School of Biomedical Science

The molecular genetics of human complement C4: implications for mapping MHC disease susceptibility genes

Mareike Puschendorf

This thesis is presented for the Degree of
Master of Science
of
Curtin University of Technology

June 2003

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

Acknowledgements

I would firstly like to thank my supervisors David Sayer and Professor John Wetherall for their support and guidance throughout the year. David, thank you for being a fantastic supervisor and an endless source of ideas. Your great enthusiasm and continued interest have been truly encouraging. Thanks to John for arising my interest in the complex field of C4 genetics and for the constructive criticism and assistance in putting together this thesis.

I am grateful to Dr Cyril Mamotte for his encouraging suggestions and his invaluable advice in designing the dHPLC experiments.

Many thanks to Silvana Gaudieri for the great help with the phylogenetic trees and for sharing her views on MHC structure and evolution.

I wish to thank Alison Castley for her advice in designing the SNP assays, Linda Smith, Beth Downsborough and Carmen Sent for their assistance in sequencing questions, Leanne Langan for helping me solve dHPLC problems and Catena Causerano for caring so much about the cell lines.

I would also like to thank all the other fantastic people at the Department of Clinical Immunology (Laila Gizzarelli, Mirna Stoklasa, Anja Prezelj, Julie Schunck, Filipa Carvalho, Damian Goodridge, Rebecca Whidborne, Benjamen Rushton, Filomena Trimboli, Annette Patterson, Dr Elizabeth Freitas, Dianne Da Santis, Blanka Cavlovic, Denise Whittle...), who assisted me with technical advice at various stages during this project and who made working in the lab an enjoyable learning experience. Special thanks go to Mirna, Anja, Leanne, Filipa and Jo for many joyful moments spent together in the lab and for those morning tea breaks.

Many thanks to Lydia Windsor for sharing the work with the SNP typing. Your help was much appreciated. The best of luck with your project! And thanks to Karey, Coral-Ann, Alvin, Silvia, Dom, Svetlana, Irma, Bree..., the Honours students, PhD students and researchers in the Department.

Special thanks to Peter for endless computer problem solving and for being so patient in teaching me how to use LaTeX, but most of all thanks for all those long discussions.

I would also like to thank all the fantastic people that I have met during my time here in Australia, and who made this year a lifetime experience. It's always hard to have to 'rebuild' a life in a place where no one really knows you. But it is a great challenge! And you have made the challenge look easy! Special thanks to Simon and Tanya for many fun

evenings, but most of all for being true friends. Thanks to Nick, Falko and Martin for the great weekends down at Molloy, to Alex F., Alex C., Alex S., Gareth, Emmy, Chris and again Simon and Tanya for all the fun trips, and to Ngaire for sharing the problems with the weirdest person I've ever met and for being a great friend.

Many thanks to my friends in Germany for spending endless hours on the phone with me and keeping me up to date with what's going on at home.

Finally, my warmest thanks go to my family for their continued support and encouragement when it was most required. I am especially grateful to my mum for supporting me in every way possible and for always being there for me. As my family is still struggling with their English, these last lines shall be repeated in German:
Abschließend noch ein riesiges Dankeschön an meine Familie. Ohne eure grenzenlose Unterstützung wäre diese Arbeit nie zustande gekommen. Ganz besonderer Danke geht an meine Ma. Danke, dass du immer für mich da warst!

Contents

List of Tables

List of Figures

Abbreviations

AH	Ancestral haplotype
BDT	Big dye terminator
Bf	Factor B
bp	base pair
C4	Complement component C4
C4BP	C4 binding protein
CAH	Congenital adrenal hyperplasia
Ch	Chido antigen
CR1	Complement receptor 1
DAF	Decay-accelerating factor
dHPLC	Denaturing high performance liquid chromatography
DNA	Deoxyribonucleic acid
Fc	Crystallisable fragment of immunoglobulins
HAART	Highly active antiretroviral therapy
HERV	Human endogenous retrovirus
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HSP	Heat shock protein
IDDM	Insulin-dependent diabetes mellitus
IFN	Interferon
IgG	Immunoglobulin G
IL	Interleukin
indel	Insertion/deletion
JRA	Juvenile rheumatoid arthritis

kb	Kilo base
kDa	Kilo Dalton
LTR	Long terminal repeat
MAC	Membrane-attack complex
MASP	MBL-associated serine protease
MBL	Mannan-binding lectin
MHC	Major histocompatibility complex
mRNA	Messenger ribonucleic acid
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleoside reverse transcriptase inhibitor
PCR	Polymerase chain reaction
PI	Protease inhibitor
RA	Rheumatoid arthritis
RCCX	Genetic module formed by RP, C4, CYP21 and TNX
Rg	Rodgers antigen
RNA	Ribonucleic acid
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphism
SSP	Sequence-specific primer
T_m	Melting temperature
TAP	Transporter associated with antigen processing
TEAA	Triethyl ammonium acetate
TNF	Tumor necrosis factor
TNX	Tenascin-X
UTR	Untranslated region

1 Abstract

The Major Histocompatibility Complex (MHC) is a gene-dense region located on the short arm of chromosome 6 (6p21.31). This region contains the highly polymorphic HLA genes as well as many other genes with immunological and non-immunological function. The susceptibility genes of many human disorders have been mapped to genes within the MHC. However, the genes themselves and indeed the locations of the genes, for many of the disorders, remain a mystery. This is a result of the high degree of linkage disequilibrium (LD) that exists between loci within the MHC. The high LD is explained by the genomic structure of the MHC. The MHC contains several blocks of DNA within which recombination is extremely rare, whereas the boundaries of the blocks are defined as "hotspots" of recombination. Most disease association studies have used the highly polymorphic HLA class I and class II genes which are separated by an, as yet, undefined number of blocks and several hundred kilobases of DNA.

The MHC gamma block resides in the central region of the MHC between the blocks that contain the HLA class I and class II genes. As such, typing for polymorphisms in the gamma block is critical for MHC disease gene mapping studies. The gamma block contains approximately 20 known genes including the complement C4 genes. The gamma block can contain between 1 and 3 tandemly arranged C4 genes. The C4 protein exists as either the C4A or C4B isotype and is polymorphic with up to 40 allotypes being reported. However, the majority of Caucasian haplotypes can be explained by the common C4A3 / C4B1 or C4AQ0 / C4B1 complotypes with the remaining haplotypes explained by just a few other complotypes. For this reason, and because C4 allotyping is a technically difficult procedure, C4 allotyping is rarely used in MHC disease association studies.

The molecular heterogeneity of human C4 genes has not been extensively studied. However, the C4A3 and C4B1 genes have been completely sequenced and are >99% identical at the DNA level across 41 exons and 15 kb of DNA. This high degree of homology and the presence of up to 3 C4 genes on any MHC haplotype makes PCR separation of the C4 genes difficult for subsequent genetic studies.

The aim of this study was to extensively characterise the molecular heterogeneity of the human C4 genes and thereby:

1. determine the extent of human C4 gene polymorphism
2. confirm previous studies which have defined isotype specific sequences
3. characterise the C4 protein polymorphisms at the DNA level

4. determine if common C4 allotypes can be subtyped on a molecular basis
5. identify C4 gene polymorphisms that can be used as targets for DNA based typing methods
6. apply DNA based C4 typing methods in MHC disease association studies
7. provide insights into MHC haplotype evolution.

In contrast to separating the C4 genes, a novel approach whereby the C4 genes were amplified and sequenced simultaneously was applied in this study. The DNA from 24 homozygous workshop cell lines, representing different ancestral haplotypes (AHs), was studied. Comparison of the C4 genes from different AHs revealed that the C4d region of the C4 α-chain is most polymorphic, but that polymorphic amino acid residues are also present in other regions of C4. The highest degree of polymorphisms was seen in the introns. In addition, the presence of the isotype specific sequences in exon 26 was confirmed and primers were designed to specifically amplify, and thereby separate, the C4A and C4B genes. Comparison of the C4 gene sequences representing the same C4 allotype revealed that most C4 allotypes are heterogeneous and may be split into several subtypes. The polymorphisms observed at the sequence level did not correlate with C4 allotypes defined by electrophoretic mobility. However, it could be shown that the differences in electrophoretic mobility of the C4 allotypes are due to cumulative charge differences. Seven polymorphic amino acids were found to account for the different migration rates of the C4 allotypes analysed in this study.

In addition, a number of haplospecific single nucleotide polymorphisms (SNPs) were identified within the C4 genes. Haplospecific SNPs are informative markers enabling the genetic mapping of recombinant AHs, an approach which can be used to identify disease susceptibility genes. Haplospecific SNPs located in the C4 gene region are important markers as they represent a separate block of the MHC (i.e. the gamma block). The frequency of one such SNP marker has been shown for a diabetes patient group and a control population. Although further studies are required to elucidate the role of the gamma block genes in susceptibility to diabetes, this study demonstrates a possible approach for the mapping of MHC disease susceptibility genes, which can also be applied in studies of other MHC associated diseases.

To conclude, the present study adds to our knowledge of the C4 gene polymorphism, provides insights into MHC and C4 gene evolution and enables future studies to examine the significance of the C4 genes and other gamma block genes in susceptibility to MHC associated diseases.

2 Literature Review

2.1 Major histocompatibility complex

The major histocompatibility complex (MHC) contains a set of highly polymorphic genes with immunological and non-immunological functions. The human MHC is located on the short arm of chromosome 6 and spans a region of about 4 Mb [1–3]. With over 200 gene loci identified, the MHC is the most gene-dense region of the human genome. The MHC has been extensively studied, yet many of the expressed gene loci are still of unknown function. The first complete sequence and gene map of the human MHC has been published by the MHC sequencing consortium in 1999 (Figure 1) [1].

Historically, the MHC has been divided into three functional regions: class II (centromeric), class III (central region) and class I (telomeric) [4]. However, further analysis revealed an extended class II region and a region telomeric to the classical MHC, which is now called class Ib region. An estimated 40% of the expressed MHC genes encode proteins with essential function in immune response. The classical class II region is particularly notable as almost all of the genes have a role in immune function.

One of the main characteristics of the MHC is its extreme polymorphism. The class I and class II HLA (human leukocyte antigen) molecules are the most polymorphic human proteins, some of which have over two hundred allelic variants [5].

2.1.1 MHC class I and II genes

The MHC class I genes encode the classical HLA class I molecules HLA-A, -B and -C. The MHC class I region also contains the non-classical HLA genes HLA-E, -F and -G as well as the pseudogenes HLA-H, -J, -K and -L. HLA-A, -B and -C molecules are cell surface glycoproteins expressed on almost all nucleated mammalian cells [6]. In contrast, the expression of the non-classical HLA genes is more restricted. For example, high levels of HLA-G are found on foetal cytotrophoblast cells during pregnancy and HLA-F is mainly expressed on B cells [7]. HLA-E is found in a wide range of fetal and adult tissues [8]. MHC class I molecules present foreign and self-antigens for recognition by T lymphocytes. Endogenous proteins, e.g. viral particles or normal cellular proteins, are targeted for degradation by binding of ubiquitin, and are subsequently digested by a large proteolytic complex termed proteasome. The resulting antigenic peptides are translocated to the endoplasmatic reticulum, which is the site of MHC biosynthesis. Class I molecules are composed of one heavy chain encoded by the polymorphic HLA-A, -B and -C genes and a non-covalently associated non-polymorphic light chain, called β_2 microglobulin [9]. The α_1 and α_2 domains of the heavy chain contain the amino acids determining the specificity of the class I molecules. Class I molecules can bind peptides of 8 to 11 amino acids [10]. Upon peptide binding, the MHC-class-I-peptide complexes are transported to the cell surface for recognition by CD8 T lymphocytes.

The classical HLA antigens encoded in the class II region are HLA-DR, -DQ and -DP. Expression of MHC class II molecules is more restricted as only antigen presenting cells, e.g. B cells, macrophages, dendritic cells and activated T lymphocytes, express class II molecules. The peptides bound by MHC class II molecules are exogenously derived like bacterial proteins or viral capsid proteins, and are recognised by CD4 T lymphocytes. The class II molecules are also synthesised in the endoplasmatic reticulum, however, for peptide binding they are transported to endosomal/lysosomal compartments. Class II molecules are heterodimers consisting of an α- and β-chain, that folds to yield the peptide binding groove with their amino terminal ends. In contrast to the class I molecules, class II molecules can bind peptides of varying length (10 to 24 amino acid residues), as the ends of the peptide binding groove are not "closed" by specific hydrophobic residues and antigenic peptides extend on either side of the binding groove [6].

In both the MHC class I and class II antigen processing pathways, accessory molecules are involved in complex formation of antigenic peptide and MHC molecule. For instance, the transporter associated with antigen processing (TAP) is encoded in the MHC class II region, and is required for the translocation of peptides present in the cytosol to the lumen of the endoplasmatic reticulum [11]. In the MHC class I pathway of antigen processing, tapasin acts as a chaperone for TAP and the MHC class I heavy chain [12]. HLA-DM is another accessory molecule that participates in the fully assembly of the MHC-class-II-peptide complex. On B cells, HLA-DM activity is modulated by HLA-DO [13]. The molecular chaperon calnexin facilitates the assembly of the MHC class I heavy chain and the β_2 microglobulin domain by making the folding of the heavy chain more efficient [9].

2.1.2 MHC class III genes

The MHC class III region contains the complement genes C4A, C4B, factor B (BF) and C2 as well as the TNF and BAT1 genes [14]. However, there are about 50 additional genes located in the region. The physiological role of many of the genes has not yet been determined, although many are involved in immune and inflammatory responses [3]. The class III region of the MHC is extremely densely packed and it is notable, that two-thirds of the intergenic regions are less than 1 kb in size. Some genes reside within

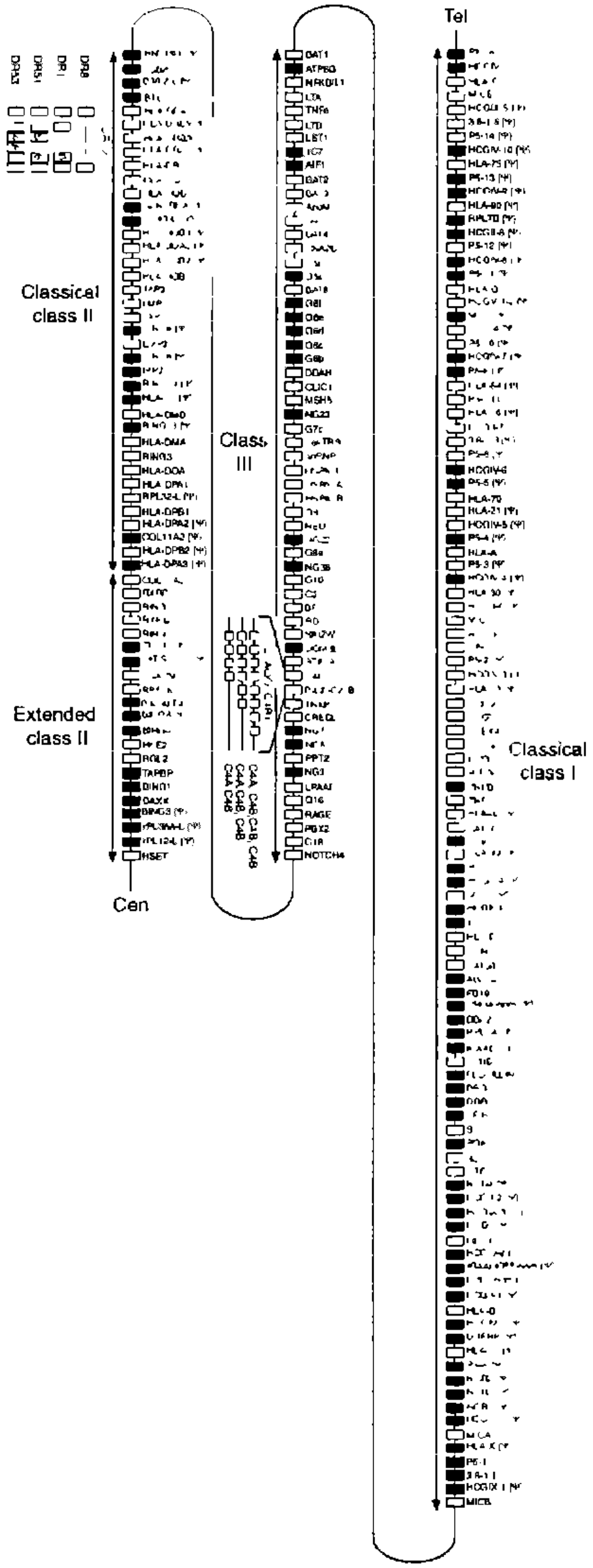

Figure 1: Gene map of the MHC. Genes are shown in order from telomere to centromere but not to scale. Gene loci that were discovered or located to the MHC as a direct result of the genomic sequence are indicated by filled boxes As will be the case for the rest of the human genome, the MHC reference sequence is a composite of different haplotypes. However, in regions with known differences in gene content (C4 region in class III and DR region in the classical class II region) only single haplotypes were sequenced (C4AQ0, C4B1 in the class III region and DR52 in the class II region). Obtained from Ref. [1].

another or have overlapping ends. For instance, the tenascin-X (TNXB) and the steroid 21-hydroxylase (CYP21B) gene overlap by 484 bp at the 3' end [2].

