

**Human chromosome 20q12-13.2: Structural, comparative and  
sequence variation studies**

**George Stavros Stavrides**

**A thesis submitted in partial fulfilment of the  
requirements of Cambridge University for the degree of  
Doctor of Philosophy**

**Wolfson College, University of Cambridge**

**September 2002**

**This dissertation is the result of my own work and includes nothing that is the outcome of work done in  
collaboration. The dissertation does not exceed the length limit set by the Biology Degree Committee.**

*To my parents*

## **Abstract**

As the human genome sequencing effort nears completion, there is a great need to identify and characterise the structural pieces of genetic information embedded in the generated sequence. The aim of this project was to explore how this goal can be achieved, in order to maximise the impact of the Human Genome Project on genome biology. A 10 Mb region on human chromosome 20q12-13.2 (representing 1/6th of the whole chromosome) provided the basis for a number of studies.

The assembly of a detailed transcript map across this region is described. Candidate gene features were identified from publicly available expressed sequences and *ab initio* gene predictions, then experimentally verified and extended. The final transcript map contains 99 coding genes, 30 putative genes and 36 pseudogenes. The expression of all novel genes was investigated by PCR screening of seven cDNA libraries. All annotated structures were studied in terms of total sequence coverage and their sequence environment was investigated. Splice sites, polyadenylation signals, isoforms and predicted transcription start sites/promoters/CpG islands are also discussed. The predicted encoded proteins were compared to various proteomes (including human), whereas data from three-species genomic sequence comparisons was used to confirm that virtually all exons and probably all genes in this region have been identified.

A gene/homology-based approach was used to construct a contiguous, 10 Mb long bacterial clone map on mouse chromosome 2 spanning the syntenic region on human 20q12-13.2. A tile path of 66 BAC clones was used to generate approximately 10.3 Mb of sequence. The mouse and human sequences were compared and the distribution of regions showing sequence conservation is discussed in the context of the annotated human sequence. The two syntenic

regions showed strong conservation of gene order and content, but no conservation of human putative genes and pseudogenes was observed within the mouse sequence. Non-exonic conserved sequences and *ab initio* predictions were used to estimate the completion of human annotation.

Human expressed sequences aligned to the annotated exons in 20q12-13.2 were used to identify over 100 exonic SNPs. A set of 2,208 SNPs mapping across the region was used to obtain allele frequencies in three populations (95 Caucasians, 12 Asians and 12 African Americans). A first generation linkage disequilibrium (LD) map of the region was constructed in Caucasians. Over half of the region is covered in “LD blocks”, segments with three or more SNPs and for which all possible SNP pairs have  $D' > 0.9$ .

## **Acknowledgements**

Many people have kindly provided me with their help and advice throughout the course of this project. The most important of these is Panos Deloukas, who has given me guidance and encouragement since the beginning of my time at the Sanger Institute. Thank you for all the time you have spent in helping me to complete this project.

A project of this type inevitably involves many people. I would like to thank Susan Rhodes, Jackie Bye and John Collins, for help with the gene discovery experiments. Thanks also to Ian Mullenger and Lisa French for their advice and guidance during the mouse map construction. A big thanks to Carol Scott, Sarah Hunt and Jilur Ghori, for being always able to help with data analysis and database management. Many thanks to Webb Miller and Francesca Chiaromonte for the help with the comparative studies and Robert Lawrence for the analysis of the genotype data. Many other people from the Sanger Institute have provided me with invaluable assistance and I would particularly like to thank Pam Whittaker, Kate Downes, Tom Dibling, Melanie Goward, Rhian Gwilliam, Elisabeth Huckle and Carol Carder. Thanks should also be extended to people in informatics, especially Jennifer Ashurst, James Gilbert, Thomas Down and Michele Clamp, as well as to all the sequencing team members for generating the human and mouse sequence.

I would like to also acknowledge the helpful discussions and critical reading of this manuscript by Panos Deloukas and Rhian Gwilliam. The transcript map fold out diagram was created by James Gilbert and printed by Richard Summers.

I would also like to thank the residents (past and present) of the G205 office, Ele Holloway, Luk Smink, Andy Mungall and Rhian Gwilliam for sharing an office with me, and putting up with me.

