

**Studies on polymorphic Alu insertions and genomic diversity
within the Major Histocompatibility Complex**

David Suliman Dunn BSc (Hons), MSc.

**This thesis is presented for the degree of Doctor of Philosophy of Murdoch
University, 2005.**

Declaration

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

.....

David Suliman Dunn

Dedication

In loving memory of my late mother (MARIAM), the light of my academic achievements, and to my late father (Dixon Suliman) .

Acknowledgments

The work described in this thesis has only been possible through the continued support of my colleagues and collaborators to whom I am extremely grateful.

I am indebted to my supervisors Professor Jerzy (Yurek) Kulski and Professor Matthew Bellgard for their intellectual support, guidance and encouragement and for the opportunity to pursue the work described in this thesis.

I am also grateful for the assistance from the many collaborators; Professor Roger Dawkins, Dr Silvana Gaudieri, Dr Guan Tay, Professor Chanvit Leelayuwat, Mr Joseph Williamson and Dr Sonia Cattely for the initial studies on MICA, Professor Frank Christiansen and Dr Campbell Witt for their willingness to share DNA samples and HLA data, Dr Shiina, Professor Hiditoshi Inoko, Dr Naruse, Dr M Ota, Dr Munkhbat, Dr Maude Phipps, Dr Brian Tait and Dr Himladevi Soodyall for supplying DNA and related data.

My sincere gratitude to my colleagues in the various research groups: Jennie Hui, Steve Iaschi, Nikitas Economou, David Sayer, David Scibechi, Adam Hunter, William Kenworthy, Paula Moolhauzjen and Kim Carter for their friendship and assistance.

This work was undertaken while holding an Australian Postgraduate Award Scholarship (January 2001 to December 2003) and a supplement research award from the Centre for Biological Computing and Bioinformatics (January 2001 to September 2003). The laboratory work was carried out in the laboratory of the State Agricultural Biological Centre (SABC) at Murdoch University.

Abstract

After the initiation of the human genome sequencing project and the introduction of the field of ‘bioinformatics’, interest in human genetic diversity studies has been increased. Sequence diversity has helped define differences between genes and genomic regions that were previously unknown or difficult to determine. In this thesis I have undertaken to study sequence diversity in the human genome in three areas; 1) investigated diversity in the MHC as represented by the MICA alleles with respect to the known HLA alleles, 2) investigated the structure and diversity in the intergenic region from an MHC related (paralogous) genomic region and related the structural and diversity findings to the knowledge available on the MHC and the wider genome, and 3) described the identification of three and characterization of five new MHC class I polymorphic markers (Alu) and their polymorphic characteristics in worldwide populations and their associations with skin cancer.

1. Phylogenetic analysis of MICA alpha-domain (extracellular) sequences demonstrated relationships with HLA-B cross-reactive serogroups. The HLA-B and MICA loci are in linkage disequilibrium. The data indicated that MICA and HLA-B have evolved in concert from their common ancestors and that the transmembrane polymorphisms have arisen independently and more recently.
2. Sequence analysis of the CD1 genomic region confirmed the presence of five CD1 genes and revealed that there are four unrelated intergenic regions (IGRs). The IGRs are composed mostly of retroelements including five full-length L1 PA sequences and various pseudogenes. Genomic and phylogenetic analyses support the view that the human CD1 gene copies were duplicated prior to the evolution of primates and the bulk of the HLA class I genes found in humans.
3. Five polymorphic Alu insertions (POALINs) were identified (two from previous studies) and located within the 1.8 megabase of the MHC class I genomic region. All five POALINs are polymorphic, and are positively associated with the HLA-A and HLA-B alleles. The Alu_{HJ} insertion was found most frequently associated with *HLA-A1* or *A24*, Alu_{HG} with *HLA-A2*, Alu_{HF} with *HLA-A2*, *A-10* or *-A26* and Alu_{TF} showed a marginal association with *HLA-A29*. The Alu_{MICB} insertion was strongly associated with *HLA-B17* (*HLA-B57*, *HLA-B58*) and *HLA-B13*. The presence of three Alu insertions

(AluyHJ, AluyHG and AluyHF) was found in only one HLA class I haplotype (*HLA-A1, -B57, -Cw6*) in the 10th IHW cell lines. A novel positive association between the presence of AluyMICB and the '*MICAdel/MICBnull/HLA-B48*' haplotype was determined. The AluyMICB insertion was also associated with at least three different MICB alleles (**0102, *0107N* and **0105*) and three different HLA-B alleles (*B13, B48* and *B57*). Based on the analysis of associations between different polymorphic markers within the beta block, the *MICB*0102* allele was inferred to be the ancestral form of the *MICB*0105* and *MICB*0107N* alleles. The AluyMICB polymorphism can be used to further investigate haplotype relationship and consequently their lineage origins. Some of the MHC POALINs are haplospecific and associate strongly with certain groups of HLA class I alleles and MHC ancestral haplotypes. The AluyTF frequency was significantly associated with skin cancer ($p < 0.005$).

