
CpG	cytidyl phosphoguanosine dinucleotide
cpm	counts per minute
cR	centiRays
CTP (dCTP, ddCTP)	cytidine 5'-triphosphate (deoxy-, dideoxy-)
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphosphate
DTT	dithiothreitol
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridisation
FP	forward primer
FPC	Fingerprinting Contig

g	gram
G banding	Geimsa banding
GDB	Genome Database
GSC	Genome Sequencing Centre, St Louis
GTP (dGTP, ddGTP)	guanine 5'-triphosphate (deoxy-, dideoxy-)
HGMP	Human Genome Mapping Resource Centre
HGP	Human Genome Project
HMM	Hidden Markov Model
HPRT	hypoxanthine phosphoribosyltransferase
kb	kilobase pairs
l	litre
LAMP2	lysosomal-associated membrane protein 2
LB	Luria-Bertani
LD	linkage disequilibrium
LINE	long interspersed nuclear element
M	molar
Mb	megabase pairs
μg	microgram
μl	microlitre
μM	micromolar
min(s)	minute(s)
mg	milligram
ml	millilitre
mm	millimetre
mM	millimolar
MRX	X-linked non-specific mental retardation
NSMR	Non-specific mental retardation
NCBI	National Centre for Biotechnology Information
ng	nanogram
nm	nanometre
O/N	overnight
OD	optical density
OMIM	On-line Mendelian Inheritance in Man

PAC	P1-derived artificial chromosome
PAR	pseudoautosomal region
PCR	polymerase chain reaction
PFAM	Protein Family
PFGE	pulsed-field gel electrophoresis
pg	picogram
plp	proteolipid protein
PMD	Pelizaeus Merchbacher Disease
pmol	picomole
poly(dT)	poly-deoxyribothymidyl oligonucleotide
R banding	Reverse Geimsa banding
RH	radiation hybrid
RFLP	restriction fragment length polymorphism
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
RP	reverse primer
Rnase A	ribonuclease A
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
SDS	sodium dodecyl sulphate
sec(s)	second(s)
seq	sequence
SINE	short interspersed nuclear element
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SSPCR	single-sided specificity PCR
STS	sequence tagged site
TEMED	N,N,N',N'-tetramethylethylenediamine
TrEMBL	Translated EMBL
TIGR	The Institute of Genome Research
Tris	tris(hydroxylmethyl)aminomethane
U	unit
UTR	untranslated region
uv	ultraviolet

V	volt
v/v	volume/volume
VNTR	variable number of tandem repeats
W	watt
w/v	weight/volume
Wash U.	Washington University
WGS	whole genome shotgun
Xace	X chromosome version of ACeDB
XCI	X chromosome inactivation
XIC	X-inactivation centre
Xist	X inactive specific transcript
XLA	X-linked agammaglobulinaemia
XLMR	X-linked mental retardation
XLP	X-linked lymphoproliferative disease
YAC	yeast artificial chromosome

Publications

Parts of this work presented in this thesis have appeared previously in the following publications:

Bentley, D. R., Deloukas, P., Dunham, A., French, L., Gregory, S. G., Humphray, S. J.*, et al.* (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J.*, et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.