A warm thank you to my dearest friends, Vassilis Koudounas, Marios Pirishis, Thanasis Athanasiou, Stelios Koursaris, Chris Markides, Simoni Ioannou, George Petropoulos, Vangelis Andreakos, Alex Papadatos, Yiannis Iliopoulos and (of course) cousin George for all the good times, home and abroad, and their support.

Finally, I would like especially to thank my parents and sister, for their support and encouragement. Your love and belief in me kept me going. Naki, good luck with your final experiments and the forthcoming write-up!

## Table of contents

<b>Abstract</b>		<b>iii</b>
<b>Acknowledgements</b>		<b>v</b>
<b>Table of contents</b>		<b>vii</b>
<b>List of Tables</b>		<b>xii</b>
<b>List of Figures</b>		<b>xiii</b>
<b>List of Abbreviations</b>		<b>xv</b>
<b>Publications arising from this work</b>		<b>xviii</b>
<b>Chapter I</b>	<b>Introduction</b>	
<b>1.1</b>	<b>Introduction</b>	<b>2</b>
<b>1.2</b>	<b>Mapping the human genome</b>	<b>5</b>
1.2.1	The genome	5
1.2.2	Genome mapping	7
<b>1.3</b>	<b>Sequencing and the landscape of the human genome</b>	<b>11</b>
1.3.1	Construction of sequence maps	11
1.3.2	The genomic landscape-sequence features	13
<b>1.4</b>	<b>Computational genomics (Bioinformatics)</b>	<b>29</b>
1.4.1	Sequence databases	29
1.4.2	Sequence analysis	30
1.4.3	Viewing genomic information	33
<b>1.5</b>	<b>Comparative genomics</b>	<b>36</b>
<b>1.6</b>	<b>Functional genomics</b>	<b>42</b>
<b>1.7</b>	<b>Human variation</b>	<b>45</b>
1.7.1	SNP identification	46
1.7.2	SNP analysis	48
1.7.3	Utilizing SNP data	51
<b>1.8</b>	<b>Chromosome 20</b>	<b>54</b>
<b>1.9</b>	<b>This thesis</b>	<b>57</b>

<b>Chapter II</b>	<b>Materials and methods</b>	
<b>2.1</b>	<b>Gene identification</b>	<b>60</b>
2.1.1	DNA manipulation methods	60
2.1.2	Clone resources	65
2.1.3	Isolation of cDNA fragments	71
2.1.4	Northern Blots	77
<b>2.2</b>	<b>Mouse studies</b>	<b>79</b>
2.2.1	Probe preparation	79
2.2.2	Screening	81
2.2.3	Fingerprinting	83
<b>2.3</b>	<b>Human variation</b>	<b>86</b>
2.3.1	DNA samples	86
2.3.2	SNP selection and primer design	86
2.3.3	Working PCR primer mix and probe dilutions	88
2.3.4	PCR amplification	88
2.3.5	SAP	89
2.3.6	Extension of primer probe	90
2.3.7	Water and resin addition	90
2.3.8	Mass spectroscopy	91
<b>2.4</b>	<b>Bioinformatics and computational support</b>	<b>92</b>
<b>2.5</b>	<b>Materials</b>	<b>96</b>
<b>2.6</b>	<b>DNA ladders</b>	<b>100</b>
<b>2.7</b>	<b>Solutions</b>	<b>101</b>
<b>Chapter III</b>	<b>Sequence and transcript map of 20q12-13.2</b>	
<b>3.1</b>	<b>Introduction</b>	<b>107</b>
3.1.1	Strategies for gene identification	107
3.1.2	Overview	111
<b>3.2</b>	<b>Sanger annotation pipeline</b>	<b>113</b>
<b>3.3</b>	<b>Experimental confirmation of 20q12-13.2 genes</b>	<b>117</b>
3.3.1	Vectorette	117
3.3.2	Single specificity PCR	121