The MHC complement genes C4A, C4B, BF and C2 are located in the centromeric part of the class III region [2]. All four genes lie in close proximity following the same pattern observed at other regions within the MHC where genes with related function are often found clustered together [15] The genes encoding C2 and BF are the result of an ancient gene duplication, and the gene products show about 40% amino acid sequence identity. C2 serves as the enzymatic subunit of the classical and lectin pathway C3/C5 convertase. BF has a function analogous to C2 in the alternative pathway of complement activation.

Present between complement genes C4A and BF are four ubiquitously expressed genes, RD, SKI2W, DOM3Z and RP1 [16]. RD is located 205 bp downstream of BF, and organised in a tail-to-tail configuration with BF. The protein encoded by the RD gene is a subunit of a negative elongation factor for transcription of mRNA. The SKI2W protein contains a RNA helicase domain and two leucine zipper motifs, that may be involved in protein-protein interactions [16]. Human SKI2W is present in the nucleus and in the cytoplasm, where it is associated with polysomes and ribosomes [17]. DOM3Z is a nuclear protein. Experiments on its homologue in yeast suggest that it interacts with a nuclear 5' to 3' exoribonuclease [2]. RP1 (G11) encodes a Ser/Thr protein kinase, however, the physiological function of the RP1 gene product is yet to be determined. The presence of a nuclear localisation signal indicates that it is probably a nuclear protein [16]. The four genes are organised as two head-to-head configured gene pairs, RD-SKI2W and DOM3Z-RP1. Their ubiquitous gene expression suggests that these are probably housekeeping genes.

Several genes present in the central part of the MHC may play a critical role in growth, development and differentiation [2]. The large extracellular matrix protein tenascin-X is encoded by the TNXB gene, and is expressed in connective tissues [18]. PBX2 (G17), a homeodomain-containing protein, forms complexes with other homeobox proteins including the proto-oncogene HOX, which increases DNA-binding specificity [19,20]. At the centromeric end of the MHC class III region lies NOTCH4, the human counterpart of the mouse mammary tumor gene int-3 [21,22].

Other important proteins are the receptor for advanced glycosylation end products (RAGE), which is a member of the immunoglobulin family [23,24], LPAAT-α required for acylation of glycerols, the palmitoyl protein thioesterase PPT2 and CREB-RP, a transcription factor inhibiting the induction of glucose-regulated proteins [2]. CYP21B encodes the steroid 21-hydroxylase, which is important in biosynthesis of glucocorticoids and mineralocorticoids [25].

In the telomeric region of the MHC class III reside several genes that are involved in inflammatory responses. Three genes encode proteins of the major heat shock family and are named HSP70-1, -2 and -Hom in humans [26]. The TNF locus contains the genes coding for TNF-α, lymphotoxin α, lymphotoxin β and the human homologue of the mouse B144-Lst1 gene [27,28]. Between the heat shock protein 70 and TNF genes, the seven genes G7-G6-G6A-G6B-G6C-G6D-G6E have been identified [29]. Three of the genes (G6C, G6D, G6E) encode proteins of the Ly-6 superfamily. Most members of this family are extracellular GPI-anchored proteins that are specifically expressed and have a definite or putative immune related function [29, 30]. The protein encoded by G6B is a putative cell surface receptor of the Ig superfamily containing a potential signal peptide, a single Ig V-like domain and a transmembrane segment [31]. The G6 gene encodes a regulatory nuclear chloride ion channel protein, while G6A encodes a putative homologue of the N^G,N^G-dimethylarginine dimethylaminohydrolase (DDAH) [29]. This enzyme has first been described in rat kidney, and is thought to be involved in the regulation of the nitric oxide-generating system [32].

2.1.3 MHC ancestral haplotypes

A highly relevant feature of the MHC is its *en bloc* mode of inheritance. Within certain blocks several hundred kilobases in length recombination seems to be inhibited, whereas the boundaries between such blocks have been shown to be "hotspots" of recombination [33]. As recombination is not found within these blocks, they are often referred to as evolutionary "frozen" [34]. HLA-A is located in the alpha block of the MHC, HLA-B and HLA-C are genes of the beta block, the C2, Bf and C4 genes are located in the gamma block and HLA-DR and HLA-DQ genes are located within the delta block (Figure 2). Additional blocks are located telomeric of HLA-A, between complement C2 and MICB, between HLA-C and HLA-J (the sigma and kappa block, respectively) and centromeric of HLA-DQ. The exact boundaries of many of these blocks (especially of the blocks in the central region of the MHC) have not yet been determined. However, SNP profiles and recombinant disease mapping studies indicate that recombination sites are present between HLA-DR and C4, C2 and HSP70, TNF and MICB, HLA-C and HLA-S, HLA-S and HLA-E, HLA-E and HLA-J, and probably also between MICB and MICA [15,35–39].

The polymorphic frozen blocks are linked to form haplotypes. A haplotype is a unique combination of alleles from closely linked loci found on a single chromosome, often having functional affinity [34]. The term ancestral haplotype (AH) has been used to describe conserved population haplotypes that are transmitted over several generations

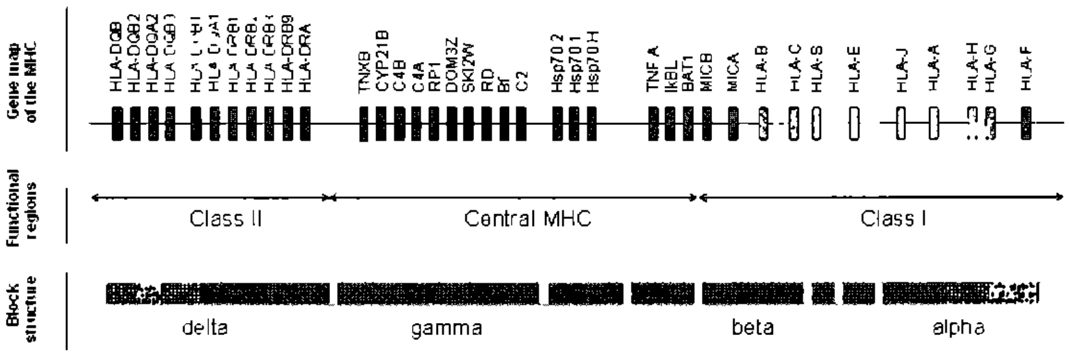

Figure 2: Organisation of the MHC in polymorphic frozen blocks. The map demonstrates that the functional regions of the MHC are structurally organised in various polymorphic frozen blocks Within these blocks recombination seems to be inhibited whereas the boundaries of the blocks have been shown to be hotspots of recombination. Block boundaries according to References [15,35–39]. The exact boundaries of most blocks are yet to be determined. Map not drawn to scale.

[40,41]. AHs have been shown to have a particular length and a unique content of alleles, deletions and duplications [42].

The number of alleles identified for many expressed MHC loci is enormous. For example, at some classical HLA loci more than 200 allelic variants have been reported [5]. Thus, the number of haplotypes expected for the human population would be immense. The number of haplotypes observed in populations is far below the theoretical expectations. Certain alleles tend to occur together on the same haplotype rather than randomly segregating. This allelic association is also termed "linkage disequilibrium" [43]. A study of a Caucasian population showed that at least 70% of the haplotypes found in this population are accounted for by 30 AHs and their recombinants [44]. Hence, a relatively limited number of AHs remained conserved during human evolution. Ancestral haplotypes have been named using the HLA-B allele as this is the most polymorphic HLA locus (e.g. the 8.1 AH is named after the HLA-B8 allele, and the 18.1, 18.2 and 18.3 AHs are characterised by HLA-B18) [45]. A list of ancestral haplotypes is shown in Table 1.

2.1.4 MHC ancestral haplotypes and associated diseases

The MHC is well-known for its association with numerous diseases, particularly autoimmune diseases like systemic lupus erythematosus, insulin-dependent diabetes mellitus (IDDM), myasthenia gravis, progression of HIV disease and IgA deficiency. Autoimmunity is probably due to immune dysregulation and defects in self-nonself discrimination [56], which may lead to the activation of autoreactive T cells that have escaped thymus negative selection.

A haplotypic association is usually stronger than an allelic association, which makes it difficult to determine the MHC loci that are involved in the various diseases mapped to

	HLA-			Central non-HLA				HLA-				
AH	**A**	**Cw**	**B**	**C2**	**Bf**	**C4A**	**C4B**	**DR**	**DQ**	**Race**	**Disease association**	**Reference**
7.1	3	7	7	C	S	3	1	15	6	C	SLE, MS, CD, HM	[40]
7.2	24	7	7	C	S	3+3	1	1	5	M		
8.1	1	7	8	C	S	Q0	1	3	2	C	IDDM, SLE, MG, IgAD..	[46,47]
13.1	30	6	13	C	S	3	1	7	2	C/M	PV	[48]
18 1	25		18	Q0	S	4	2	15	6	C	SLE, C2D	[40]
18.2	30	5	18	C	F1	3	Q0	3	2	C	IDDM (Sar)	[40,49–51]
18.3			18		S	3	1	11	7	C		
35 2	11	4	35	C	F	3+2	Q0	1	5	C	HIV rapid progression	[40,52]
35.3	11	4	35		S	3	Q0	1	5	C	HIV rapid progression	[40,52]
35 4			35		S	3	1	5				
35.5			35		S	3	1	5				
37.1	1	6	37	C	F	3	1	10		C		
38 1	26		38		S	2	1	4	8	C		
42 1		2	42	C	F	12–91	Q0	3	4	N		
44 1	2	5	44	C	S	3-3	Q0	4	7	C	RA	[40]
44.2	29		44	C	F	3	1	7	2	C	IgAD, CD	[40]
44.3	29		44		S	Q0	1	7	2	M		
44.4	33		44	C	F	3	1	13	6	M		
46.1	2	1	46	C	S	4	2	9	9	M	MG (Th,S)	[41]
46.2	2	1	46	C	S	4	2	8	6	M		
47 1	3	6	47	C	F	91	Q0	7	2	C	21D	[40,53]
50.1			50	C	S07	2	1+2	7		C	IDDM (I)	[41]
51.1			51	C	F	3	Q0	4	3	C		
52 1	24		52	C	S	3+2	Q0	15	6	M		
52 2			52	C	FT	3	1	9	9	M	IDDM (J)	[53]
54 1		1	54	C	S	3	5	4	4	M	IDDM (J)	[51,54]
55 1		3	55	B	S	4	5	14		C	21D	[53]
57 1	1	6	57	C	S	6	1	7	9	C/M/N	IgAD, PV	[40,41,48]
58.1	33	3	58		S	3	Q0	3	2	M	IDDM (Ch)	[40,51]
58.2	33		58		F	Q0	1	13		M		
59.1		1	59	C	S	3	5	9	9	M		
60 1		3	60	C	S	3	1	4	3	C		
60.2		3	60	C	S	3	Q0	8	4	C		
60 3	2	3	60	C	S	Q0	2	13	6	C		
61 1	26	3	61	C	S	3	1	9	9	M	IDDM (J), RA	[40,51,53]
62.1	2	3	62	C	S	3	3	4	8	C	IDDM, RA	[40,49–51]
62 2		3	62	B	S	4	2	4	8	C	IDDM	[53]
62.3			62	C	F	3	1	13	6	C		
62 4t	11	4	62	C	S	3	1	7		C/M		
64.1		8	64	C	S	3	1	4	2	C		
65.1		8	65	C	S	2	1+2	1	5	C	21D, IgAD	[40,53]
65.2		8	65	C	F	3	1	13	6	C		

Table 1: List of ancestral haplotypes, their alleles and disease associations. Reproduced from Ref. [55].

IDDM, insulin-dependent diabetes mellitus, SLE, systemic lupus erythematosus, 21D, 21-hydroxylase deficiency; MG, myasthenia gravis; MS, multiple sclerosis; IgAD, IgA deficiency; RA, rheumatoid arthritis; PV, psoriasis vulgaris; CD, celiac disease; HM, hemochromatosis; C2D, C2 deficiency; C, Caucasoid; M, Mongoloid; N, Negroid [45,51], Ch, Chinese, I, Indian; J, Japanese; S, Singapore; Sar, Sardinians, Th, Thai; t, tentative assignment the region of the MHC. Association between a particular MHC gene and disease may therefore be due to the direct affect of a gene or can be caused by the effects of linked genes carried by the haplotype.

The 8.1 AH is a particularly interesting haplotype as it has been extensively studied and is associated with a number of autoimmune diseases. The 8.1 AH (HLA-A1, Cw7, B8, C2C, Bfs, C4AQ0, C4B1, DRB1*0301, DRB3*0101, DQA1*0501, DQB1*0201) is characterised by a C4A null allele, i.e. resulting in the lack of C4A protein [46]. Thus, the 8.1 AH carries only one functional short C4B gene The consequences arising from the C4AQ0 allele include a prolonged persistence of immunising antigens and a reduced clearance of circulating immune complexes. The 8.1 AH is especially common in northern Europe, where it is the haplotype carried by most Caucasians who type for HLA-B8. Although this haplotype appears to contribute to a higher susceptibility to various autoimmune disorders, including insulin-dependent diabetes mellitus (IDDM), systemic lupus erythematosus (SLE), myasthenia gravis, rapid progression of HIV infection and IgA deficiency [47], over 10 million Europeans carry the 8.1 AH.

Various other ancestral haplotypes have been reported in association with autoimmune diseases. For example, the 8.1 AH has an increased frequency in Caucasians with IDDM, whereas in other populations the strongest association with IDDM has been found for the 18.2 AH (Sardinians), the 52.2 and 54.1 AHs in the Japanese population and the 50.1 AH in Indians [41, 53]. In contrast, the 7.1 AH in Caucasians and the 52.1 AH in Japanese are protective haplotypes and are rarely found in IDDM [51, 53]. The 46.1 AH in Thai and Singapore Chinese has an increased frequency in patients with myasthenia gravis, and the 57.1 AH is associated with psoriasis vulgaris [41]. For an overview of ancestral haplotypes and disease associations see Table 1.

2.2 Complement system

The complement system is an important component of the immune defence against infection. It is a strong antimicrobial system and can respond to pathogens before an adaptive immune respond has developed, but it is also essential for the operation of the antibody mediated response [57].

Complement was first discovered in the late nineteenth century as a heat-labile component of normal plasma and shown to have bacteriolytic activity [58]. Since then, more than 30 distinct plasma and cell surface proteins have been identified [59], which interact with each other resulting in a cascade of reactions. A number of complement proteins are proteases that are themselves activated by proteolytic cleavage at sites of infection. Complement activation results in the opsonization of pathogens for engulfment by phagocytes, some complement proteins act as chemoattractants for phagocytic cells and the terminal complement components damage pathogens by creating pores in the bacterial membrane [60]. Activation of the complement system promotes the clearance of immune complexes, and complement also participates in the regulation of the B cell response to antigens [61,62]. Furthermore, complement opsonizes apoptotic cells for fast clearance by phagocytic cells [63].

