

**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**

**1<sup>st</sup> 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 Dhani, 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!)

<b>Table of Contents</b>	<b>page</b>
Abstract	iii
Acknowledgments	v
Table of Contents	vii
List of Figures	xii
List of Tables	xvii
Glossary of abbreviations	xxi
Publications	
<b>Chapter One: Introduction</b>	<b>1</b>
1.1 Mapping and sequencing of model organisms	2
1.2 Mapping and sequencing the human genome	6
1.3 Interpreting the human genome sequence	17
<i>1.3.1 Gene identification</i>	19
1.4 The human X chromosome	33
1.4.1 <i>Xq22</i>	37
1.4.2 <i>Xq23-24</i>	38
1.4.3 <i>Non-specific X-linked mental retardation</i>	38
1.5 Aims of this thesis	39
<b>Chapter Two: Materials and Methods</b>	<b>41</b>
<u>Materials</u>	44
2.1 Chemical reagents	44
2.2 Enzymes and commercially prepared kits	44
2.3 Nucleotides	44
2.4 Solutions	45
2.4.1 <i>Buffers</i>	45
2.4.2 <i>Electrophoresis and Southern blotting solutions</i>	46
2.4.3 <i>Media</i>	47
2.4.4 <i>DNA labelling and hybridisation solutions</i>	48
2.4.5 <i>General DNA preparation solutions</i>	48

2.5	Size markers	49
2.6	Hybridisation membranes and X-ray and photographic film	49
2.7	Sources of genomic DNA	49
2.8	Bacterial clone libraries	50
	2.8.1 <i>Cosmid libraries</i>	50
	2.8.2 <i>PAC and BAC libraries</i>	50
	2.8.3 <i>cDNA libraries</i>	50
2.9	Primer sequences	51
2.10	World Wide Web addresses	63
	<u>Methods</u>	64
2.11	Isolation of bacterial clone DNA	64
	2.11.1 <i>Miniprep of cosmid, PAC and BAC DNA</i>	64
	2.11.2 <i>Microprep of cosmid, PAC and BAC DNA for restriction digest fingerprinting</i>	64
2.12	Bacterial clone fingerprinting	66
	2.12.1 <i>Radioactive fingerprinting</i>	66
	2.12.2 <i>Fluorescent fingerprinting</i>	66
	2.12.3 <i><u>Hind</u> III fingerprinting</i>	67
2.13	Marker preparation	68
	2.13.1 <i>Radioactive fingerprinting</i>	68
	2.13.2 <i>Fluorescent fingerprinting</i>	68
	2.13.3 <i><u>Hind</u> III fingerprinting</i>	69
2.14	Gel preparation and electrophoresis	69
	2.14.1 <i>Agarose gel preparation and electrophoresis</i>	69
	2.14.2 <i>Gel preparation and electrophoresis for radioactive fingerprinting</i>	69
2.15	Applications using the polymerase chain reaction	70
	2.15.1 <i>Primer design</i>	70
	2.15.2 <i>Oligonucleotide preparation</i>	71
	2.15.3 <i>Amplification of genomic DNA by PCR</i>	71
	2.15.4 <i>Colony PCR of STSs from bacterial clones</i>	71



