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

List of tables

List of figures

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., Leversha 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.