There are three distinct pathways through which complement can be activated. The classical pathway participates in antibody-mediated immune response or can be initiated by the direct binding of complement components to pathogen surfaces [64]. The lectin pathway uses mannan-binding lectin to recognise sugar residues that are present on the surfaces of many pathogens [65], and the alternative pathway is initiated by the binding of spontaneously activated complement components to pathogens [66]. All three pathways converge to initiate the assembly of the terminal components of complement to form a membrane-attack complex (MAC), which creates pores in lipid membranes and thus causes damage to target cells. Figure 3 gives an overview of the complement pathways.

As activation of complement results in a cascade of reactions in which a high degree of amplification occurs, the system needs to be strictly regulated in order to prevent uncontrolled activation. Therefore, a large number of control proteins is required to specifically protect host cells and allow complement activation to proceed on pathogen surfaces.

The importance of the complement system in host defence becomes evident when observing the severe symptoms caused by complement deficiencies. Patients with deficiencies of classical pathway components for example have increased susceptibility to systemic lupus erythematosus-like diseases and bacterial infections. Those patients with defects in terminal complement components or the lectin pathway of complement activation are particularly prone to neisserial infections [67].

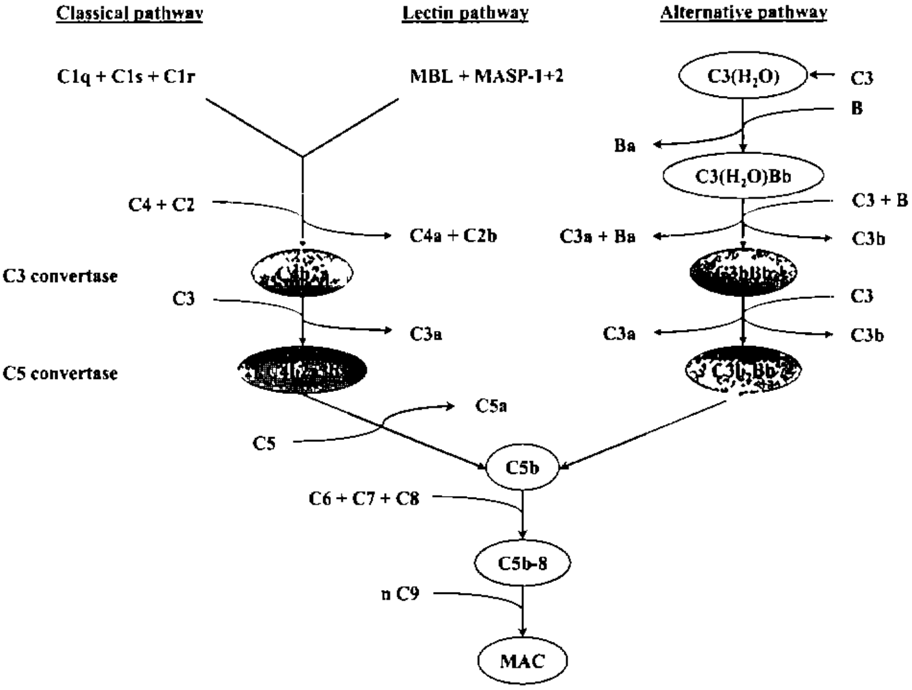

Figure 3: The complement pathways. Activation of the complement system occurs through three distinct pathways, the classical pathway, lectin pathway and alternative pathway, and leads to the assembly of the membrane-attack complex (MAC)

2.2.1 Classical pathway

The classical pathway plays a role in both innate and adaptive immunity. The first component of the classical pathway is C1q, which is part of a large protein complex called C1. C1q binds to the C_H2 domain of IgG or the C_H3 domain of IgM antibodies complexed with antigen, and therefore links the humoral immune response to the complement system [64] However, many other substances can also activate the classical pathway. These include nucleic acids, gram-positive and gram-negative bacteria, some viruses, soluble immune complexes, C-reactive protein [68] and apoptotic cells [69].

The C1 complex comprises a single C1q molecule bound to two molecules each of the zymogens C1r and C1s joined by Ca^{2+}. C1q itself has a complex quaternary structure with 6 tulip-like globular regions linked together by collagen-like stems [69]. The heads bind to the Fc part of surface-bound antibody or directly to the pathogen surface, causing a conformational change in C1r Subsequent activation of the autocatalytic enzymatic activity of C1r results in cleavage of C1s. Activated C1s then uses the next two components of the classical pathway, C4 and C2, as substrates. Proteolytic cleavage of C4 leads to the formation of a major fragment, C4b, and to the release of a smaller fragment with anaphylotoxin activity, C4a. The conformational change results in the exposure of an internal thioester in C4b, which is able to react with target surfaces. Covalently attached C4b then binds C2 [70, 71], making it susceptible to cleavage by C1s to produce C2a and C2b. C2a itself is an active serine protease and due to the binding to C4b it remains on the surface of the pathogen. The complex of C4b2a acts as the C3 convertase of the classical pathway, cleaving C3 to produce large amounts of C3b molecules. C3b binds covalently through its thioester bond to adjacent molecules, which leads to the opsonization of pathogen surfaces to mark them for destruction by phagocytes. C3b also binds to the C4b2a complex forming the C5 convertase, C4b2a3b. C3b in this complex binds and orients C5 for cleavage by C2a, resulting in the generation of C5a and C5b. C5a acts as a mediator of inflammation, whereas C5b initiates the assembly of the membrane-attack complex.

The small complement fragments C3a, C4a and C5a act on specific receptors to produce local inflammatory responses. They are often referred to as anaphylotoxins, as production of large amounts can lead to a syndrome called anaphylactic shock. C4a has the lowest activity as mediator of inflammation, but all three peptides induce smooth muscle contraction and increase vascular permeability [72]. C3a and C5a also stimulate neutrophils, eosinophils, phagocytes and endothelial cells at sites of infection.

2.2.2 Lectin pathway

The lectin pathway of the innate immune system is the most recently discovered pathway of complement activation and is very similar to the classical pathway. The lectin pathway uses a protein, called mannan-binding lectin (MBL), which is structurally related to C1q of the classical pathway [73]. Mannan-binding lectin consists of three to six identical subunits, each containing three identical chains with a globular carbohydrate recognition domain and a collagen-like region.

MBL binds specifically to terminal non-reducing sugars, including N-acetylglucosamin, mannose, fucose and glucose residues represented by a wide range of pathological bacteria, viruses, fungi and parasites [74]. Like the classical pathway, complement activation through MBL involves two serine proteases, called MBL-associated serine proteases, MASP-1 and MASP-2. The amino acid sequence of MASP-1 and MASP-2 is closely homologous to that of C1r and C1s, and all four proteins are modular serine proteases exhibiting homologous structural organisation [73]. When MBL binds to pathogen surfaces, MASP-1 and MASP-2 are activated, resulting in the cleavage of C4 and C2. Thus, the C3 convertase of the lectin pathway is the same as the C3 convertase of the classical pathway, and hence the rest of the complement cascade.

Recently, a third MBL-associated serine protease (MASP-3) has been discovered, which shares domain organisation with the other two MBL-associated proteases. Also found in the MBL complex is a small protein of 19 kDa referred to as MAp19, but no physiological function has been reported yet [75].

2.2.3 Alternative pathway

The alternative pathway is considered phylogenetically earliest of the complement pathways [76] and does not depend on a pathogen binding protein for its initiation. Instead, the thioester bond in C3 is spontaneously hydrolysed to form C3(H_2O). The conformational change allows the binding of the plasma protein factor B to C3(H_2O), which in turn becomes cleaved by factor D. The large fragment, Bb, remains associated with C3(H_2O) to form C3(H_2O)Bb, the fluid-phase C3 convertase of the alternative pathway. The exposed thioester of C3b is highly reactive with a half-life significantly less than one second [77]. Most of the activated molecules are neutralised by molecules in the surrounding medium (especially water), however, some of the C3b molecules attach covalently to nearby target surfaces. Deposition of C3b to the surfaces of pathogens enhances the binding of more factor B, allowing its cleavage by factor D to yield the small fragment Ba and the active protease Bb. This results in the formation of the C3 convertase, C3bBb, which is stabilised by a positive regulatory factor called properdin. The C3bBb complex rapidly cleaves yet more C3, thus establishing a positive feedback loop. The addition of another C3b molecule to the C3bBb complex results in the formation of the C5 convertase, $C3b_2Bb$, which cleaves C5 and therefore initiates the assembly of the terminal complement components to form the membrane-attack complex. However, the monomeric convertase C3bBb has also been shown to cleave C5 without the help of an additional C3b molecule, but has much lower affinity for C5 [78].

C3 is an abundant protein in plasma, and due its spontaneous activation the alternative pathway has continuously low activity. To prevent complement activation on host cell surfaces, a number of complement regulatory proteins are present in plasma and on host cell membranes. For example, factor H preferentially binds to C3b bound to vertebrate cells. When C3b forms a complex with factor H, it is rapidly cleaved by factor I to form its inactive derivative iC3b. Decay-accelerating factor (DAF or CD55) and CR1 are both linked to host cell surfaces and enhance the decay of the C3 and C5 convertases [79, 80].

2.2.4 Terminal pathway

The first step in the formation of the membrane-attack complex is the cleavage of C5 by a C5 convertase to release the large fragment C5b and a small peptide, C5a, which acts as a mediator of inflammation. While still attached to the cleaving enzyme, C5b binds one molecule of C6, and the C5b6 complex then binds one molecule of C7. Binding of C7 induces a conformational change resulting in the exposure of a hydrophobic site through which the complex can bind to lipid bilayers. C5b of the membrane-bound complex binds to the β-chain of C8, and that allows the hydrophobic domain C8αγ to insert into the lipid bilayer. C8αγ then binds the final component of the pathway, C9, which again leads to the exposure of a hydrophobic site on C9 and also reveals a binding site allowing additional C9 molecules to attach. Ten to 16 molecules of C9 polymerise to form a pore in the target cell membrane, leading to the free passage of water and other small molecules across the lipid bilayer. The eventual destruction of the pathogen is caused by the disturbance of homeostasis, a change in ion gradient and the penetration of lysing enzymes into the cell.

2.2.5 Complement deficiency and associated disease

Activation of the complement system involves a cascade of several reactions. Therefore, deficiency in one of the complement components can lead to a defect of the whole system. Genetic polymorphisms and deficiency states have been described for the majority of human complement proteins [67] and result in reduced protein level, abnormal protein synthesis or complete lack of protein production. Deficiency of complement proteins can be caused by inherited gene deletions, mutations that induce frame-shifts and lead to premature stop codons or splice site mutations that interfere with the processing of the mRNA. Depletion of circulating complement components may be caused by autoantibodies against complement components, thus resulting in a secondary deficiency [81]. Insufficiency in the regulation of the complement cascade may also lead to unrestricted complement consumption. The significance of complement in inflammation becomes evident as patients with complement deficiencies almost invariably have increased susceptibility to infections and immune complex diseases [82].

2.3 Complement component C4

The complement protein C4 is a non-catalytic subunit of classical pathway and lectin pathway C3/C5 convertases. Covalent binding of C4 opsonizes antigens for phagocytosis, enhances the solubilization of immune aggregates and is also involved in the clearance of immune complexes through complement receptor 1 (CR1) on erythrocytes [83]. Human C4 is among the most complex and polymorphic molecules of the complement system.

2.3.1 C4 protein

The C4 protein is synthesised from a 5.5 kb mRNA, and mainly expressed and secreted in the liver as a ~ 200 kDA glycoprotein. However, biosynthesis of C4 has also been reported in peripheral blood monocytes, skin fibroblasts, glomerular mesangial cells and epithelial cells of the lung, intestine and kidney [84]. Synthesis of C4 at extrahepatic sites may be important for local protection and inflammatory response. C4 expression is regulated by the proinflammatory cytokine INF-γ [85], whereas other cytokines, TNF-α, IL-2 and IL-6, were reported to have no effect on C4 levels. Stimulation with IFN-γ results in a two- to threefold upregulation of C4 protein synthesis, which is probably due to an increase in mRNA stability [84]

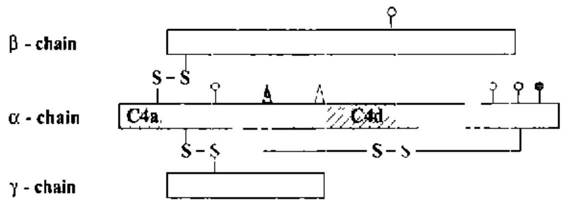

Figure 4: Schematic structure of the C4 protein. The C4 molecules are composed of three disulfide-linked polypeptide chains: a 93 kDa α-chain, a 75 kDa β-chain and a 33 kDa γ-chain. The thioester site is marked with a filled triangle, whereas the open triangle identifies the isotypic residues. Glycosylation sites are indicated by an open circle, sulphation sites are marked by a filled circle. Interchain disulfide bonds are represented by S-S.

C4 is synthesised as a single chain precursor protein and processed to a disulfide-linked heterotrimer made up of a 93 kDa α-chain, a 75 kDa β-chain and a 33 kDa γ-chain (Figure 4) [86]. The C4 molecules found in plasma differ in size as the proteolytic cleavages are incomplete reactions, although no decrease in functional activity of the partially processed C4 molecules has been found [87]. Post-translational modifications also involve sulphation at three tyrosine-*O*-sulphation sites and glycosylation at four *N*-linked glycosylation sites (at residue 207 on the β-chain and residues 843, 1309 and 1372 on the α-chain) [83]. Upon proteolytic cleavage by C1s an internal thioester becomes exposed in the α-chain of the major fragment C4b. Inactivation of C4b by protein factor I in the presence of one of the cofactors, CR1 or C4BP, results in the formation of the C4d fragment [88]. C4d is known to be the most polymorphic domain of C4.

A wide range of C4 concentrations has been observed in the blood plasma of different haplotypes. Serum levels of C4 vary between 0.08 and 0.67g/l [89]. The primary cause for this variation is the different number of C4 alleles present on different haplotypes. However, other mechanisms may also be involved as in general C4B proteins have a higher abundance in the blood plasma compared to C4A protein levels [83]. Analysis of different C4 allotypes has revealed that the C4B genes resulting in high expression levels are all short C4 genes (i.e. are 6.4kb smaller than long C4 genes). Therefore, it has been suggested that the rate of C4 transcription could be influenced by the length of the C4 genes

2.3.1.1 Isotypes

C4 exists as two isotypes, C4A and C4B. Although sharing 99% homology at the sequence level, they have different chemical reactivities and binding affinities towards antigens and immune complexes. The activated form of C4A has a high binding affinity to amino group containing substrates, whereas C4B preferentially forms a covalent ester bond with hydroxyl groups [90]. C4B is much more hemolytically active than C4A due to the high number of hydroxyl groups present on the surface of erythrocytes [90, 91]. Activated C4B is very reactive with a half-life of less than one second [92]. Therefore, C4B activates the classical complement pathway specifically at sites of infection, and hence plays an important role in the defence against infection. C4A functions more in immunoclearance through binding to IgG in antibody-antigen aggregates or to antigens of immune complexes [83]. Covalently bound C4A is then caught by erythrocytes and phagocytes through CR1.

The structural differences that account for the functional properties of C4A and C4B are determined by four isotype specific residues in the C4d region of the α-chain. C4A is characterised by **Pro-Cys**-Pro-Val-**Leu-Asp** at residues 1101 - 1106, whereas **Leu-Ser**-Pro-Val-**Ile-His** are found on C4B [56, 91]. Experiments using site directed mutagenesis have identified His^{1106} as the key residue responsible for the binding properties of C4B [93]. Also, the reaction mechanism of the internal thioester has been found to be different for C4A and C4B. C4A binding occurs through a direct reaction between amino-nucleophils and thioester, while a two-step mechanism accounts for the covalent binding reactivity of C4B. First, the thioester is attacked by His^{1106} to form an acyl-imidazol intermediate. The released thiol anion acts as a base to catalyse the reaction between acyl-intermediate and hydroxyl nucleophile, including water [92].