2.16	Radiolabelling of DNA probes	72
	2.16.1 <i>Random hexamer labelling</i>	72
	2.16.2 <i>Radiolabelling of PCR products</i>	72
	2.16.3 <i>Pre-reassociation of radiolabelled probes</i>	73
2.17	Hybridisation of radiolabelled DNA probes	73
	2.17.1 <i>Hybridisation of DNA probes derived from whole cosmids</i>	73
	2.17.2 <i>Hybridisation of DNA probes derived from STSs</i>	73
	2.17.3 <i>Hybridisation of DNA probes to gridded zebrafish library</i>	74
	2.17.4 <i>Stripping radiolabelled probes from hybridisation filters</i>	74
2.18	Restriction endonuclease digestion of cosmid DNA	74
	2.18.1 <i>Restriction endonuclease digestion of cosmid DNA</i>	74
	2.18.2 <i>Restriction endonuclease digestion of PAC or BAC DNA</i>	75
2.19	Generation of vectorette libraries of PACs and BACs	75
2.20	Rescue of clone ends by PCR amplification of vectorette libraries	74
2.21	Preparation of high density colony grids	76
2.22	Clone library screening	76
	2.22.1 <i>Bacterial clone library screening</i>	76
	2.22.2 <i>cDNA library screening by PCR</i>	77
	2.22.3 <i>Single-sided specificity PCR (SSPCR) of cDNA</i>	78
	2.22.4 <i>Vectorette PCR on cDNA</i>	81
	2.22.5 <i>Reamplification of vectorette PCR products</i>	81
2.23	Mapping and sequence analysis software and databases	84
	2.23.1 <i>IMAGE</i>	84
	2.23.2 <i>FPC</i>	84
	2.23.3 <i>Xace</i>	85
	2.23.4 <i>BLIXEM</i>	86
	2.23.5 <i>RepeatMasker</i>	87
<b>Chapter Three: Construction of a Sequence-Ready Bacterial Clone Contig</b>		<b>88</b>
3.1	Introduction	89
3.2	Contig construction	91
3.3	Comparison of the published maps	106
	3.3.1 <i>Genetic Map</i>	106
	3.3.2 <i>RH map</i>	108

3.3.3	<i>YAC maps</i>	111
3.4	Sequence composition and repeat content analysis	113
3.4.1	<i>Sequence composition analysis</i>	113
3.4.2	<i>Analysis of previously identified low copy repeats</i>	116
3.4.3	<i>Analysis of previously unidentified low copy repeats</i>	117
3.4.4	<i>Analysis of clone instability</i>	123
3.5	Discussion	125
<b>Chapter Four: Genome Landscape of Xq23-Xq24</b>		<b>131</b>
4.1	Introduction	132
4.2	Identification of genes	133
4.3	Evaluation of genes in region	155
4.3.1	<i>Evaluation of the 5' ends</i>	159
4.3.2	<i>Evaluation of the 3' ends</i>	159
4.3.3	<i>Alternative Splicing</i>	162
4.3.4	<i>Genes in their genomic context</i>	165
4.4	Predicting the function of novel gene products	171
4.5	Analysis of the sequence composition of the region in Xq23-Xq24	179
4.6	Mutation screening for MRX23	185
4.7	Discussion	191
4.8	Appendix	196
<b>Chapter 5: Comparative sequence analysis between human and mouse</b>		<b>201</b>
5.1	Introduction	202
5.2	Construction of bacterial clone contig	207
5.3	Identification of orthologous genes in the region	217
5.4	Comparison of the genome landscape in human and mouse	226
5.5	Analysis of conserved sequences	229
5.5.1	<i>Evaluating the methods for sequence comparison</i>	229
5.5.2	<i>Potential function for novel conserved sequences</i>	238

5.6	Evaluation of whole genome shotgun (WGS)	241
5.7	Discussion	245
5.8	Appendix	249
<b>Chapter 6: Comparative sequence analysis between human and zebrafish</b>		<b>250</b>
6.1	Introduction	251
6.2	Identification of zebrafish genomic clones	254
6.3	Evaluation of strategy for the identification of orthologous genes	259
6.4.	Identification of BAC clones using orthologous zebrafish EST sequence	263
6.5	Sequence analysis	267
6.6	Identification of 20 novel repeat elements in the zebrafish genome	277
6.7	Multiple sequence analysis	280
6.8	Discussion	288
6.9	Appendix	292
<b>Chapter 7: Discussion</b>		<b>296</b>
7.1	Advances in mapping technology and strategy	297
7.2	Mining the human genome sequence	302
7.3	Comparing different genomes to aid human genome sequence analysis	308
7.4	Functional analysis of gene products	310
7.5	Conclusion	316
<b>Chapter 8: References</b>		<b>317</b>