3.3.3	RACE	121
3.3.4	Summary of experimental efforts	123
<b>3.4</b>	<b>Combining all computational and experimental data – re-annotation of 20q12-13.2</b>	<b>124</b>
3.4.1	Broad genome landscape	127
3.4.2	Supporting evidence for annotated loci	130
3.4.3	First-pass expression data for novel and putative genes	132
3.4.4	Gene features	134
3.4.5	Splice sites	137
3.4.6	Splice isoforms	138
<b>3.5</b>	<b>Investigating the annotated 5' and 3' ends of coding genes</b>	<b>140</b>
3.5.1	Polyadenylation signals (3' end)	141
3.5.2	Promoters (5' end)	142
3.5.3	Summary	146
<b>3.6</b>	<b>Measuring completion of annotation</b>	<b>147</b>
3.6.1	Homology searches	147
3.6.2	Genscan and FGENESH	148
<b>3.7</b>	<b>Protein analysis</b>	<b>150</b>
<b>3.8</b>	<b>Discussion</b>	<b>153</b>
<b>Chapter IV</b>	<b>Comparative mapping, sequencing and analysis</b>	
<b>4.1</b>	<b>Introduction</b>	<b>161</b>
4.1.1	The mouse genome	161
4.1.2	Comparative studies	164
4.1.3	Overview	166
<b>4.2</b>	<b>Mouse clone map construction</b>	<b>168</b>
4.2.1	Marker selection and development	170
4.2.2	Bacterial clone identification	175
4.2.3	Fingerprint analysis	178
4.2.4	Landmark content mapping	181
4.2.5	BAC contig assembly in FPC	181
4.2.6	Gap closure	183

4.2.7	Genetic markers	185
<b>4.3</b>	<b>The sequence-ready bacterial clone map</b>	<b>186</b>
<b>4.4</b>	<b>Tile path selection and sequencing</b>	<b>187</b>
<b>4.5</b>	<b>Long range comparative sequence analysis</b>	<b>189</b>
4.5.1	Repeat content analysis	189
4.5.2	BLAST searches	191
4.5.3	An evaluation of the current human sequence annotation	196
<b>4.5.4</b>	<b>PipMaker analysis</b>	<b>197</b>
<b>4.6</b>	<b>Finished mouse sequence analysis</b>	<b>200</b>
<b>4.7</b>	<b>Discussion</b>	<b>213</b>
<b>Chapter V</b>	<b>Human variation</b>	
<b>5.1</b>	<b>Introduction</b>	<b>221</b>
5.1.1	Human variation	221
5.1.2	Theoretical aspects of linkage disequilibrium	222
5.1.3	Allelic associations and common disease	226
5.1.4	Mass spectrometry	227
5.1.5	This chapter	231
<b>5.2</b>	<b>Exonic SNP discovery in 20q12-13.2</b>	<b>232</b>
5.2.1	Identification of exonic SNPs in silico	232
5.2.2	Features of exonic SNPs	234
<b>5.3</b>	<b>Studying sequence variation across 20q12-13.2</b>	<b>237</b>
5.3.1	SNP selection and high-throughput genotyping	237
5.3.2	Error checks and quality assessment of data	240
5.3.3	Estimation of allele frequencies in three populations	241
<b>5.4</b>	<b>A first generation LD map of 20q12-13.2 in Caucasians</b>	<b>246</b>
<b>5.5</b>	<b>Discussion</b>	<b>261</b>
<b>Chapter VI</b>	<b>Discussion</b>	
<b>6.1</b>	<b>Summary</b>	<b>266</b>
<b>6.2</b>	<b>Analysis of genomic sequence</b>	<b>266</b>
<b>6.3</b>	<b>Mouse genomics</b>	<b>268</b>

<b>6.4</b>	<b>Human variation and linkage disequilibrium</b>	<b>269</b>
<b>6.5</b>	<b>Conclusions and future work</b>	<b>270</b>

<b>Chapter VII</b>	<b>References</b>	<b>273</b>
--------------------	-------------------	------------

## **Appendices**

<b>Appendix 1</b>	<b>A Genomics Timeline</b>	<b>I</b>
<b>Appendix 2</b>	<b>List of primers designed and used for gene identification</b>	<b>III</b>
<b>Appendix 3</b>	<b>cDNA probe repository</b>	<b>XII</b>
<b>Appendix 4</b>	<b>cDNA sequences</b>	<b>XVII</b>
<b>Appendix 5</b>	<b>Gene data</b>	<b>XXII</b>
<b>Appendix 6</b>	<b>Novel gene expression results and isolation of cDNA sequences</b>	<b>XXX</b>
<b>Appendix 7</b>	<b>Putative genes expression results</b>	<b>XXXIII</b>
<b>Appendix 8</b>	<b>Supporting evidence for annotated genes</b>	<b>XXXIV</b>
<b>Appendix 9</b>	<b>The sequences of gene-based, working STSs</b>	<b>XXXVIII</b>
<b>Appendix 10</b>	<b>Mouse BAC-end sequence-based STSs</b>	<b>XLI</b>
<b>Appendix 11</b>	<b>Mouse genetic markers mapped on mouse contig</b>	<b>XLVI</b>
<b>Appendix 12</b>	<b>Human clone sequences</b>	<b>XLVIII</b>
<b>Appendix 13</b>	<b>Mouse clone sequences</b>	<b>LI</b>
<b>Appendix 14</b>	<b>Exonic cSNPs</b>	<b>LIII</b>
<b>Appendix 15</b>	<b>Verified polymorphic SNPs</b>	<b>LVI</b>
<b>Appendix 16</b>	<b>Variability of D' and r<sup>2</sup></b>	<b>LXVI</b>