2.3.1.2 Allotypes

In addition to the isotypic variation, typing of C4A and C4B revealed more than 40 C4 allotypes. The separation of C4 allotypes is mainly based on differences in electrophoretic mobility and hemolytic activity. The most common allotypes in the Caucasian population are C4A3 and C4B1. Other frequent allotypes include C4A2, C4A4 and C4A6 for C4A and C4B2, C4B3 and C4B5 for C4B [56, 94]. In total, 27 polymorphic residues have been identified in the C4 protein sequence. However, only a relative small number of C4A and C4B genes have been characterised at the DNA level. Serological typing and DNA sequencing of C4A3, C4B1, C4B3 and C4B5 revealed that these allotypes are not homogeneous and may be split into several subtypes [83, 95, 96]. Therefore, systematic analysis of further C4 DNA sequences may elucidate similar heterogeneities for other C4 alloytpes.

2.3.1.3 Rodgers and Chido antigenic determinants

The Rodgers (Rg) and Chido (Ch) antigenic determinants detected on erythrocytes originate from the deposition of C4d fragments on erythrocyte surfaces [95,97]. Alloantibodies against Rodgers and Chido antigens may be produced by blood transfusion recipients who are deficient for C4A or C4B [83]. Two Rodgers (Rg1 and Rg2) and six Chido (Ch1 to Ch6) determinants have been defined as well as one rare antigenic determinant termed WH [98–100].

Four polymorphic sites encoded within exons 25, 26 and 28 of the C4 gene account for the sequential epitopes Rg1, Ch1, Ch4, Ch5 and Ch6 (Figure 5). Rg2, Ch2, Ch3 and WH are conformational epitopes resulting from the interaction of two polymorphic sites. Rg2 is only found in combination with Rg1 and the hypothetical Rg3 epitope. The antigenic determinants Ch4 and Ch5 form the second Chido epitope, Ch2, whereas Ch3 depends on the presence of Ch1 and Ch6 on one C4 molecule. The low-frequency allele WH is a result of the combination Rg1 and Ch6 [101]. Generally, Rodger determinants are associated with C4A and Chido determinants with C4B, but reversed antigenicity has been observed for most determinants [95, 101, 102].

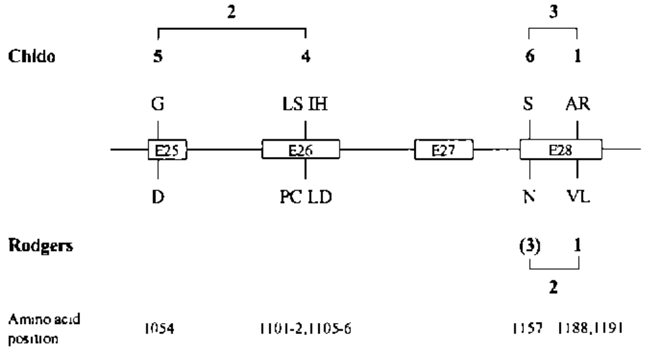

Figure 5: Schematic representation of the Rodgers- and Chido-specific amino acid differences in the C4d region. The amino acid positions are shown at the bottom The Rg3 determinant in parenthesis is not detectable directly by serology. Conformational epitopes are marked by a horizontal line, except for the WH epitope which is formed by Rg1 and Ch6

2.3.2 C4 genes

The complement C4 genes are arranged in tandem and are ~ 10 kb apart. Each C4 gene consists of 41 exons, altogether encoding 1744 amino acid residues of the pre-pro-C4 molecule. Exon 1 codes for a 19 residues leader sequence as well as for the first 3 amino acids of the β-chain. The proteolytic cleavage site for the β-α-chain junction is encoded in exon 16. Exons 16 to 33 code for the α-chain, the anaphylotoxin domain C4a being encoded by the exons 16 and 17. The sequences coding for the two factor I cleavage sites are located in exon 23 and 30. Cleavage of C4b by factor I results in the formation of the C4d fragment. The 291 amino acids of the γ-chain are encoded by exon 33 to 41. The first exon contains a 51 bp 5' untranslated sequence and the last exon includes 140 bp 3' untranslated sequence [87].

Comparison of the regions upstream of the C4 genes of different C4 alleles revealed an extensive conservation of the C4 promoter sequence [103, 104]. The 5' regulatory sequence of the human C4A and C4B genes does not involve a TATA-box. Studies using reporter gene assays found that maximal reporter gene expression is associated with the sequence contained within the -178 to -39 region [105]. The Sp1 site and the three E-boxes encoded in this region have been shown to be important for basal transcription of C4 in HepG2 cells as site-directed mutagenesis of the Sp1 binding site (-57 to -49) results in total loss of promoter activity. The E-box binding factors between -78 to -73 have been suggested to be responsible for IFN-γ induction of C4 gene expression [106].

2.3.2.1 HERV-K(C4)

The C4 genes can be either long or short due to the presence of the endogenous retrovirus HERV-K(C4) in intron 9 of the long genes. The length of the short C4 genes is 14.2 kb and the long C4 genes are 20.6 kb in size. Approximately 75% of the C4 genes contain the HERV insertion, 25% do not. The retroviral insertion is constantly present in C4A genes, whereas C4B genes may be long or short [107]. HERV-K(C4) contains a primer binding site for tRNA, two long terminal repeats (LTR) and the *gag*, *pol* and *env* genes [108]. HERV-K(C4) lies in opposite transcriptional direction with respect to the C4 coding sequence. Therefore, transcription of the long C4 genes results in the production of HERV-K(C4) antisense RNA. Antisense RNA specific for HERV-K(C4) has been found in cells constitutively expressing C4 [109]. The expression of other retroviruses is down-regulated in these cells, suggesting that the endogenous retrovirus might represent a cellular defence mechanism against further retroviral infections. It has also been suggested that HERVs play a role in the transcription regulation of adjacent genes. For example, expression of the human amylase gene is believed to be regulated by the HERV-E insertion in the 5'-flanking region of the gene [110,111]. Retroelements like human endogenous retroviruses may also contribute to allelic variation by providing sites for recombination and translocation events [111,112]. Other studies have focused on the potential involvement of HERVs in MHC associated diseases. However, no convincing evidence has been provided yet. A study on insulin-dependent diabetes mellitus (IDDM) in Germans could not reveal any preferential transmission of HERV-K(C4) to affected offspring [113].

2.3.2.2 RCCX module

The C4 genes lie adjacent to the genes encoding the serine/threonine nuclear protein kinase RP, steroid 21-hydroxylase CYP21 and extracellular matrix protein tenascin TNX. The four genes are usually duplicated together, and therefore form a genetic unit known as RCCX module [114]. Duplicated RCCX modules are generated by the addition of a long or short C4 gene, the CYP21A pseudogene or CYP21B gene, and the truncated gene fragments TNXA and RP2. The three pseudogenes/gene fragments are located between the two C4 loci and probably do not encode functional proteins. The truncated gene segment TNXA corresponds to intron 32 to exon 45 of TNXB, but contains a 120 bp deletion at the exon 36/intron 36 boundary resulting in a frame-shift and the generation of a premature stop codon [115]. RP2 is also a partially duplicated gene segment and corresponds to the last two and one-half exons of RP1. The CYP21A pseudogene contains three deleterious mutations (a 8 bp deletion in exon 3, a T nucleotide insertion in exon 7 and a C to T transition in exon 8) that probably render the gene non-functional [114]. A molecular map of a bimodular RCCX module is presented in Figure 6.

The human MHC varies in the number of RCCX modules with most containing two but many containing one or three modules. The frequency of the RCCX modular variation has been analysed in a study population of 150 healthy Caucasians. Seven different structures of the RCCX modules have been identified, including monomodular L (long) and S (short), bimodular LL and LS, and trimodular LLL, LLS and LSS modules [94]. This supports a dynamic "1-2-3 loci" model for the human C4 genes in the Caucasian population, rather than the two-loci theory proposed previously [56]. The C4A-C4B configuration only accounts for 55% of RCCX haplotypes. Bimodular haplotypes carrying either C4 isotype (C4A-C4A or C4B-C4B) have a frequency of 14%. The remaining 31% of RCCX haplotypes occur as monomodular (17%) or trimodular (14%) [94]. In addition, two other rare RCCX length variants have been identified. A haplotype carrying four long C4 genes (LLLL) has been reported in a CAH patient and in an Asian patient with SLE. Bimodular RCCX modules with two short C4 genes (SS) have been found in two white individuals and the native tribes of Brazil only [83]. Moreover, a recent study using novel or improved techniques identified a quadromodular RCCX structure with a SLSL configuration in a white patient with JRA, and a trimodular RCCX module expressing three C4A3 proteins has been found in a white female [116, 117].

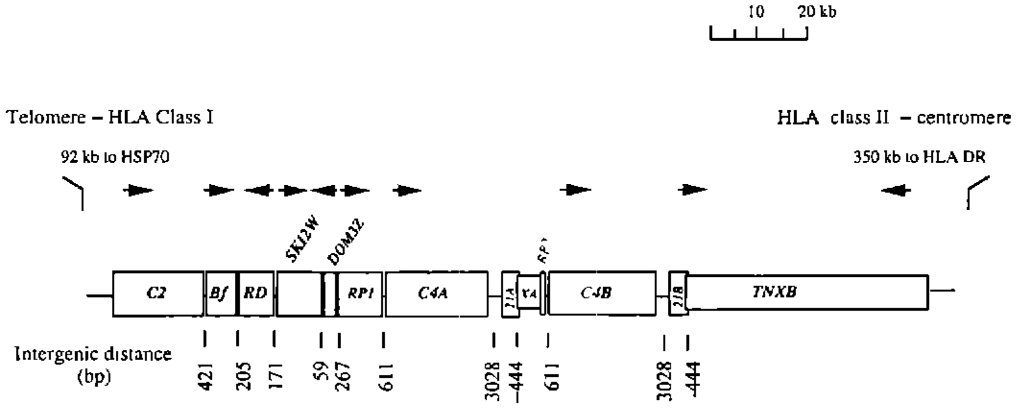

Figure 6: Molecular map of the human MHC class III complement region. A bimodular RCCX module (RP, complement C4, steroid hydroxylase CYP21 and tenascin TNX) is shown The directions of gene transcription are indicated by horizontal arrows Pseudogenes or partially duplicated gene segments are shaded. The negative signs for the intergenic distances between CYP21A and TNXA and between CYP21B and TNXB represent overlaps at the 3' ends of these genes Obtained from Ref. [114].

The high frequency of trimodular and monomodular structures may promote recombination or unequal crossovers between misaligned homologous chromosomes during meiosis. Recombination events further contribute to the diversity and polymorphism of RCCX modules and may lead to the acquisition of deleterious mutations from pseudogenes or gene fragments. Deletions of functional genes due to unequal crossovers have been reported in patients with CAH and JRA [94,114,115]. Recently, a de novo gene conversion has been also described between a C4A3a and a C4B1b gene resulting in a C4B5-like protein [115].

2.3.2.3 C4A and C4B null alleles

Partial deficiency of C4A or C4B has been observed in all populations studied so far and has a combined frequency of approximately 31% [83, 118]. In contrast, complete deficiency of C4 is a rare condition and has only been reported in a small number of individuals. C4AQ0 and C4BQ0 alleles may be due to the presence of monomodular RCCX modules with single C4A or C4B genes, or can be the result of C4A or C4B homoexpression at bimodular C4 loci. Non-expressed C4 genes (pseudogenes) caused by point mutations have a frequency of less than 1% [94].

A two base pair insertion (TC) has been identified after nucleotide position 5880 in exon 29 of C4AQ0 genes which leads to a frame-shift and generates a premature stop codon in exon 30 [119]. Recently, this insertion has also been found in a C4BQ0 allele [120], therefore providing the first molecular basis of a C4B pseudogene. It has been suggested that the C4BQ0 gene could have been acquired from a C4A pseudogene by an unequal cross-over or gene conversation event [120]. Characterisation of the non-expressed C4 genes in a patient with complete C4 deficiency detected a one base pair insertion in exon 20 of one of the C4 pseudogenes [121]. Deletion of a cytosine at position 3317 or 3318 results in a premature stop codon, which terminates translation of the C4 transcript. Another study investigating the molecular basis of complete C4 deficiency in a patient with SLE identified a novel single C nucleotide deletion in exon 13 of a C4B gene, causing a frame-shift mutation and premature termination [119].

C4A and C4B null alleles have a higher prevalence in patients with autoimmune or immune complex diseases [122].

2.3.2.4 Nucleotide polymorphisms in the C4 genes

To date, a limited number of complete C4 DNA sequences have been published at Genbank (http://www.ncbi.nlm.nih.gov/). Table 2 represents the nucleotide differences of the C4 coding region derived from a comparison of complete C4 sequences and also includes previously reported polymorphisms [83,87,104,119–121,123,124].

Twenty-seven polymorphic amino acid residues have been identified, most of which are located in the C4d region of the α-chain. Four amino acid changes represent the isotype

Exon	Codon	Substitutions	Comment
2	63	Phe (C/T)	
3	122	Leu/Val	
9	328	Tyr/Ser	
11	399	Val/Ala	
12	458	Arg/Trp	
12	459	Pro/Leu	
12	476	Ala (C/T)	
13	522	1-bp del (C)	Detected in C4BQ0 genes
15	616	Cys/Ser	
17	707	Pro/Leu	
17	708	Asp/Asn	
17	716	Cys (C/T)	
20	806	Val (C/T)	
20	811	1-bp del (C)	Detected in C4AQ0 genes
21	853	Val/Ala	
21	863	Gly (G/T)	
21	888	Ala/Thr	
24	1018	Leu (G/T)	
25	1054	Asp/Gly	Rodgers and Chido antigenic determinant
26	1076	Gly (A/C)	
26	1090	Ser/Ile	
26	1091	Gln/Ala	
26	1101	Pro/Leu	Isotypic residue
26	1102	Cys/Ser	Isotypic residue
26	1105	Leu/Ile	Isotypic residue
26	1106	Asp/His	Isotypic residue
28	1157	Asn/Ser	Rodgers and Chido antigenic determinant
28	1159	Phe(C/T)	
28	1182	Thr/Ser	
28	1186	Ala(G/C)	
28	1188	Val/Ala	Rodgers and Chido antigenic determinant
28	1191	Leu/Arg	Rodgers and Chido antigenic determinant
29	1213	2-bp ins (TC)	Detected in C4AQ0 and C4BQ0 genes
29	1223	Ser(G/A)	
29	1226	Pro(G/A)	
29	1267	Ala/Ser	
29	1281	Arg/Val	
30	1286	Thr/Gly	
30	1287	Val/Gly	
30	1298	Ile/Phe	
33	1401abc	AspTyrGlu/-	Deletion includes a tyrosine sulphation site
34	1478	Tyr/Asp	
40	1669	Leu(G/A)	

Table 2: Polymorphisms present in the C4A and C4B coding sequence. Polymorphisms have been derived from a comparison of complete C4 sequences available at GenBank. Polymorphisms reported in Ref. [83, 87, 104, 119–121, 123, 124] have also been included.

specific substitutions and four are responsible for the Rodgers and Chido antigenic determinants. The other polymorphic residues have been suggested to account for allotypic variations of the C4 proteins. For example, the Arg to Trp change in codon 458 has been suggested to be characteristic for the C4A6 allotype [123] and the Ala/Ser substitution at position 1267 is specific for C4A3a and C4A3b, respectively. In addition, 11 nucleotide polymorphisms are present in the coding region, that do not result in an amino acid change. One 2-bp insertion and two 1-bp deletions have been detected in non-expressed C4 genes. Moreover, a deletion of three amino acid residues (Asp, Tyr and Glu) has been reported near the carboxy terminal end of the α-chain, which also includes a tyrosine sulphation site [87]. Nucleotide substitutions and single-base insertions and deletions exist in the intronic sequence. The highest degree of polymorphism is found in intron 9 which contains the retrovirus HERV-K(C4) in the long C4 genes [104].