MICA gene diversity is derived from two different evolving paths, therefore one or the other alone cannot reliably mark an ancestral haplotype. The CD1 duplicons originated well before the HLA class I duplicons. The MHC POALINs provide new lineage and linkage markers for the fine mapping study of different haplotypes and variations in linkage groups across 1.8 Mb of the MHC class I region. The POALINs may also prove useful in investigating the origins and history of human populations and in determining the role of human genetic diversity in disease risk.

TABLE OF CONTENTS

Author's declaration	i
Dedication	ii
Acknowledgments	iii
Abstract	iv
Table of Contents	vi
Abbreviations and symbols	ix
Glossary	xi
CHAPTER 1 INTRODUCTION	1
1.1 OVERVIEW	1
1.2 THESIS STRUCTURE	2
1.3 AIMS	2
CHAPTER 2 LITERATURE REVIEW	5
2.1 BRIEF HISTORY OF GENETIC DIVERSITY	5
2.1.1 Darwin and beyond	5
2.1.2 Genetic Diversity	5
2.1.3 Bioinformatics	7
2.2 HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX (MHC) - 6P21.3	7
2.2.1 Role of the MHC in Disease	9
2.2.2 Genomics of the MHC	10
2.2.2 The Human Chromosome Sequencing Project	13
2.2.3 MHC Paralogy	13
2.2.5 Diversity and population studies	15
2.2.6 MHC and MHC-related genes	15
2.2.6.1 Human Leucocyte Antigens (HLA)	15
2.2.6.2 "MHC Class I Related Chain" (MIC)	16
2.2.6.3 CD1 genomic region on chromosome 1	18
2.3 MHC CLASS I GENES AND GENOMICS	20
2.4 HUMAN RETROELEMENTS AND REPEAT SEQUENCES	21
2.4.1 Definition	21
2.4.3 Short Interspersed Nuclear Element (SINE)	22
2.4.3.1 Subtypes of Alu elements	22
2.4.3.2 Polymorphic Alu Insertion (POALIN)	23
2.4.4 SVA elements	23
2.5 MHC EXTENDED HAPLOTYPES	25
2.5.2 Ancestral Haplotypes	26
2.5.3 Disease Association	26
2.6 PURPOSE, DESCRIPTION AND PUBLICATIONS	27
CHAPTER 3 MATERIALS AND METHODS	31
3.1 BIOINFORMATICS MATERIALS (TOOLS) AND METHODS	31

3.1.1	MICA gene diversity	31
3.1.2	Intergenic diversity within the CD1 genomic cluster	33
3.1.3	POALINs	35
3.1.3.1	Location of MHC class 1 POALINS	35
3.1.3.2	Sub-family identification	35
3.2	BIOLOGICAL MATERIALS AND METHODS	36
3.2.1	DNA samples	36
3.2.1.1	Human cell-lines	36
3.2.1.2	Australian Caucasian samples	37
3.2.1.3	Population variation within Japan, Thailand and Malaysia	37
3.2.1.4	Sub-Saharan Africa Populations	37
3.2.1.5	HLA-B48 and HLA-B57/HLA-B7	38
3.2.1.6	Busselton population	38
3.2.2	PCR strategy	38
3.2.3	PCR analysis for POALINS	41
3.2.4	Sequencing of POALINS	42
3.2.5	Sequencing reactions	43
3.2.6	Purification of extension products for sequencing	43
3.3	POPULATION DATA ANALYSIS	43
3.3.1	POALIN Allele Frequency calculations	43
3.3.2	POALIN haplotype construction and calculations	45
3.3.3	Genetic distance and phylogeny	46
3.3.4	Linkage Disequilibrium Analysis	46
3.3.5	Differences Between Populations	48
CHAPTER 4 RESULTS AND DISCUSSION		49
4.1	COEVOLUTION OF HLA-B AND MICA (PERB11.1)	50
4.1.1	Coevolution of HLA-B and PERB11.1 (MICA): significance of independent triplet expansion within the transmembrane region of PERB11.1 (MICA).	53
4.2	INTERGENIC DIVERSITY	60
4.2.1	Genomic and phylogenetic analysis of the human CD1 and HLA class I multicopy genes.	63
4.3	MHC CLASS I POALIN DESCRIPTION and CHARACTERISATION	72
4.3.1	Alu polymorphism within the MICB gene and association with HLA-B alleles.	75
4.3.2	The association between HLA-A alleles and young Alu dimorphisms near the HLA-J, -H and -F gene loci in workshop cell lines and Japanese and Australian populations.	80
4.3.3	Characterisation of a dimorphic Alu element located between the TFIID and CDSN genes within the MHC by association studies using workshop cell lines and Japanese and Australian populations.	89
4.4	POPULATION GENETICS of the MHC CLASS I POALINs	98
4.4.1	Diversity of the polymorphic Alu insertions and their associations with MHC class I alleles and haplotypes in the Northeastern Thais.	101
4.4.2	The distribution of dimorphic Alu insertions within the MHC class I region and associations with HLA alleles in a Chinese population from Malaysia.	110