## List of tables

Table 1.1	Properties of Giemsa (G) and Reverse (R) bands	6
Table 1.2	Blast types	31
Table 1.3	Overview of the main sequence-feature prediction programs	33
Table 1.4	Genome sizes of the model organisms initially proposed	37
Table 1.5	Essential SNP facts	46
Table 2.1	PCR mixes and cycling programs	63
Table 2.2	cDNA resources	66
Table 2.3	Universal primer sequences	72
Table 2.4	Northern Blots	77
Table 2.5	Details of the mouse genomic library	79
Table 2.6	Touch-down PCR programs	80
Table 2.7	European, Asian and African American samples	87
Table 2.8	Software used in this study	92
Table 2.9	People involved in sequence analysis and data storage and management	93
Table 2.10	URLs used in this study	94
Table 3.1	Number, Coverage and Density of different classes of repeats	127
Table 3.2	Supporting evidence for annotated features	132
Table 3.3	Size of gene loci	134
Table 3.4	Structural features of annotated gene features	135
Table 3.5	Structural features of genes annotated in chromosomes 20, 21 and 22	136
Table 3.6	Polyadenylation signals found in the 3'UTR of annotated coding genes	141
Table 3.7	Correlation of predicted regions and annotation	143
Table 3.8	Correlation of types of annotated features and predictions	144
Table 3.9	Analysis of predicted coding sequence	148
Table 3.10	Comparison of Genscan and FGENSH predictions and annotated, supported, coding exons	149
Table 3.11	Most common InterPro domains in 20q12-13.2 and their abundance in other species	152
Table 4.1	Gene-based (working) STS markers	172
Table 4.2	Repeat content analysis	190
Table 4.3	Human:mouse BLAST searches	194
Table 4.4	PipMaker analysis	199
Table 4.5	Sequence features of three gene pairs	204
Table 4.6	Percentage identities of human and mouse sequences	209
Table 5.1	Expected distribution of transitions and transversions	234
Table 5.2	Distribution of transitions and transversions	234
Table 5.3	Coding changes for exonic SNPs	236

## List of figures

Figure 1.1	Basic gene structure	25
Figure 1.2	The Ensembl and UCSC genome browsers	34
Figure 1.3	'Modular' design of some of the assays for SNP genotyping	49
Figure 3.1	Sequence analysis pipeline	114
Figure 3.2	ACeDB view of sequence features	115
Figure 3.3	Vectorette library construction	118
Figure 3.4	Example of cDNA-end isolation using the vectorette method	119
Figure 3.5	Complementing annotation using RACE	122
Figure 3.6	The sequence map of human chromosome 20q12-13.2	126
Figure 3.7	Repeat content distribution of 20q12-13.2	128
Figure 3.8	Repeat distribution for each family	128
Figure 3.9	Dot plot of the two regions present in the sequences of clones RP5-1057D4 and RP5-991B18	129
Figure 3.10	Positive gene features for cDNA libraries tested	133
Figure 3.11	Positive cDNA libraries per gene feature	133
Figure 3.12	3' intron-5' exon splice sites	137
Figure 3.13	3' exon-5' intron splice sites	138
Figure 3.14	Northern blots	140
Figure 3.15	Correlation of genes with predicted promoters at their 5' end and predictions by the three methods.	146
Figure 3.16	SEMG1 and SEMG2 protein alignment	151
Figure 4.1	Strategy for contig construction, involving landmark content mapping and restriction enzyme fingerprinting	169
Figure 4.2	Design of mouse STSs	171
Figure 4.3	Primer testing	174
Figure 4.4	Overview of BAC identification strategy	176
Figure 4.5	Positive clone identification and scoring	177
Figure 4.6	Viewing and editing fingerprint data in Image	178
Figure 4.7	Example of landmark content mapping	182
Figure 4.8	Example of PCR-based library screen	184
Figure 4.9	Hybridising marker D2MIT413 (stSG104981) to BAC polygrid 2	185
Figure 4.10	The mouse clone map	188
Figure 4.11	Size distribution of mouse BLAST hits	191
Figure 4.12	Acedb view of human:mouse BLAST search results	193
Figure 4.13	Pip-plots for two genomic regions	198
Figure 4.14	Scatter plots for the exon sizes between human and mouse	201
Figure 4.15	Scatter plots for the intron sizes between human and mouse	202
Figure 4.16	Coding sequence alignments	205
Figure 4.17	Protein sequence alignments	207
Figure 4.18	CpG island comparison	209
Figure 4.19	GC- (A) and Repeat content analysis (B-D) of orthologous gene pairs	211