**List of Figures:****Chapter 2**

Fig. 2.1	Strategy for SSPCR on cDNA libraries	80
Fig. 2.2	Strategy for vectorette PCR on cDNA libraries	83

**Chapter 3**

Fig. 3.1	The status of the region of interest before the generation of the bacterial clone contig began	90
Fig. 3.2	Strategy for the construction of the bacterial clone contig.	92
Fig. 3.3	Cosmid fingerprinting and assembly	93
Fig. 3.4	Fingerprinting of DXS101-positive cosmids	95
Fig. 3.5	PAC isolation by whole cosmid hybridisation	97
Fig. 3.6	PAC isolation using STSs taken from YAC map of Srivastava, A. K., <i>et al.</i> (1999)	99
Fig. 3.7	PAC isolation using STSs generated by vectorette PCR or end sequencing	103
Fig. 3.8	FPC diagram of bacterial clone contig between DXS366 and DXS1230	105
Fig. 3.9	Comparison of the genetic map	107
Fig. 3.10	Comparison of the gene map	110
Fig. 3.11	Comparison of the YAC map	112
Fig. 3.12	A graph showing the relative abundance of the GC content, LINES and SINES across the region of interest.	115
Fig. 3.13	An image from the computer program ACT, showing the position of low copy repeat sequences within the final sequence map	119
Fig. 3.14	Analysis of 140 kb indirect repeat	122
Fig. 3.15	Analysis of clone instability showing the region around DXS24 and the status of the mapping	124
Fig. 3.16	Status of mapping at each stage of contig construction	126

**Chapter 4**

Fig. 4.1	Status of the region between Xq21.3 and Xq25	134
Fig. 4.2	Genomic sequence analysis	135
Fig. 4.3	ACEDB and BLIXEM	137
Fig. 4.4	Examples of features for which STSs were designed for cDNA isolation	140
Fig. 4.5	cDNA isolation by SSPCR	147
Fig. 4.6	cDNA isolation by vectorette PCR	148
Fig. 4.7	Confirmation of novel gene	150
Fig. 4.8	Example of a pseudogene	152
Fig. 4.9	A summary of the gene map between DXS7598 and DXS7333	154
Fig. 4.10	Evaluation of gene structures	161
Fig. 4.11	Genes in their genomic context (1)	164
Fig. 4.12	Genes in their genomic context (2)	167
Fig. 4.13	Analysis of 50 kb duplication	170
Fig. 4.14	Functional analysis of genes	177
Fig. 4.15	Genome landscape of the region of interest	181
Fig. 4.16	Unclassified diseases mapping to region of interest	187
Fig. 4.17	Mutation screening for MRX23	189
Fig. 4.18	Identification of a potential silent mutation	193
Fig. 4.19	Contributions of cDNA sequencing projects and prediction programs	195
Fig. 4.20	Examples of unconfirmed genes	195

**Chapter 5:**

Fig. 5.1	A schematic representation of the syntenic relationship between human and mouse	204
Fig. 5.2	Summary of the region for comparative analysis	206
Fig. 5.3	Examples of alignment between human and mouse orthologues	208
Fig. 5.4	Strategy for contig construction	209
Fig. 5.5	BAC clone isolation with mouse-specific STSs	211
Fig. 5.6	Contig construction by fingerprinting	212
Fig. 5.7	Summary of the mapping	214
Fig. 5.8	Summary of the gene map constructed in mouse	216

Fig. 5.9	Comparative analysis of the region in human and mouse	219
Fig. 5.10	Comparative analysis of novel orthologous genes	221
Fig. 5.11	Analysis of the homeobox genes	223
Fig. 5.12	Comparative analysis of the genome landscape in human and mouse	228
Fig. 5.13	Examples of comparative sequence analysis tools	232
Fig. 5.14	Identification of conserved sequences	237
Fig. 5.15	Analysis of the promoter region of ANT2	240
Fig. 5.16	Evaluation of whole genome shotgun	243
Fig. 5.17	Analysis of predicted and incomplete genes	247