2.3.3 C4 associated diseases

The complex organisation and great genetic diversity of human C4A and C4B genes render C4 an excellent candidate gene for MHC associated disease studies. Partial and complete deficiencies of C4A and C4B occur with increased frequency in patients with autoimmune or immune complex diseases. For instance, complete C4A deficiency is almost invariably associated with systemic lupus erythematosus-like diseases [125]. However, also partial deficiency of C4A and/or C4B has been shown to be a risk factor for the development of SLE [126–129]. Homozygous deficiency of C4B is associated with susceptibility to recurrent viral and bacterial infections [130, 131]. In addition, insulin-dependent diabetes mellitus (IDDM) [114, 120], IgA deficiency and common variable immuno deficiency [132, 133], IgA nephropathy and Henoch-Schönlein purpura [134], autoimmune hepatitis [135], rapid progression of HIV infection [136, 137], sudden infant death syndrome [138, 139], multiple sclerosis [140] and vitiligo [141] have all been suggested to be associated with C4A or C4B null alleles. However, the disease associations could be due to various effects and do not necessarily reflect a direct effect of C4 null alleles on disease pathogenesis. C4 null genes could serve as a marker for other gamma block genes or reflect the effects of linked genes carried by the haplotype.

2.4 Recombinant mapping of disease susceptibility loci

A number of autoimmune diseases have been shown to be associated with various regions of the major histocompatibility complex. However, linkage disequilibrium across the MHC makes it extremely difficult to identify individual genes that contribute to disease susceptibility. One approach to determine disease susceptibility genes is based on genetic mapping of recombinant ancestral haplotypes [142, 143]. The approach involves the identification of useful markers that are specific for a disease associated haplotype. Markers include polymorphic microsatellites, single nucleotide polymorphisms (SNPs) and traditionally HLA alleles. HLA typing commonly involves typing for HLA-A, HLA-B and HLA-DR alleles. The MHC is organised in polymorphic frozen blocks within which recombination seems to be inhibited, whereas the boundaries of the blocks have been shown to be hotspots of recombination. HLA-A is located in the alpha block, HLA-B is located in the gamma block and HLA-DR is found in the delta block. However, the region between HLA-B and HLA-DR also contains the gamma block which is not represented by the classical HLA alleles. This region contains more than 50 genes, some of which may be relevant in susceptibility to MHC associated diseases. Therefore, more recently microsatellite and SNP markers have been used in recombinant mapping studies in addition to the classical HLA markers. Microsatellites are highly polymorphic tandem repeats occurring every 30-60 kb in euchromatic regions of the human genome [144]. More than 100 microsatellites have been described in the MHC region, most of which are $(CA)_n$ repeats [145]. Their abundance and high degree of informativeness make them useful markers in disease mapping studies. In addition, a high number of haplotypic and haplospecific SNPs have been identified within the MHC [146–148]. SNP markers in the central region of the MHC have, for example, been identified in the BAT1 and TNF genes [27, 149, 150] However, these genes are located at the telomeric end of the central MHC, outside the gamma block. Few of the genes located in the gamma block have been characterised at the molecular level, resulting in a lack of well characterised molecular markers for this region. By typing recombinant haplotypes for informative markers, the region shared by all recombinant haplotypes found in the disease group can be identified, and thus the region containing disease susceptibility genes.

Using similar approaches, several disease susceptibility genes have been mapped to the region of the central MHC [47]. For instance, the class III region haplotype D6S273, HSP70c, Bat2 138, TNFα2 is a significant risk factor for the development of rheumatoid arthritis (RA) [151]. A study on Italian patients with myasthenia gravis identified a region between the MHC class III genes C4 and TNF that might contain genes predisposing to the disease [152]. Two regions of the central MHC have been shown to contain genes that may affect IgA levels, and hence are associated with susceptibility to

IgA deficiency [153]. Furthermore, deficiency of TN-X due to mutations in the TNXB gene is a cause of Ehlers-Danlos syndrome [18]. Deletion of the CYP21 gene, which leads to 21-hydroxylase deficiency, is a common cause of congenital adrenal hyperplasia (CAH) [25].

In addition, susceptibility genes contributing to insulin-dependent diabetes mellitus have been suggested to be located in the central region of the MHC [51, 154].

2.4.1 Insulin-dependent diabetes mellitus

Insulin-dependent diabetes mellitus (IDDM) is an autoimmune disease that results from the T cell mediated destruction of the insulin producing pancreatic β cells [155]. The disease can exist for years in a pre-clinical phase before the characteristic symptoms of IDDM become evident [156]. These symptoms include hyperglycemia and an altered glucose metabolism associated with vascular complications as well as neuropathy. Several years before onset of IDDM, autoantibodies to β cell autoantigens can be detected in the serum of patients. Anti-insulin autoantibodies are present in approximately 50% of patients with newly diagnosed diabetes [157]. In 70-80% of patients with newly diagnosed IDDM, islet cell cytoplasmic autoantibodies are present that react with antigens located in the cytoplasm of pancreatic islet cells. Their frequency in normal subjects and non-diabetic relatives of patients with IDDM is 0.5% and 3-4%, respectively [156]. Critical autoantibodies are directed against glutamic acid decarboxylase (GAD65/67), ICA512 and IA-2β . Additional autoantigens seen during the development of IDDM include carboxypeptidase H, HSP60, glycolipids and other less characterised molecules [157]. Although autoantibodies play a major role in early diagnosis of IDDM, they appear to have no crucial pathologic role in the destruction of β cells. The component of the immune system that triggers β cell destruction is still unknown. However, some important insights into the pathogenesis of IDDM have been obtained from animal models of diabetes [158, 159]. Studies of the non-obese diabetic (NOD) mouse have revealed that both CD4 and CD8 T cells participate in β cell destruction. Furthermore, several regulatory and proinflammatory cytokines, including IFN-γ, IL-2 and TNF-α, may have an important pathogenic function [160].

2.4.1.1 Genetics of IDDM

IDDM is a polygenic disorder and a number of susceptibility loci have been identified. However, even in monozygotic twins the concordance rate is only 50%, indicating the importance of environmental factors in the development of the disease [161]. Although a number of environmental risk determinants, including viral infections, early infant diet and toxins, have been suggested to trigger disease, most studies have failed to find an association between these factors and anti-islet autoimmunity [162].

The genetic influences on the development and progression of IDDM are very complex, with a combination of several genes contributing to disease. Random genome searches have identified more than 15 susceptibility loci, termed IDDM1 to IDDM15 [163]. However, the most important genetic determinants are located within the major histocompatibility complex on chromosome 6p21 (IDDM1). This region accounts for about 45% of genetic susceptibility for the disease [164]. Another locus associated with IDDM is the insulin gene region on chromosome 11p5 (IDDM2), which contributes about 10% to the familial inheritance of IDDM [165]. For most of the other putative IDDM susceptibility loci, contribution to familial risk is small and significant linkage has been difficult to confirm in replicated studies [166].

2.4.1.2 MHC association

The association of the MHC region with IDDM has been known since the early 1970s, however, the strong linkage disequilibrium between genes in the MHC has made it extremely difficult to identify individual genes [167]. Now however, it has been clearly established that susceptibility or resistance to IDDM is associated with different MHC class II genes. HLA-DR4/DQ8 and HLA-DR3/DQ2 have been found to confer the highest risk of the disease in Caucasians. In contrast, HLA-DR15/DQ6 is associated with protection to IDDM [160]. While MHC class II haplotypes are the strongest genetic determinants contributing to disease susceptibility, they do not explain all MHC associations with IDDM [157]. Therefore, it has been suggested that central MHC genes may be involved in modulation of the disease [3, 51] Studies using single nucleotide polymorphisms (SNPs) and polymorphic microsatellite markers to map the boundaries of the IDDM susceptibility locus in the central MHC found that the region between HLA-B and TNF might be important [168–170]. However, few studies have included genes or markers within the gamma block, despite the high degree of polymorphism found for some gamma block genes (e.g. C4A and C4B). This may be in part related to the restricted number of defined SNP markers available for this region.

In addition, various ancestral haplotypes (AHs) have been identified that are increased in patients with IDDM. Other AHs were found to be protective. As ancestral haplotypes differ in their degree of conservation and frequency between ethnic populations, different AHs are associated with IDDM in different populations [168]. For example, in Sardinians the strongest association with IDDM has been found for the 18.2 AH [40], whereas in other Caucasian populations the 8.1 AH has an increased frequency in patients compared to controls [46]. In addition, the 62.1 and 62.2 AH have been characterised as diabetogenic haplotypes in Caucasians, whilst the 7.1 AH shows a negative association with IDDM, i.e. is protective [40, 51]. In the Japanese, the 52.2, 54.1 and 61.1 AHs confer high risk to the development of the disease. In contrast, the 52.1 AH is protective in the Japanese [53]. In both Caucasians and Japanese, the 44.1 AH is neutral, giving an example of a DR4 containing haplotype that is not associated with IDDM. Different diabetogenic haplotypes have been identified in other populations, for example the 58.1 AH in the Chinese [40] and the 50.1 AH in Indians [41].

3 Materials and methods

3.1 Characterisation of polymorphisms within 3.3 kb of sequence

3.1.1 Cell lines

Twenty-four Epstein-Barr Virus (EBV)-transformed cell lines from the 4th Asia-Oceania Histocompatibility workshop (4AOH) [55] and the 10th International Histocompatibility Workshop (10IHW) [171] panel were used in this study. Cells have been typed at various MHC loci, including HLA-A, HLA-B, HLA-C, C2, Bf, C4, HLA-DR and HLA-DQ, and were predicted to be homozygous for the MHC loci. The cell lines represent conserved ancestral haplotypes, spanning different racial groups. A list of cell lines used in this study is shown in Table 3.

AH	Local ID	10IHW	4AOH	HLA-A	HLA-C	HLA-B	C2	Bf	C4A	C4B	DR	DQ
7 1	R86 012367	9082	100041B	3	7	7	C	S	3	1	15	6
7 2	R87 004708	9130	100042Z	24	7	7	C	S	3+3	1	1	5
8 1	R85 001518	9132	100044V	1,24		8		S	Q0	1	3	2
13.1	R86 012333	9048	100048M	30	6	13	C	S	3	1	7	2
18.1	R86 012293	9008	100050A	25		18	Q0	S	4	2	15	6
18.2	R85 005054	9135	100051Y	2,19	5	18	C	F1	3	Q0	3	2
35.2	R86 012291	9006	100052W	11	4	35	C	F	3+2	Q0	1	5
38.1	R86 012311	9026	100009X	26		38	C	S	2	1	4	8
42 1	R86 012306	9021	100007B	68,30	2	42	C	F	12,91	Q0	18	4
44.1	R90 009217	9090	100054S	2	5	44	C	S	3+3	Q0	4	7
44.2	R86 012336	9051	100057K	29		44	C	F	3	1	7	2
44 3	R86 012335	9050	100058H	29		44		S	Q0	1	7	2
44.4	R86 012338	9053	100145P	33		44	C	S	3	1	13	6
46.1	R86 012361	9076	100059F	2	1	46	C	S	4	2	9	9
46.2	R86 012351	9066	100063R	2	1	46	C	S	4	2	8	6
47.1	R86 012332	9047	100064F	3	6	47	C	F	91	Q0	7	2
52 1	R87 004709	9142	100065M	24		52	C	S	3+2	Q0	15	6
54.1	R88 015375		100062T	24	1	54	C	S	3	5	4	4
55 1	R85 000862			1,10	3	55		S	4	5	6	
57.1	R92 023091	9052	100084G	2	6	57	C	S	6	1	7	9
58.1	R92 020751	9157	100087A	33	3	58		S	3	Q0	3	2
62 1	R86 012316	9031	100072Q	2	3	62	C	S	3	3	4	8
62.2	R90 026468		100074K	2	3	62		S	4	2	4	8
65.1	R86 012364	9079	100002N	33	8	65		S	2	1+2	1	5

Table 3: Detailed description of the cell lines used in this study. Cell lines were chosen from the 4th Asia-Oceania Histocompatibility workshop (4AOH) and the 10th International Histocompatibility Workshop (10WS) panel. R88 015375 and R90 026468 were not included in the 10WS panel, R85000862 was not included in the 4AOH and 10WS panels Blanks in the allele assignment represent blanks in the description of these cell lines.

3.1.2 DNA extraction

High molecular weight DNA was extracted from EBV cell lines using the QIAamp DNA Mini Kit. 200 µl of cells (20 × 10^6 cells/ml) were mixed with 20 µl Protease K and 200 µl Buffer AL. The mixture was vortexed for 30 s to ensure sufficient lysis of cells and incubated at 56°C for 10 min. Samples were centrifuged briefly, 200 µl ethanol (99.5%) was added and the samples were mixed by pulse-vortexing for 30 s. The samples were then transferred to QIAamp spin columns and centrifuged at 13000 rpm for 1 min. Columns were placed in a clean 2 ml collection tube and washed with 500 µl Buffer AW1. The wash step was repeated using 500 µl Buffer AW2. The columns were then placed in a 1.8 ml sterile eppendorf tube and 200 µl Buffer AE was added. After incubation at room temperature for 5 min, DNA was eluted by centrifugation at 8000 rpm for 1 min.

Quantitation of genomic DNA was performed on the Beckman DU530 spectrophotometer by measuring the optical density (OD) of a 1:10 DNA dilution at 260 nm. Purity of DNA samples was assessed by determining the absorbance ratio A260nm/A280nm and accepted if ratio was within target range of 1.5 and 2.0. DNA was stored at 4°C.

3.1.3 Co-amplification of C4A and C4B genes

Recently, a co-amplification approach of both C4 isotypes, C4A and C4B, was developed by David Sayer. The method covers the region from intron 16 to intron 28 of the C4 genes, which is 3.3 kb in size. This approach was used as screening method to identify polymorphic positions within the C4d region of the C4 genes from different ancestral haplotypes. In the previously performed study, 22 ancestral haplotypes have been described (AHs 7.1, 7.2, 13.1, 18.1, 18.2, 35.2, 38.1, 42.1, 44.1, 44.2, 44.3, 44.4, 46.1, 46.2, 47.1, 52.1, 54.1, 57.1, 58.1, 62.1 and 65.1, represented by the cell lines listed in Table 3).

In the present study, two additional cell lines were characterised using the same approach. These cell lines were R85 000862 and R86 012364 representing AHs 55.1 and 62.2, respectively (see Table 3).

3.1.3.1 PCR

For amplification of the 3.3 kb of sequence spanning from intron 16 to intron 28 of the C4 genes, six primer pairs were used: C4In16F/C4In19R, C4In19F/C4In20R, C4In20F/ C4In21R, C4Ex21F/C4In23R, C4In23F/C4In26R and C4In26F/C4In28R. The sequences of these primers are presented in Table 4. Each PCR reaction was carried out in a final volume of 50 µl containing 625 ng genomic DNA, 12.5 pmol of each forward and reverse primer, 10 pmol of each dNTP, 2.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.3, 50 mM KCl and 1 unit of Platinum Taq DNA polymerase. Thermocycling was performed in the Applied Biosystems GeneAmp PCR System 9700 using the following PCR conditions: 1 cycle at 96°C for 6 min; 35 cycles at 96°C for 30 s, 65°C for 30 s and 72°C for 2 min; 1 cycle at 72°C for 10 min and a final hold at 4°C.

Primer ID	Sequence	Sequ. tag	Fragment size
C4In16F	TCA CCC CCA CCT GGC CCT GCA G	M13F	
C4In19R	GGC CCA GAC AGG GTG ACA TC	M13R	606
C4In19F	TGT CTG GGC CTC AGG TGA CC	M13F	
C4In20R	CTC CTG TAT GCT CAG GCT C	M13R	418
C4In20F	CGG CCT GTC CTC TAT AAC TAC C	M13F	
C4In21R	AAC TCC AGG GAC AGA GTT GG	M13R	521
C4Ex21F	CCA AGG TTC TGC AGA TTG AG	M13F	
C4In23R	AGG TCT GAG GAC TCT GTG TC	M13R	620
C4In23F	TGA CAC AGA GTC CTC AGA CC	M13F	
C4In26R	GTC CTC CGA CAG GCG CTT C	M13R	870
C4In26F	GAA GCG CCT GTC GGA GGA C	M13F	
C4In28R	CCT CCT CTG AGT CTT CAT CC	M13R	563

Table 4: Oligonucleotide primers used for the amplification of the C4d region. All primers contain tags for sequencing with M13F or M13R primers, respectively. Fragment sizes range between 400 bp and 900 bp.