Figure 5.1	The erosion of linkage disequilibrium by recombination	224
Figure 5.2	MALDI-TOF MS	229
Figure 5.3	Blixem view of homologous sequences	232
Figure 5.4	Supporting evidence for exonic SNPs	233
Figure 5.5	Codon position changes for coding exonic SNPs	235
Figure 5.6	Viewing genotyping results in SpectroAnalyser	238
Figure 5.7	Overall breakdown of SNP assay results	242
Figure 5.8	Breakdown of “complete” SNP assay results	243
Figure 5.9	Distribution of polymorphic and monomorphic SNPs	244
Figure 5.10	Distribution of polymorphic SNPs in the three groups	245
Figure 5.11	Distribution of minor allele frequencies in ethnic populations	246
Figure 5.12	SNP distribution across the region of interest	247
Figure 5.13	Distance between neighbouring SNPs (SNP pairs)	248
Figure 5.14	Proportion of sequence occupied by the various types of SNP pairs	249
Figure 5.15	Variability of $D'$ and $r^2$	251
Figure 5.16	Average $D'$ and $r^2$	253
Figure 5.17	Linkage disequilibrium across 20q12-13.2	256
Figure 5.18	The Marshfield genetics maps	257
Figure 5.19	Correlation of “LD blocks” and SNPs	258
Figure 5.20	Size correlation of “LD blocks”	259
Figure 5.21	The distribution of “LD blocks” across 20q12-13.2	260

## List of abbreviations

20ace	chromosome 22 implementation of ACeDB
aa	amino acid
ACeDB	A C. elegans DataBase
AITDs	AutoImmune Thyroid Diseases
ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BSA	Bovine Serum Albumin
cDNA	complementary DNA
CDS	CoDing Sequence
CEPH	Centre d'Etude du Polymorphisme Humain
cM	centiMorgan
CpG	5'CG3' dinucleotide
cR	centiRay
cRSC	coding Region of Sequence Conservation
cSNP	complementary SNP
dATP	2'-deoxyAdenosine 5'-TriPhosphate
dbEST	database of ESTs
dbSNP	database of SNPs
dCTP	2'-deoxyCytidine 5'-TriPhosphate
DDBJ	Dna DataBase of Japan
ddCTP	2', 3'-dideoxyCytidine 5'-TriPhosphate
dGTP	2'-deoxyGuanosine 5'-TriPhosphate
DNA	DeoxyriboNucleic Acid
dsDNA	double strand DNA
DTT	DiThioThreitol
dTTP	2'-deoxyThymidine 5'-TriPhosphate
EDTA	EthyleneDiamineTetraAcetic acid
EMBL	European Molecular Biology Laboratory
ePCR	Electronic PCR
EST	Expressed Sequence Tag
FBS	Fetal Bovine Serum
FEN	Flap EndoNucleases
FISH	Fluorescent In Situ Hybridisation
FMF	Familial Mediterranean Fever
FPC	FingerPrinting Contigs
G-band	Giemsa band
GD	Graves disease
GSS	Genome Survey Sequence
HERV	Human Endogenous RetroVirus-like elements
HGP	Human Genome Project
HGSP	Human Genome Sequencing Project

