Genome Evolution:
a study of MHC paralogous genes in the human genome

Vikki Rand

This dissertation is submitted for the degree of Doctor of Philosophy

September 2003

Gonville and Caius College,
University of Cambridge

Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus,
Hinxton, Cambridge

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 word limit set by the Biology Degree Committee.

Abstract

One of the interesting findings of the Human Genome Project was that approximately 10% of the genome has arisen by duplication. This is exemplified by the clusters of genes, on chromosomes 1q21-q25, 9q32-q34.3 and 19p13, paralogous to genes located within the Major Histocompatibility Complex (MHC) region, on 6p22.2-p21.3. By definition, paralogues are genes within the same species that have originated through duplication of an ancestral gene. The survey of the human genome identified 82 MHC paralogues based on sequence similarity and conserved gene structure. Analysis of the distribution of the paralogues identified clusters on chromosomes 1q21-q25, 9q32-q34.3 and 19p13 (38/82), and revealed paralogues located elsewhere in the genome (44/82). In total, 44% of the paralogues identified are novel discoveries, of which 89% are located outside the previously known clusters.

Evidence from my phylogenetic analyses indicates that the MHC paralogues located within the regions on 1, 9 and 19 arose by two ancient duplication events, either by duplication of the whole genome or of chromosomal segments, prior to vertebrate emergence. Expansion of paralogous gene families has occurred by additional duplications involving individual loci or chromosomal regions resulting in paralogues outside the clusters. In-depth analysis of the chromosomal region 9q32-q34.3 revealed that the order of paralogues is not conserved and that they are interspersed by other genes, indicating the region has been subjected to genomic rearrangements.

Comparison of the expression profiles of a selected set of MHC paralogues revealed that some have functionally diverged since duplication; with members of the same paralogous gene family being ubiquitously expressed, and others, having an

expression profile restricted to only a few tissues. Evidence of co-expression of paralogues in some tissues suggests a similar function and involvement in the same pathways. This thesis highlights the importance of understanding paralogy, particularly for future investigations of phenotypes associated with paralogous genes.

Acknowledgements

I would like to thank my supervisor Stephan Beck for his tireless enthusiasm, encouragement and support throughout this project. Particular thanks go to my second supervisor at Cambridge University, John Trowsdale. I am also grateful to the past and present members of the Immunogenomics group (a.k.a Team 50) who have helped me throughout my PhD: Karen Novik, Ruth Younger, Roger Horton, Melanie Stammers, Karen Crum, Karen Halls, Jennifer Sambrook, Penny Coggill, Marcos Miretti and Vardhman Rakyan. Thankyou for all your advice, encouragement and help. I am also very grateful for your tolerance – particularly of my moaning over the last few months – and for providing regular distractions in the form of tea breaks, coffee breaks, lunch, pub..... this also goes for the members of team 30, Gavin Wright, Mark Bushall and Nick Bockett, who foolishly share our lab and office!

I am indebted to everyone who has contributed to the mapping and sequencing of the human genome - without this data the work in this thesis would not have been possible. I would also like to thank everyone who has helped me with the different aspects of the project. Special thanks goes to Sean Humphray and the members of the Chromosome 9 Sequencing and Mapping group who helped me with the work presented in chapter 3. Particular thanks go to Andrew Knight for his help with the sub-cloning, Keith Porter for the fingerprinting and the Sanger Institute Cytogenetics laboratory, namely Sheila Clegg and Pawan Dhami, who helped me with the fluorescent *in-situ* hybridisation (FISH) experiments. I am also very grateful to the members of team 41 and, the now extinct, team 49 who helped me with the sequencing and to Gavin Laird and Adrienne Hunt for their 'finishing' tutorials.

Thanks also go to Rosemary Ward for all her assistance with the tissue culture, Alison Coffey for her help with the blots and to Dave Vetrie, Cordelia Langford and the Sanger Institute Microarray Facility for all their help and guidance with the microarray experiments. Thanks also go to Ewan Birney and Michele Clamp for their help with ENSEMBL and to Kevin Howe for maintaining the FINEX database and writing some useful perl scripts. I realise I have been privileged to have worked in an environment where help has been offered so freely.

On a personal note, thanks must go to all my family and friends who have put up with me over the last few months – and years! Particular thanks go to my mam, dad and brother Simon for their pillar-like support and their unwavering belief and confidence in me – without this I don't think I would have made it this far! This also goes for all my friends who have offered encouragement and wine along the way!!! Cheers!

Table of contents

List of tables

List of figures

Chapter 6 Expression analysis of extended MHC paralogous gene families

Abbreviations

aa	amino acid
AIF	Allograft inflammatory factor
ATP	adenosine 5'-triphosphate
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair
BRD	Bromodomain containing protein
°C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
CLIC	Chloride intracellular chloride channel
CTP	cytidine 5'-triphophate
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphophate
DTT	dithiothreitol
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FISH	Fluorescent *in-situ* hybridisation
FPC	fingerprinting contig
GPX	Glutathione peroxidase
GTP	guanine 5'-triphosphate
HGMP	Human Genome Mapping Resource Centre
HGP	Human Genome Project
HLA	human leukocyte antigen
IHGSC	International Human Genome Sequencing Consortium
kb	kilobase pairs
l	litre
-L	-like
LB	Luria-Bertani
LINE	long interspersed nuclear element

M	molar
mA	milliamps
Mb	megabase pairs
µg	microgram
µl	microlitre
µM	micromolar
min(s)	minute(s)
MIPS	Munich Information Centre for Protein Sequences
mg	milligram
MHC	Major Histocompatibility Complex
ml	millilitre
mm	millimetre
mM	millimolar
NCBI	National Centre for Biotechnology Information
ng	nanogram
NOTCH	Neurogenic locus Notch homologue
OR	Olfactory receptor
PCR	polymerase chain reaction
PFAM	protein family database
PBX	Pre-B cell leukaemia transcription factor
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
rpm	revolutions per minute
RT-PCR	reverse transcription polymerase chain reaction
RXR	Retinoic acid receptor
SDS	sodium dodecyl sulphate
sec(s)	second(s)
SINE	short interspersed nuclear element
STS	sequence tagged site
TEMED	N, N, N', N'-tetramethylethylenediamine
Tris	tris(hydroxymethyl)aminomethane
U	unit
UTR	untranslated region
V	volt