Amplification was checked by agarose gel electrophoresis. 5 µl of PCR product were mixed with 3 µl of loading buffer type IV and loaded into a 1% agarose gel with ethidium bromide. 5 µl of lambda plus DNA ladder were loaded in every gel as size standard. Electrophoresis was carried out at 150 volts for approximately 30 min. DNA was visualised under UV light and a pictures was taken.

The UltraClean PCR clean-up kit was used for purification of DNA from PCR reactions. 500 µl of SpinBind buffer were added to each 50 µl PCR reaction, mixtures were transferred to spin filter units and centrifuged for 1 min at 13000 rpm. The liquid flow-through was discarded and samples were washed with 300 µl of SpinClean buffer. For elution of DNA, 100 µl of Elution buffer (1:2) were added to each spin filter unit and tubes were centrifuged for 1 min at 13000 rpm. DNA was stored at -20°C until required.

3.1.3.2 Sequencing

DNA samples were sequenced using the Big Dye Terminator cycle sequencing kit. Sequencing reactions were set up in a final volume of 20 µl containing 2 µl purified PCR product, 2 µl of sequencing primer (1 ng/µl), 7 µl of 2.5 × Sequencing buffer and 2 µl of Big Dye Terminator mix. Each PCR product was sequenced forward and reverse with M13F and M13R sequencing primers, respectively (Table 5). Thermocycling conditions were: 25 cycles at 96°C for 10 s, 50°C for 5 s, 60°C for 4 min and a final hold at 4°C.

Big Dye Terminator amplified products were precipitated with 125 µl 65% ethanol containing 0.01 M NaOAc pH 5.2. After leaving samples for 15 min at room temperature in the dark, DNA was pelleted by centrifugation at 3200 rpm for 45 min. To remove the ethanol, tubes were inverted onto paper towels and centrifuged at 2000 rpm for 1 min. Pellets were washed with 200 µl of 70% ethanol, centrifuged at 3200 rpm for 10 min, tubes were inverted onto paper towels and inverted tubes were centrifuged at 2000 rpm for 1 min. The wash step was repeated. DNA pellets were vacuum dryed for 3 min and then stored at -20°C until required.

Further processing of samples was performed by the Western Australia Genome Resource Centre of the Department of Clinical Immunology, Royal Perth Hospital, including analysis of samples on the ABI 3100 sequencer.

Primer ID	Sequence
M13F	TGT AAA ACG ACG GCC AGT
M13R	CAG GAA ACA GCT ATG ACC

Table 5: Sequencing primers M13F and M13R.

3.1.3.3 Comparison of DNA sequences

The resulting sample files were transferred to a Macintosh system using the WS_FTP Limited Edition program and converted to Macintosh format with the conversion program File Type Win to Mac. Alignment of DNA sequences was performed with the MT Navigator PCC program. Heterozygote codes were assigned according to the IUPAC designation.

3.1.4 Isotype specific sequence analysis

To separate C4A and C4B genes, two isotype specific PCR methods were developed. The first PCR spans from intron 16 to the isotypic site in exon 26, the second covers the region from the isotypic site in exon 26 to intron 28. This method was used to assign the previously identified polymorphisms within the 3.3 kb of sequence (intron 16 to 28) to either C4 isotype. All 24 cell lines listed in Table 3 were included in this analysis.

3.1.4.1 Isotype specific PCR

The 2.7 kb fragment spanning from intron 16 to the C4A- and C4B-isotypic site in exon 26 was amplified with primers C4In16F and C4Aspec2R or C4Bspec2R, respectively (Table 6), using the Expand Long Template PCR system. PCR reactions contained 500 ng genomic DNA, 15 pmol of each primer, 17.5 pmol of each dNTP, 1.75 mM $MgCl_2$, 2.6 units of enzyme mix, and were carried out in a final volume of 50 µl. PCR conditions were: 1 cycle at 94°C for 2 min; 30 cycles at 94°C for 10 s, 65°C for 30 s and 68°C for 2 min; 1 cycle at 68°C for 7 min and a final hold at 4°C. The 0.6 kb fragment spanning from the isotypic site in exon 26 to intron 28 was amplified with the forward primers C4A26F and C4B26F, respectively, and reverse primer C4In28R (Table 6). A final volume of 50 µl containing 625 ng genomic DNA, 12.5 pmol of each forward and reverse primer, 10 pmol of each dNTP, 2.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.3, 50 mM KCl and 1 unit of Platinum Taq DNA polymerase was reacted using the following conditions: 1 cycle at 96°C for 6 min; 35 cycles at 96°C for 30 s, 65°C for 30 s and 72°C for 2 min; 1 cycle at 72°C for 10 min and a final hold at 4°C.

To determine if a PCR product was present, 5 µl of each sample were resolved on a 1% agarose gel. PCR products were purified using the UltraClean PCR clean-up kit as described above (see 3.1.3.1). For use in sequencing reactions, DNA was eluted with 100 µl of Elution buffer (1:2) and stored at -20°C until required. PCR products for cloning were eluted with 20-50 µl of Elution buffer (1:2) depending on the intensity of DNA bands on the agarose gel.

Primer ID	Sequence	Sequ. tag	Fragment size
C4In16F	TCA CCC CCA CCT GGC CCT GCA G	M13F	
C4Aspec2R	GCA CCT GCA TGC TCC TGT CTA A	M13R	
C4Bspec2R	GCA CCT GCA TGC TCC TAT GTA T	M13R	2.7 kb
C4A26F	GAC CTC TCT CCA GTG ATA C	-	
C4B26F	GAC CCC TGT CCA GTG TTA G	-	
C4In28R	CCT CCT CTG AGT CTT CAT CC	M13R	0 6 kb

Table 6: Oligonucleotide primers used for isotype specific amplification of C4 genes. Primers C4Aspec2R and C4A26F as well as C4Bspec2R and C4B26F bind specifically to the isotypic site in exon 26 of C4A and C4B genes, respectively. Primers C4In16F and C4In28R are conserved primers.

3.1.4.2 Cloning

PCR products of samples representing ancestral haplotypes with more than one C4A or C4B gene were cloned into the pGEM-T Easy vector to separate the genes. The ligation reactions were set up in a final volume of 10 µl containing 5 µl of purified PCR product, 12 mM Tris-HCl pH 7.8, 4 mM $MgCl_2$, 4 mM DTT, 0.4 mM ATP, 2% PEG and 50 ng pGEM-T Easy vector. Reactions were incubated over night at 4°C. 2 µl of each ligation reaction were mixed with 50 µl of JM109 competent cells. Heat shock transformation was carried out in a water bath at 42°C for 45 s and cells were returned to ice immediately. 950 µl of room-temperature SOC medium were added to each tube and cells were incubated for 1.5 hours at 37°C with shaking at 150 rpm. To obtain a sufficient number of colonies, cells were pelleted by centrifugation at 1000 rpm for 10 min and resuspended in 200 µl of SOC medium. 100 µl of each cell suspension were plated on LB plates with ampicillin/IPTG/X-Gal and plates were incubated overnight at 36°C. Transformants were screened for inserts by blue-white-selection. Successful cloning of a PCR product into the vector usually results in white colonies due to the interruption of the lacZ gene, which codes for β-galactosidase. Blue colonies are an indication for vectors without an insert, but may also result from PCR products that are in-frame with the lacZ gene.

The TempliPhi DNA amplification kit was used to extract and amplify DNA for cycle sequencing. About 10 white colonies were picked from each plate and resuspended in 20 µl of sterile TE buffer. 1 µl of each cell suspension was transferred into tubes containing 5 µl of Sample buffer. Samples were denatured at 95°C for 3 min. After cooling to 4°C, 5 µl of Reaction buffer and 0.2 µl of enzyme mix were added to each sample. The DNA was amplified at 30°C for 6 hours. The enzyme was then inactivated by heating the samples to 60°C for 10 min. The amplified DNA was diluted 5 fold with TE buffer for use as template in sequencing reactions and stored at -20°C until required.

3.1.4.3 Sequencing

Sequencing reactions were performed using the Applied Biosystems strategy for automated sequencing as described above (see 3.1.3.2). 2 µl of purified PCR product or 4 µl of amplified vector DNA were used as template. Primers M13F and M13R (Table 5) as well as internal primers C4In19R, C4In19F, C4In20R, C4In20F, C4In21R, C4Ex21F, C4In23R, C4In23F, C4In26R and C4In26R (Table 4) were used as sequencing primers at a concentration of 1 ng/µl. Primers were chosen to cover previously identified polymorphic sites.

3.1.4.4 Comparison of DNA sequences

Sample files were transferred to the Macintosh and analysed with the MT Navigator PCC program as described above (3.1.3.3). Isotype specific sequences were compared to composite sequences and previously identified polymorphisms were assigned to either C4 isotype.

The polymorphic sequences were used to generate a number of phylogenetic trees using the MEGA software (Molecular Evolutionary Genetics Analysis version 2.1). Polymorphic sites located within the 3.3 kb of sequence (spanning from intron 16 to 28) were used to construct the trees. Distances were calculated by the two-parameter method of Kimura' and phylogenetic trees were obtained by the neighbour-joining method (pairwise deletion, Bootstrap analysis). The phylogenetic trees were not rooted.

3.2 Identification of coding polymorphisms outside of C4d

3.2.1 Cell lines and DNA extraction

The twenty-four EBV cell lines described under 3.1.1 were used in this study. All cell lines represent different ancestral haplotypes from various ethnic origins. For a detailed description of the cell lines see Table 3. DNA was extracted with the QIAamp DNA Mini Kit as described under 3.1.2.

3.2.2 Amplification of coding regions

The most polymorphic region of the C4 genes was characterised as described under 3.1. However, polymorphisms coding for the variations observed at the protein level may also be located outside this region. Therefore, a method was designed to identify polymorphisms in all further exonic regions.

3.2.2.1 Primer design

The complete nucleotide sequences of human C4 available at GenBank (accession numbers AL 662849, AL 049547, AL 645922, NG 000013, AF 019413 and M 59815) were aligned to identify conserved regions within the polymorphic C4 genes. Primers were designed to bind within these conserved regions using Primer3 software (freely available at http://www-genome.wi.mit.edu/genome-software/other/primer3.html). Primer pairs were chosen to span one or two exons each, covering the coding regions from exon 1 to 16 and from exon 29 to 41. The size of the resulting 23 amplicons ranged from 250 bp to 600 bp. All oligonucleotide primers were obtained from Geneworks. A list of the primers and expected sizes of PCR fragments is presented in Table 7.

3.2.2.2 PCR

For each of the 23 regions, 625 ng of genomic DNA were amplified in a total volume of 50 µl. Each reaction contained 12.5 pmol of each forward and reverse primer, 10 pmol of each dNTP, 2.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.3, 50 mM KCl and 1 unit of Platinum Taq DNA polymerase. Amplification was performed in an Applied Biosystems GeneAmp PCR System 9700. The reaction mix was denatured at 96°C for 6 min, followed by 35 cycles at 96°C for 30 s, 65°C for 30 s, 72°C for 2 min and a final amplification cycle at 72°C for 10 min. The reaction mix was held at 4°C until required.

Primer ID	Sequence	Sequ. tag	Frag. size	Melting temp.	2nd temp.
E1F	GGA CAG GGT TAT TTC TGG GC	-			
E1R	TCT CCC TCA CTC CTG AAT CG	-	375	62°C	64°C
E2F	ATT CAG GAG TGA GGG AGA GC	-			
E2R	GAT GAC ACT TAC AAG ACA GAT GGG	-	349	61°C	62°C
E3F	TTC TCC TTC CAC GTT TCT CC	-			
E3R	GTT AAA GGT TGA GGC CCT GG	-	330	62°C	63°C
E4+5F	TCC CTC TGT GGA GTT TGA CC	-			
E4+5R	GGA GAG CCT AAC AGG AAT TGG	-	376	62°C	63°C
E6F	ATT CCT GTT AGG CTC TCC ACC	-			
E6R	TCT CCT TCC ACC CTT ATT TCC	-	319	61°C	62°C
E7F	TCT TTG AGC TGG AGT CTG ACC	-			
E7R	CTT CCC ATA GAT GTA CCT GTC G	-	297	61°C	-
E8F	CAG TAT GAA TGG GCT CCT GC	-			
E8R	ACT GAG TCT CCC ACC TCA CC	-	254	61°C	-
E9F	GGG CTC CTA GAT GAG GAT GG	-			
E9R	CTC AGA GGT CAG AGG CAA GG	-	342	61°C	-
E10+11F	CTC CTG TCC CTC TCT TCT TGG	-			
E10+11R	CAG GTG CGA ATA GGG TAG TAG C	-	586	61°C	62°C
E12F	CTA CCC TAT TCG CAC CTG ACC	-			
E12R	AGA GTG GTT GCC TCT TCA TGG	-	323	62°C	63°C
E13F	AAG AGG CAA CCA CTC TTG TCC	-			
E13R	CTA AAT CCA TGC CCT GTT GG	-	303	62°C	63°C
E14F	ATA CCG GGA CTG AAG GAA GC	-			
E14R	AGG AAG GAT ACA GAG CCA GG	-	377	63°C	64°C
E15+16F	CTG TGG TCT CCA TCT CCT GG	-			
E15-16R	TGG AGA GCC CAA GCT ACT GC	-	536	61°C	62°C
E29F	TAT AAG CAG GGG TGG GTT GG	-			
E29R	CCC TTG GTC TGA GGA CTA CC	-	404	63°C	65°C
E30F	ATT CCG CAG TAC CCA AGT AGG	-			
E30R	AGT GGT TCA CCA GGG AGT GG	-	291	62°C	63°C
E31F	CTT TGT GGA AAT GTG AGG TGG	-			
E31R	ACC AAC CCT GAG GTG TCT GC	-	302	62°C	64°C
E32+33F	ACA TGT CCC ACG TCC TCT CC	-			
E32+33R	AGA CGT GTG AGC TGT CGT CC	-	525	63°C	65°C
E34+35F	ACG ACA GCT CAC ACG TCT CC	-			
E34+35R	ATA CTC AGT AAA CCC GGT GCC	-	337	63°C	64°C
E36+37F	AGT GGG TCC CTC ATC TCT CC	-			
E36+37R	GAA CCC ATC AGA CAG TGT GG	-	469	63°C	64°C
E38F	CAA GTA AGA GCA GAC TCT TGG C	-			
E38R	GTT GGT GTC AGA GCA AAC AGG	-	366	63°C	64°C
E39F	TTT GCT CTG ACA CCA ACT TCC	-			
E39R	TCA CAC TTC CAG ATG GTC AGG	-	249	61°C	-
E40F	GCG AAG GTG GAA TGA GAG G	-			
E40R	AGG CAT CTG GCT TCT GAG G	-	268	62°C	-
E41F	TGT GCT CTC CGT TTC CAC C	-			
E41R	ACA CAG CAG TGC TTC CAG C	-	276	63°C	-

Table 7: Oligonucleotide primers used for the amplification of exons in non-C4d regions. Twenty-three primer pairs were used to amplify the coding regions spanning from exon 1 to exon 16 and from exon 29 to exon 41. As these primers do not contain the M13F or M13R sequencing tag, they have been used for amplification of DNA in PCR and sequencing reactions. The expected fragment size of all 23 PCR products is indicated Fragment sizes range between 250 bp and 600 bp. The melting temperature and the second temperature shown in the last columns represent the temperature at which amplicons were analysed on the dHPLC.

Isotype specific fragments spanning from exon 12 to the isotypic site in exon 26 were amplified using the Expand Long Template PCR system with E12F forward and C4Aspec2R or C4Bspec2R reverse primers, respectively (Tables 6 and 7). PCR reactions contained 500 ng genomic DNA, 15 pmol of each primer, 17.5 pmol of each dNTP, 1.75 mM $MgCl_2$, 2.6 units of enzyme mix, and were carried out in a final volume of 50 µl. PCR conditions were: 1 cycle at 94°C for 2 min; 10 cycles at 94°C for 10 s, 65°C for 30 s and 68°C for 2 min; 20 cycles at 94°C for 10 s, 60°C for 30 s and 68°C for 2 min plus an extension of 20 s per cycle; 1 cycle at 68°C for 7 min and a final hold at 4°C.

5 µl of each reaction mix were resolved on a 1% agarose gel with ethidium bromide to confirm the presence of amplification products.