HSA20	Homo Sapiens chromosome 20
HT	Hashimoto's Thyroiditis
iATG	Translation Initiation site
IHGMC	International Human Genome Mapping Consortium
IHGSC	International Human Genome Sequencing Consortium
INSD	International Nucleotide Sequence Databases
ISNPMWG	International SNP Map Working Group
Kb	Kilo base pairs
LINE	Long INterspersed repeat Element
LTR	Long Terminal Repeat
MaLR	Mammalian LTR
Mb	Mega base pairs
MDS	Myelodysplastic Syndromes
MER	Medium Reiterative Repeat
MGC	Mouse Genome Consortium
MGD	Mouse Genome Database
MGSC	Mouse Genome Sequencing Consortium
MIR	Mammalian-wide Interspersed Repeat
MIR	Mammalian-wide Interspersed Repeat
MMU2	Mus MUsculus chromosome 2
MPD	MyeloProliferative Disorders
mRNA	messenger RNA
MS	Mass Spectroscopy
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NIH	National Institute of Health
nt	nucleotide
OMIM	Online Mendelian Inheritance In Man
ORF	Open Reading Frame
PAC	P1 Artificial Chromosome
PCR	Polymerase Chain Reaction
PIP	Percentage Identity Plot
Q-banding	Quinacrine banding
R	Purine
R-banding	Reverse banding
RCS	Region of Sequence Conservation
RefSNP	Reference SNP
RFLP	Restriction Fragment Length Polymorphism
RH	Radiation Hybrid
RNA	RiboNucleic Acid
RNAi	RNA interference
rRNA	ribosomal RNA
RT-PCR	Reverse Transcription PCR
SAGE	Serial Analysis of Gene Expression
SDS	Sodium Dodecyl Sulphate
SINE	Short INterspersed repeat Element
SNP	Single Nucleotide Polymorphism

snRNA	small nuclear RNA
SRS	Sequence Retrieval System
SSR	Simple Sequence Repeat
STS	Sequence Tagged Site
TIR	Terminal Inverted Repeat
TrEMBL	Translated EMBL
tRNA	transfer RNA
TS site	Transcription Start site
TSC	The Snp Consortium
UTR	UnTranslated Region
VNTR	Variable Number Tandem Repeat
WGS	Whole Genome Shotgun
WWW	World Wide Web
Y	Pyrimidine
YAC	Yeast Artificial Chromosome

## Publications arising from this work

Bench A. J., Nacheva E. P., Hood T. L., Holden J. L., French L., Swanton S., Champion K. M., Li J., Whittaker P., *Stavrides G.*, Hunt A. R., Huntly B. J., Campbell L. J., Bentley D. R., Deloukas P., and Green A. R. (2000). Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG). *Oncogene* **19**: 3902-13.

Deloukas P., Matthews L. H., Ashurst J., Burton J., Gilbert J. G., Jones M., *Stavrides G.*, Almeida J. P., Babbage A. K., Bagguley C. L., Bailey J., Barlow K. F., Bates K. N., Beard L. M., Beare D. M., Beasley O. P., Bird C. P., Blakey S. E., Bridgeman A. M., Brown A. J., Buck D., Burrill W., Butler A. P., Carder C., Carter N. P., Chapman J. C., Clamp M., Clark G., Clark L. N., Clark S. Y., Clee C. M., Clegg S., Cobley V. E., Collier R. E., Connor R., Corby N. R., Coulson A., Coville G. J., Deadman R., Dhami P., Dunn M., Ellington A. G., Frankland J. A., Fraser A., French L., Garner P., Grafham D. V., Griffiths C., Griffiths M. N., Gwilliam R., Hall R. E., Hammond S., Harley J. L., Heath P. D., Ho S., Holden J. L., Howden P. J., Huckle E., Hunt A. R., Hunt S. E., Jekosch K., Johnson C. M., Johnson D., Kay M. P., Kimberley A. M., King A., Knights A., Laird G. K., Lawlor S., Lehvaslaiho M. H., Lerversha M., Lloyd C., Lloyd D. M., Lovell J. D., Marsh V. L., Martin S. L., McConnachie L. J., McLay K., McMurray A. A., Milne S., Mistry D., Moore M. J., Mullikin J. C., Nickerson T., Oliver K., Parker A., Patel R., Pearce T. A., Peck A. I., Phillimore B. J., Prathalingam S. R., Plumb R. W., Ramsay H., Rice C. M., Ross M. T., Scott C. E., Sehra H. K., Shownkeen R., Sims S., Skuce C. D., et al. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**: 865-71.