4.4.3	The distribution of POALINs within the MHC class I region and associations with HLA alleles in Mongolians.	127
4.4.4	Genetic variation of five MHC class I POALINs in Southern African human populations.	143
4.5	MHC CLASS I POALIN HAPLOTYPE ANALYSIS	158
4.5.1	Association of MHC dimorphic Alu insertions with HLA class I and MIC genes in Japanese HLA-B48 haplotypes.	161
4.5.2	The distribution of POALINs within the MHC class I HLA-B7 and HLA-B57 haplotypes.	165
4.6	MHC CLASS I POALIN and SKIN CANCER	169
4.6.1	The association between young dimorphic Alu elements within the Major Histocompatibility Complex (MHC) in Skin Cancer in an Australian population	171
4.7	REVIEW OF MHC Class I POALINS	184
4.7.1	Polymorphic Alu insertions within the Major Histocompatibility Complex class I genomic region. A brief review.	186
CHAPTER 5 DISCUSSION		214
CHAPTER 6 REFERENCES		220

ABBREVIATIONS

4AOH	Fourth Asia Oceania Histocompatibility Workshop
10thIHW	Tenth International Hitocompatibility Workshop
AH	Ancestral Haplotype
Alu	<i>Arthrobacter luteus</i> restriction enzyme (<i>AluI</i>) cleaved conserved insertion sequence
BLAST	Basic local alignment tool
BLUR	Bam-linked ubiquitous repeat
bp	base pair
CD1	Cluster domain 1
CEH	Conserved extended haplotype
CEPH	Centre d'Etude du Polymorphisme Humain
CREG	Cross-reactive group
DDR	Discoidin domain receptor
DNA	Deoxyribonucleic acid
EBV	Epstein-Barr virus
FcRn	Neonatal Fc receptor molecule
GvHD	Graft-versus-host disease
HERV	Human endogenous retrovirus
HSP	Heat shock protein
HLA	Human leucocyte antigen
H-W	Hardy-Weinberg equilibrium
IGRS	Intervening genomic region
Indel	Insertion/deletion
IVS	intervening sequence (region)
Kb	Kilobase
LD	Linkage disequilibrium
LINE	Long interspersed element
LTR	Long terminal repeat
MEGA	Molecular evolutionary genome analysis
MER	Medium reiterated frequency repeat
MHC	Major histocompatibility complex

MIC	MHC class I chain related
MIR	Medium interspersed repeat
MLR	Mixed lymphocyte reaction
mya	Million years ago
PCR	Polymerase chain reaction
POALIN	Polymorphic Alu insertion
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SBT	Sequence-based typing
SINE	Short interspersed elements
SSCP	Single strand conformational polymorphism
SSOP	Sequence-specific oligonucleotide probing
SSP	Sequence-specific priming
SVA	composite retrotransposon of SINE, VNTR and Alu
TE	Transposable element
TNF	Tumor necrosis factor
tRNA	Transfer ribonucleic acid
VNTR	Variable number tandem repeat

Glossary

Allo-antisera	Different human or different animal sera containing antibodies that are specific for one or more antigens.
Diversity	The presence of a wide range of variation in the qualities or attributes under discussion.
Haplotype	Combination of alleles at several linked loci on the same chromosome which are inherited <i>en block</i> from a parent.
Isozyme	A different electrophoretic form of the same multi-subunit enzyme. Unlike allozymes, isozymes are due to differing subunit configurations rather than allelic differences.
Panmixis	Random mating within a breeding population.
Paralogous	Two genes or clusters of genes at different chromosomal locations in the same organism that have structural similarities indicating that they derived from a common ancestral gene and has since diverged from the parent copy by mutation and selection or drift.
Paralogy	Relationship between two genes related through gene duplication.
Retroelement	A genetic element that transposes via an RNA intermediate.
Submetacentric	Having the centromere near the center of a chromosome but not in the middle, so that one arm is shorter than the other.
Transposable element	A segment of genetic material that is capable of changing its location in the genome