3.2.3 Screening for polymorphic positions by dHPLC

Denaturing high performance liquid chromatography (dHPLC) is a method for detecting unknown single base substitutions and small insertions or deletions. Under non-denaturing conditions, DNA fragments are separated depending on the size of the fragment. Under partially denaturing conditions however, DNA sequence variations can be detected. The HPLC system provides a rapid and highly sensitive method, and was therefore used to screen the coding region of the C4 genes for polymorphic positions. Samples were analysed on the Varian Helix System.

3.2.3.1 Verification of product yield and purity

Unpurified PCR products of a reference sample were analysed under non-denaturing conditions to determine the size, yield and purity of amplification products. 3 µl of each of the 23 amplicons were injected and run at 50°C. The pUC18 standard was used to confirm the size of the fragments. Fragment sizes detected on the instrument were compared with expected fragment sizes for each amplicon to verify that the expected product is present. For a list of fragment sizes see Table 7. PCR reactions were optimised to give a sharp, single peak under non-denaturing conditions. The yield of PCR product was assessed by measuring the absorbance at 260 nm and was accepted if intensity was greater than 30 mV.

3.2.3.2 Temperature optimisation

The sensitivity of the method is strongly dependent on the temperature at which the amplification products are analysed. Therefore, the DHPLCMelt software (freely available at http://insertion/stanford.edu/melt.html) was used to determine the optimum temperature for detection of sequence variation in each of the 23 amplicons. For a given sequence, the DHPLCMelt software predicts the melting temperature and gradient conditions that will resolve heteroduplexes on the dHPLC. The predicted melting temperatures for all 23 amplicons ranged between 61°C and 63°C. A list of all individual melting temperatures is given in Table 7.

However, some amplicons may have multiple melting domains as GC-rich regions melt at higher temperatures than AT-rich regions. Therefore, polymorphisms located in various positions along a fragment may be detected at different temperatures. To achieve sensitivity of >96%, Jones et al. [172] recommend to run samples at two different temperatures. These are the predicted melting temperature (T_m) and the predicted T_m plus 2°C. As high sensitivity was required in this study, all amplicons were analysed at both recommended temperatures. In some cases, however, samples were melted completely at the predicted T_m plus 2°C or resulted in peaks with an intensity of less than 15 mV and could therefore not be interpreted. These particular amplicons were analysed at the predicted T_m plus 1°C as second temperature. Six of the 23 amplicons could only be analysed at one temperature as any further increase above the melting temperature resulted in complete melting. The temperatures used for analysis of all 23 amplicons are shown in Table 7.

3.2.3.3 Sample preparation

For analysis of unknown mutations on the dHPLC, it is necessary to mix each sample with a known homozygous reference. The 8.1 AH (local ID R85 001518) was selected as a reference as it is predicted to be homozygous and has only one C4 gene. To confirm that there are no heterozygous positions within the sequence of this sample, all 23 amplicons were sequenced in forward and reverse direction.

For each amplicon, 5 µl of unpurified PCR product were mixed with 5 µl of the PCR product of the 8.1 AH. Samples were denatured at 95°C for 4 min followed by slow renaturation (decrease of temperature in 5°C steps to 65°C every 4 min) to facilitate heteroduplex formation.

3.2.3.4 Analysis of samples

Analysis of samples was performed on the Varian Helix System. Prior to each run, performance of the system was evaluated using the pUC18 size standard or an internal heterozygote control. Given a sufficient resolution of the expected peaks, samples were injected using the autosampler of the Varian Helix System. For each run, 3 µl of sample were required. All samples were run at the temperatures indicated in Table 7 using the methods univ61, univ62, univ63, univ64 or univ65. The universal methods are designed for the analysis of amplicons with a size of 150 bp to 550 bp using the Varian Helix Column at a flow rate of 0.45 ml/min. DNA fragments were eluted from the column by an increasing acetonitrile gradient. The gradient was created by mixing Buffer A (100 mM TEAA pH 7.0, 0.1 mM EDTA) and Buffer B (100 mM TEAA pH 7.0, 0.1 mM EDTA, 25% (v/v) acetonitrile). For a detailed description of the universal program see Table 8. Eluted DNA fragments were detected by the ProStar 310 UV detector at 260 nm and data was stored by the Star software.

Chromatograms of each sample were compared to the chromatogram of the homozygote reference (8 1 AH). Differences in the peak pattern of the sample/reference mixture indicate heteroduplex formation. Heteroduplexes melt at the mismatch site, generating a single-stranded region under partially denaturing conditions [173]. This results in lower affinity to the column and therefore, heteroduplexes elute earlier than homoduplexes. The heteroduplex peak can be observed as separate peak or as shoulder of the homoduplex peak [174].

	Time	Buffer A	Buffer B	Flow rate	Detection
begin	0:00 min	55%	45%	0.45 ml/min	260 nm
	0:30 min	50%	50%	0.45 ml/min	260 nm
	6:00 min	25%	75%	0.45 ml/min	260 nm
	7:36 min	55%	45%	0.45 ml/min	260 nm
end	9:00 min	55%	45%	0.45 ml/min	260 nm

Table 8: The universal method for analysis of PCR fragments (150 bp to 550 bp) on the dHPLC. DNA is eluted by an increasing acetonitrile gradient and detected under UV light. Buffer A: 100 mM TEAA pH 7.0, 0 1 mM EDTA; Buffer B: 100 mM TEAA pH 7.0, 0.1 mM EDTA, 25% (v/v) acetonitrile.

3.2.4 Characterisation of detected polymorphisms

Samples that were identified to contain polymorphisms were further analysed by DNA sequencing. 20 µl of PCR product were purified with the UltraClean PCR clean-up kit as described under 3.1.3.1. DNA was eluted with 40 µl of Elution buffer (1:2). Sequencing reactions were performed with the Big Dye Terminator Cycle Sequencing Kit using 2 µl of purified PCR product as template. As PCR primers did not contain the sequencing tags M13F or M13R, the same primers were used in sequencing reactions at a concentration of 1 ng/µl. Thermocycling was performed in an Applied Biosystems GeneAmp PCR System 9700 using the following PCR conditions: 25 cycles at 96°C for 10 s, 50°C for 5 s and 60°C for 4 min and a final hold at 4°C. Amplified products were concentrated as described above (3.1.3.2) and further processed by the Western Australia Genome Resource Centre of the Department of Clinical Immunology, Royal Perth Hospital. The obtained sample files were analysed with the MT Navigator PCC program (see 3.1.3.3). DNA sequences of samples were aligned to the sequence of the C4B1 gene of the 8.1 AH and checked for the presence of polymorphisms.

3.3 Typing of a 62.1 specific single nucleotide polymorphism

3.3.1 Control subjects and IDDM patients

Control subjects were recruited from the Busselton Health study population. The Busselton population has been characterised by a number of studies during the past 30 years. Individuals were previously typed at various MHC loci including HLA-A, HLA-B, HLA-C, C4, HLA-DR and HLA-DQ, although typing was incomplete for some individuals.

IDDM patients were recruited from patients attending the outpatient diabetes clinic at Royal Perth Hospital (RPH), Western Australia. The 58 patients investigated in this study were typed for HLA-A, HLA-B, HLA-DR and HLA-DQ, and some patients were also typed for complement C4 and Bf.

3.3.2 Sequencing based SNP typing

Comparison of DNA sequences from different AHs revealed a 62.1 haplospecific single nucleotide polymorphism (SNP) in exon 17 (first base of codon 695) of the C4B3 gene (see Table 13). To confirm that the SNP is haplospecific, nine C4B3 positive and ten C4B3 negative control subjects were typed for the SNP allele by sequencing.

The region spanning from intron 16 to exon 26 was amplified by C4B specific PCR using the Expand Long Template PCR System. PCR reactions contained 500 ng genomic DNA, 15 pmol of C4In16F forward and C4Bspec2R reverse primer (Table 6), 17.5 pmol of each dNTP, 1.75 mM $MgCl_2$, 2.6 units of enzyme mix, and were carried out in a final volume of 50 μl. PCR conditions were: 1 cycle at 94°C for 2 min; 30 cycles at 94°C for 10 s, 65°C for 30 s and 68°C for 2 min; 1 cycle at 68°C for 7 min and a final hold at 4°C. PCR amplification was checked by agarose gel electrophoresis (see Section 3.1.3.1).

Sequencing reactions were performed using the Applied Biosystems strategy for automated sequencing (see 3.1.3.2). 4 μl of purified PCR product were used as template; the M13F primer was used as sequencing primer. DNA sequences were analysed with the MT Navigator PCC program (see 3.1.3.3).

3.3.3 SNP typing by PCR-SSP

SNP typing of patients and controls was carried out using an PCR-SSP typing assay. PCR-SSP is a highly sensitive method that allows detection of one allele in the presence of an excess of another allele, which is very important as an individual can have up to six C4 genes of which only one might have the SNP.

PCR amplification was carried out in a final volume of 20 µl. Reaction mixes contained 100 ng genomic DNA, 12.5 pmol of each C4E17AF and C4E17R primers (Table 9), 7.5 pmol of each HGH I and II primers (Table 9), 2.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.3, 50 mM KCl and 1 unit of Platinum Taq DNA polymerase. Amplification was performed in an Applied Biosystems GeneAmp PCR System 9700. The reaction mix was denatured at 96°C for 6 min, followed by 8 cycles at 96°C for 30 s, 70°C for 30 s, 72°C for 2 min, 10 cycles at 96°C for 30 s, 65°C for 30 s, 72°C for 2 min, 17 cycles at 96°C for 30 s, 70°C for 30 s, 72°C for 2 min, an amplification cycle at 72°C for 10 min and a final hold at 4°C.

PCR products (20 µl) were mixed with 3 µl of loading buffer type IV and resolved on a 1% agarose gel with ethidium bromide. 5 µl of lambda plus DNA ladder were loaded in every gel as size standard. The band at about 420 bp indicated adequate PCR amplification. The presence of the A allele (first base of codon 695 of the C4B3 gene) was indicated by a second band at about 200 bp.

The typing assay was developed by myself. SNP typing of IDDM patients and controls was carried out by Lydia Windsor.

Primer ID	Sequence	Fragment size
C4E17AF	GAC ACG TCT GCC CAT GAT GAG	
C4E17R	ACC GTT CTG CCT TTC CAA G	205 bp
HGH I	CAG TGC CTT CCC AAC CAT TCC CTT A	
HGH II	ATC CAC TCA CGG ATT TCT GTT GTG TTT C	414 bp

Table 9: Oligonucleotide primers used for PCR-SSP typing of a 62.1 haplospecific SNP. The C4E17AF primer binds specifically to the A allele in exon 17 (first base of codon 695) of the C4B3 gene. The HGH primers were used as PCR controls.

4 Characterisation of polymorphisms within 3.3 kb of sequence

4.1 Introduction

Complement component C4 is one of the most polymorphic proteins of the complement system. There are two isotypic forms of C4, C4A and C4B, which show differential chemical reactivities and binding affinities towards target surfaces. Four amino acid residues encoded within exon 26 have been shown to account for the isotypic differences of C4A and C4B. Further variation comes from the Rodger and Chido antigenic determinants encoded within exons 25, 26 and 28 of the C4 genes. In addition to this, typing of C4 based on differences in electrophoretic mobility and hemolytic activity revealed more than 40 C4 allotypes. However, there is limited DNA sequence information and the sequences which encode specific C4 allotypes have not been described. Serological typing of C4 proteins has shown that some common C4 allotypes can be split into several subtypes. Characterisation of C4 gene sequences might reveal an even greater level of polymorphisms, as has been observed in other studies of HLA gene polymorphisms.

The C4 genes are located in the central region of the MHC. As many MHC associated diseases have been mapped to the central region of the MHC, this region is of special interest. Previous studies have examined the role of several genes located within this region, but few studies have included C4, despite the high degree of C4 protein polymorphism. This may be in part related to the technical difficulties associated with C4 allotyping. Therefore, elucidation of the C4 variations at the DNA level would facilitate investigations into the role of C4 and other genes in the central region of the MHC with MHC associated diseases.

The aim of the present study was to systematically characterise the heterogeneity of the C4 gene sequences. Novel approaches were used for the detection of polymorphisms on a panel of extensively characterised standard homozygous DNAs. DNAs were chosen to represent various C4 allotypes from a range of different ancestral haplotypes (AHs), spanning all racial groups. Due to the modular variation of RCCX modules, each AH carries between one and three C4 genes. The C4 allotypes represented by the various AHs include six different C4A and four different C4B allotypes. A detailed description of DNAs used in this study is shown in Table 3.

Comparison of published DNA sequences from different C4 allotypes had identified the region spanning from intron 16 to intron 28 to be the most polymorphic. Therefore, a method was designed to characterise the polymorphisms present in this region. The approach covered 3.3 kb of sequence, including the isotype specific sites in exon 26 and the Rodgers and Chido antigenic determinants in exons 25 and 28.

4.2 Co-amplification of C4A and C4B genes

A novel approach, developed by David Sayer, was used to identify polymorphic sites within 3.3 kb of sequence. The approach is based on co-amplification of all C4 genes present on a haplotype and subsequent sequencing of PCR products. The co-amplification approach involves screening of C4A and C4B genes for polymorphisms as the separation of the C4 genes is difficult due to the high sequence homology of C4A and C4B. DNA sequencing was used as polymorphism screening assay because it also allows quantitative analysis of mutations. PCR amplification produced 6 overlapping amplicons, spanning the region from intron 16 to intron 28 (see Figure 7). PCR products were sequenced forward and reverse with BDT, and DNA sequences were analysed for the presence of nucleotide polymorphisms. Using this approach, 22 DNAs have been characterised. In the present study, two additional DNAs from 55.1 and 62.2 AHs were included.

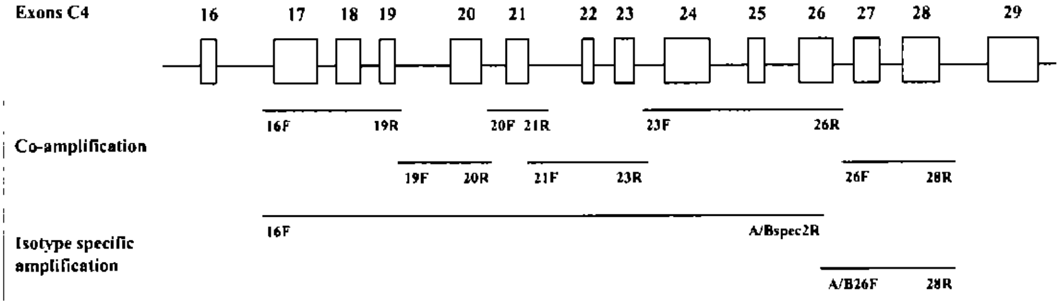

Figure 7: PCR amplification of 3.3 kb of sequence. The region spanning from intron 16 to intron 28 was screened for polymorphic positions using a co-amplification approach. Polymorphisms were assigned to either C4 isotype by C4A and C4B specific amplification.

All sequences were compared to the C4B1 gene of the 8.1 AH. Table 12 represents sequence differences detected within the 3.3 kb of sequence. All together, 35 polymorphic sites have been identified, including 19 intronic and 16 exonic nucleotide differences. Exons 26 and 28, known to code for the isotype specific residues and Rodgers and Chido antigenic determinants, were found to be the most polymorphic exons. Outside the C4d region (which spans from exon 23 to exon 30), 6 exonic differences have been identified (in exons 17, 19, 20 and 22). The nucleotide differences in exons 17 and 21 resulted in an amino acid change. The other three polymorphisms observed in the non-C4d region did not result in an amino acid change. Intronic sequence differences included 17 nucleotide substitutions and two 1-bp deletions (relative to the 8.1 C4B1). The highest degree of intronic polymorphism was seen in introns 19 (7 polymorphic sites) and 21 (4 polymorphic sites.) Other intronic differences were located in introns 17, 20, 23 and 28. Interestingly, no polymorphism was found in any of the introns spanning the polymorphic exons 25 to 28, which encode the isotypic and Rg/Ch specific residues. However, this correlates with initial findings by Ulgiati et al. [104], indicating that these introns show a very low degree of polymorphism.