## Chapter 6:

Fig. 6.1	Synteny between human and zebrafish	253
Fig. 6.2	BAC isolation by reduced stringency hybridisation	258
Fig. 6.3	Evaluation of false positives	260
Fig. 6.4	Evaluation of false negatives	262
Fig. 6.5	Identification of BAC clones using an STS designed to the zebrafish EST wz3779	266
Fig. 6.6	Summary of the gene map constructed in zebrafish	268
Fig. 6.7	Comparison of the genes identified in zebrafish with the genes in the region of interest between HPR6.6 and ZNF-Kaiso in human	270
Fig. 6.8	Analysis of orthologous in human, mouse and zebrafish (1)	271
Fig. 6.9	Analysis of orthologous in human, mouse and zebrafish (2)	273
Fig. 6.10	DOTTER of bZ74M9 against itself showing the presence of five copies of a direct repeat	276
Fig. 6.11	Comparison of genes in human, mouse and zebrafish	282
Fig. 6.12	Identification of conserved sequences	284
Fig. 6.13	Evidence of a novel conserved exon	287

## Chapter 7:

Fig. 7.1	A representation of how speciation and gene duplication can influence comparative genome analysis	313
----------	---	-----

**List of Tables:****Chapter 1**

Table 1.1	Complete DNA sequence	3
Table 1.2	Comparison of G-bands and R-bands	6
Table 1.3	Genes in the human genome	20

**Chapter 2**

Table 2.1	Clones and appropriate antibiotics	47
Table 2.2	cDNA libraries used	51
Table 2.3	Vector-specific primer sequences and ‘bubble’ sequences for primers used in vectorette PCR and SSPCR (performed on clone DNA and cDNA)	52
Table 2.4	STSs from Srivastava <i>et al</i> (1999) used for contig construction	53
Table 2.5	STSs derived from clone ends used for walking as described in chapter 3	54
Table 2.6	STSs used for gene identification as described in chapter 4	57
Table 2.7	STSs used for mouse mapping as described in chapter 5	61
Table 2.8	STSs for conserved sequence analysis in human and mouse	62
Table 2.9	STSs used for zebrafish mapping as described in chapter 6	62
Table 2.10	Primer combinations used in SSPCR	78

**Chapter 3**

Table 3.1	Position of the DXS101 loci in the genomic sequence	117
Table 3.2	Low copy duplications between DXS366 and DXS1230	118
Table 3.3	Example of probability of overlaps, comparing clones of different sizes	127

**Chapter 4**

Table 4.1	Known genes with confirmatory human cDNA sequence	136
Table 4.2	Experimental verification of predicted genes	142

Table 4.3	Evaluation of the gene structures	157
Table 4.4	Functional characterisation of Genes	173
Table 4.5	Link information as described in Figure 4.9	196
Table 4.6	Information of pseudogenes	197
Table 4.7	STSs used for mutation screening of MRX23 patients	198
Table 4.8	Information on cDNA sequencing projects	200
 <b>Chapter 5</b>		
Table 5.1	Comparison of orthologous genes	220
Table 5.2	Results of conserved sequence analysis	235
Table 5.3	Comparison of read number for various genome equivalents	242
Table 5.4	Information on sequence shown in Figure 8	249
 <b>Chapter 6</b>		
Table 6.1	Summary of bacterial clone isolation	256
Table 6.2	Breakdown of known repeats	277
Table 6.3	Summary of novel repeat sequences in the zebrafish genome	278
Table 6.4	Summary of conserved sequence analysis in human, mouse and zebrafish	285
Table 6.5	Comparison of orthologous genes in human, mouse and zebrafish	292



## Glossary of Abbreviations

ACeDB	<i>A. C. elegans</i> database
ANT2	adenosine nucleotide transporter 2
<i>Alu</i> -PCR	<i>Alu</i> -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
$\beta$ -ME	$\beta$ -mercaptoethanol
bp	base pair
BSA	bovine serum albumin
BTK	Bruton's tyrosine kinase
$^{\circ}$ 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 <i>in situ</i> 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 Merzbacher 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(hydroxymethyl)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.