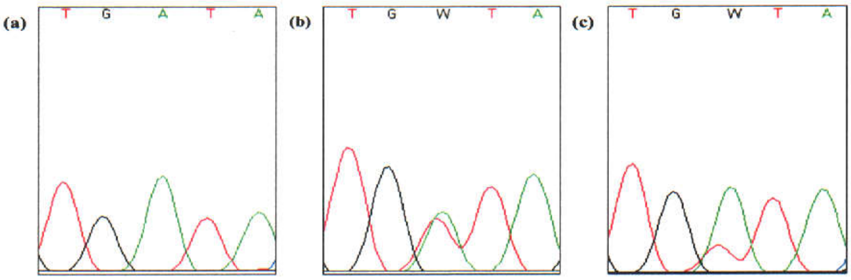

Figure 8: Electropherogram of composite DNA sequences. The PCR approach involves co-amplification of C4A and C4B genes, and therefore DNA sequences are composite sequences. DNA sequences of AHs carrying different C4 copy numbers are shown: (a) 8.1 (one C4 gene: C4B1), (b) 46.2 (two C4 genes: C4A4, C4B2) and (c) 65.1 (three C4 genes: C4A2, C4B1, C4B2).

Sequences were further analysed to confirm that the number of C4 genes expected from the description of the AHs correlates with the nucleotide pattern observed in the sequence electropherograms. The PCR approach applied here, included co-amplification of all C4 genes present on a haplotype. Therefore, all sequences shown in Table 12 were composite sequences. Some of the polymorphisms detected within the 3.3 kb of sequence were present on all C4 genes of a haplotype, whereas others were only found on part of the C4 genes. Analysis of relative quantities of nucleotide peaks in sequence electropherograms allowed estimation of the number of genes that showed the polymorphism. An example is given in Figure 8. At the polymorphic site shown, AHs 46.2 and 65.1 had an A to T substitution on part of their C4 genes. While the A and T nucleotide peaks of 46.2 were identical in height, the T nucleotide peak of 65.1 was reduced to about half the intensity of the A nucleotide peak. The 46.2 AH carried two C4 genes, C4A4 and C4B1, one of which had a T nucleotide at the polymorphic positions shown in Figure 8, whereas the other gene had an A nucleotide. In contrast, the 65.1 AH had three C4 genes (C4A2, C4B1 and C4B2). Analysis of the relative nucleotide quantities of this haplotype showed a 2:1 ratio for the A to T substitution, suggesting that the A nucleotide was present on two of the C4 genes and the T nucleotide on the third C4 gene. Thus, quantitative sequences reflected the C4 copy number of both haplotypes. A similar pattern of relative nucleotide quantities was found at most other polymorphic sites of 46.2 and 65.1. Sequence analysis of all other AHs with two C4 genes showed comparable intensities of nucleotide peaks at heterozygous positions. In addition to 65.1, there was one other AH with three C4 genes. The 7.1 AH had C4A3, C4A3 and C4B1. Quantitative sequence analysis of this haplotype showed a 2:1 ratio of nucleotide intensities at heterozygous positions, again reflecting the C4 copy number of the haplotype. No heterozygous positions were found on AHs predicted to have only one C4 gene. These AHs had either C4A specific sequences (i.e. encoding Pro^{1101}, Cys^{1102}, Leu^{1105}, Asp^{1106}) or C4B specific sequences (i.e. encoding Leu^{1101}, Ser^{1102}, Ile^{1105}, His^{1106}) at the isotypic site in exon 26, suggesting that the C4AQ0 and C4BQ0 alleles of these haplotypes were due to gene deletions rather than the presence of pseudogenes. Thus, quantitative sequences reflected the C4 copy numbers expected from the description of the AHs, suggesting that no additional genes (i.e. genes not included in the description of the AHs) were present in our samples.

In summary, this study confirmed the presence of the isotype specific sequences in exon 26 of the C4 genes. The polymorphic sites within 3.3 kb of sequence were identified, with most polymorphisms being located in the introns. However, the polymorphisms could not be correlated with the different C4 allotypes and hence, additional studies were required to identify the allotype specific sequences.

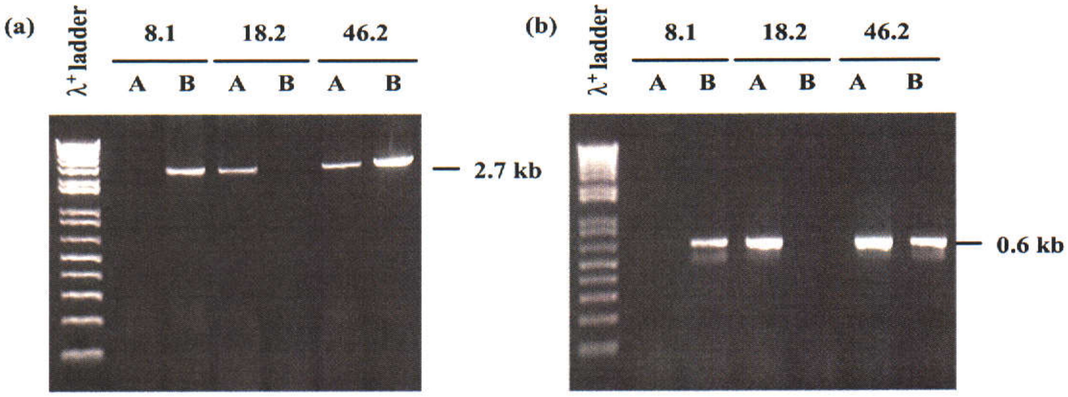

Figure 9: Isotype specific amplification of C4 genes. PCR amplification (a) from intron 16 to the isotypic site in exon 26, and (b) from the isotypic site in exon 26 to intron 28. Amplification with C4A specific primers is shown in lanes 2, 4 and 6, C4B specific primers were used in lanes 3, 5 and 7. Specificity is demonstrated using three different DNAs: 8.1 AH (C4AQ0, C4B1), 18.2 AH (C4A3, C4BQ0) and 46.2 AH (C4A4, C4B2).

4.3 Separation of C4A and C4B genes

The polymorphisms identified within the 3.3 kb of sequence were further analysed by isotype specific amplification of C4. Two PCR methods were designed to separate C4A and C4B genes, the first spanning from intron 16 to the isotype specific region in exon 26 and the second spanning from exon 26 to intron 28 (see Figure 7). The specificity of the primers was tested using DNA from 8.1, 18.2 and 46.2 AHs. The 8.1 AH carries a C4B1 gene and has a deletion at the C4A locus, 18.2 has a C4A3 gene, the C4B gene is deleted, and 46.2 carries both C4 isotypes (C4A4 and C4B2). The results of the isotype specificity are presented in Figure 9 and demonstrate that specific primers only amplified in the presence of the respective C4A and C4B isotypic sequence on the haplotype. By sequencing the specific PCR products, the previously identified polymorphisms could be assigned to either C4 isotype. As demonstrated in Figure 10, composite sequences could be split into C4A and C4B specific sequences. However, some AHs carried two C4A or two C4B genes, and hence the genes could not simply be separated by isotype specific amplification. Therefore, PCR products of these haplotypes were cloned to separate the genes. A comparison of all separated sequences is presented in Table 13.

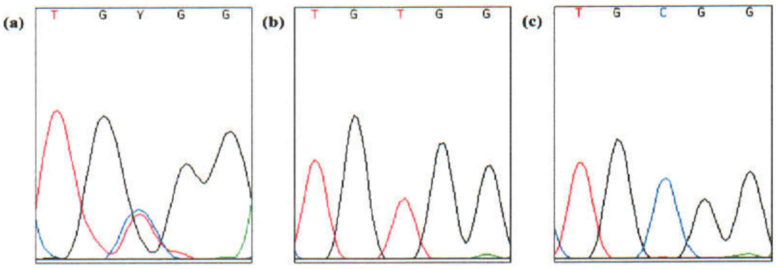

Figure 10: Electropherogram of isotype specific DNA sequences. Polymorphic positions were identified by co-amplification (a) of C4 genes. C4A specific (b) and C4B specific (c) amplification was used to separate sequences at polymorphic sites.

4.4 Rodgers and Chido antigenic determinants

The sequences of the C4A and C4B genes were analysed for the presence of Rodgers and Chido determinants. Rodgers and Chido antigenic determinants can be detected on red blood cells and result from the deposition of C4d fragments on the cell surface. All together, two Rodgers, six Chido and one rare determinant called WH have been described. The Rg/Ch determinants are encoded by four polymorphic sites located in exons 25, 26 and 28 (illustrated in Figure 5). In general, Rodger determinants are associated with C4A and Chido determinants with C4B. A list of the Rg/Ch determinants identified by sequence analysis of all C4 genes included in this study is shown in Table 10.

Both Rodgers and all six Chido determinants could be identified in our samples. The low-frequency allele WH was present on four C4A3 genes of 18.2, 44.1, 52.1 and 58.1 AHs, which is consistent with previous findings of C4A3 WH$^+$ phenotypes [98]. In addition, the WH antigen has been reported on C4A2, C4A5 and C4B5 allotypes [98, 100]. However, none of the four C4A2 genes included in our study had Asn^{1157}, Val^{1188} and Leu^{1191}, which in combination encode the WH determinant. Also, the two C4B5 genes of the 54.1 and 55.1 AHs were WH$^-$. Analysis of the C4B genes showed that most C4B allotypes were Rg$^-$ and Ch$^+$, as would be expected for C4B. Rodger determinants could only be

	Rg/Ch specific amino acids								Antigenic determinants			
Allele	1054	1101	1102	1105	1106	1157	1188	1191	Rg	Ch	WH	Ancestral haplotype
C4A2	D	P	C	L	D	N	V	L	1, 2, (3)	-	-	35.2
C4A2	G	P	C	L	D	N	V	L	1, 2, (3)	5	-	38.1, 52.1, 65.1
C4A3	D	P	C	L	D	N	V	L	1, 2, (3)	-	-	7.1, 7.2, 13.1, 35.2, 44.1, 44.2, 44 4, 54 1, 62 1
C4A3	D	P	C	L	D	S	V	L	1	6	WH	18.2, 44.1, 52.1, 58.1
C4A4	D	P	C	L	D	N	V	L	1, 2, (3)	-	-	18.1, 46 1, 46 2, 55.1, 62.2
C4A6	D	P	C	L	D	N	V	L	1, 2, (3)	-	-	57 1
C4A12	D	P	C	L	D	S	A	R	-	1, 3, 6	-	42 1
C4A91	G	P	C	L	D	S	A	R	-	1, 3, 5, 6	-	42.1, 47.1
C4B1	G	L	S	I	H	S	A	R	-	1, 2, 3, 4, 5, 6	-	7 1, 7.2, 8.1, 38.1, 44.2 44.3, 44.4
C4B1	G	L	S	I	H	N	A	R	(3)	1, 2, 4, 5	-	13.1, 57.1, 65 1
C4B2	D	L	S	I	H	S	A	R	-	1, 3, 4, 6	-	18 1, 46.1, 46.2, 62.2, 65 1
C4B3	G	L	S	I	H	S	A	R	-	1, 2, 3, 4, 5, 6	-	62 1
C4B5	D	L	S	I	H	S	A	R	-	1, 3, 4, 6	-	54 1, 55.1

Table 10: Rodgers and Chido antigenic determinants for various C4 allotypes. C4 allotypes are shown on the left, the AHs carrying these allotypes are indicated on the right The Rodgers 3 determinant shown in parenthesis is a hypothetical allele and is not detectable directly by serology.

detected on three C4B1 genes (AHs 13.1, 57.1 and 65.1), which had Asn^{1157} corresponding to the hypothetical Rg 3 determinant. In contrast, reversed antigenicity was found for two C4A allotypes, C4A12 and C4A91. Both allotypes were Rg^{-}, but expressed Ch 1, 3, 6 (C4A12) and Ch 1, 3, 5, 6 (C4A91), respectively. It is notable that C4A12 and C4A91 share the isotype specific residues Pro^{1101}, Cys^{1102}, Leu^{1105}, Asp^{1106} with other C4A allotypes, but resemble C4B allotypes with respect to their electrophoretic mobility. All other C4A allotypes were Rg^{+}, and most were Ch^{-}. Chido determinants were found on three C4A2 genes, which had Gly^{1054} characteristic for the Ch 5 determinant. Ch 6 was found on the four WH^{+} C4A3 genes.

As shown in Table 10, some C4 allotypes could be split into several subtypes based on the presence of different Rg and Ch antigenic determinants. For example, C4A2 on the 35.2 AH carried Rg 1, 2 and 3, whereas C4A2 on 38.1, 52.1 and 65.1 AHs expressed Ch 5 in addition to Rg 1, 2 and 3. Similarly, C4A3 could be divided into two subtypes as ten of the C4A3 genes had Rg 1, 2 and 3, and the other four C4A3 genes carried Rg 1, Ch 6 and the WH determinant. Of the C4B allotypes, C4B1 could be split due to the presence of two different Rg/Ch combinations. These findings are consistent with previous reports by Blanchong et al. [83] suggesting that some C4 allotypes are not homogeneous and may be split into several subtypes by serological typing. Roos et al. [95] found, that C4B5 could be subdivided into C4B5 Rg^{+} and C4B5 Rg^{-} by serological Rg and Ch typing. However, the C4B5 genes included in our study were both Rg^{-}. The C4B5 Rg^{+} haplotypes identified by Roos et al. had HLA-B60, C4A4, C4B5, DR4, whereas our C4B5 Rg^{-} genes were present on the 54.1 AH (HLA-B54, C4A3, C4B5, DR4) and 55.1 AH (HLA-B55, C4A4, C4B5, DR14). This is consistent with the C4B5 Rg^{-}

haplotypes identified by Roos et al. which had HLA-B35, C4A4, C4B5, DR4; HLA-B47, C4A4, C4B5, DR5 or HLA-B55, C4A4, C4B5, DR4. Aside from C4B5, it has been reported that three other common allotypes, C4A3, C4B1 and C4B3, may be split into subtypes [83]. These finding could be confirmed for C4A3 and C4B1, as Rg and Ch analysis of our samples revealed two C4A3 and two C4B1 subtypes. In addition to this, the present study revealed that C4A2 genes are not homogeneous and may be split into two subtypes on the basis of different Rg and Ch determinants.

4.5 Comparison of allotypic sequences

Sequences of all individual C4 allotypes were aligned to further investigate the presence of sequences which encode the allotypic variation seen by protein electrophoresis. However, apart from the isotype specific residues and the residues encoding the Rodger and Chido determinants, there were only three additional nucleotide substitutions present within the 3.3 kb of sequence that resulted in an amino acid change. Of the three amino acid changes, one was found on the C4B3 gene of the 62.1 AH (arginine to serine substitution at residue 695). This change was not found on any other C4 gene, suggesting that it might be specific for the C4B3 allotype which is only present on the 62.1 AH. Therefore, this arginine to serine substitution might account for the distinct electrophoretic mobility of the C4B3 allotype. Another amino acid change was identified at residue 710 of the 55.1 C4B5 gene (arginine to tryptophan substitution). However, this change was not present on the C4B5 gene of the 54.1 AH, and hence, does not explain the electrophoretic mobility of the C4B5 allotype. The amino acid change in exon 21 (residue 888) resulted in a threonine residue being replaced with an alanine. The threonine residue was present on the 7.1 C4B1 gene, the 8.1 C4B1 gene and on the C4A2 or C4A3 gene of the 52.1 AH. As the threonine to alanine substitution was only found on two of the ten C4B1 genes included in the study, and in addition was also present on a C4A2 or C4A3 gene, this change does not explain allotypic separation of C4.

To examine whether intronic sequences reflect the allotypic separation of C4 proteins, DNA sequences of all C4 genes included in the analysis were sorted according to their C4 allotype. A list of C4B1 and C4A3 intronic sequences is shown in Table 11. Comparison of nine C4B1 genes revealed eight unique intronic sequences. Apart from the two sequences of the 13.1 and 57.1 AH, all other C4B1 sequences were found to be unique. Similarly, comparison of C4A3 intronic sequences revealed that the C4A3 allotype, as defined by serology, can be split at the sequence level. Comparison of thirteen C4A3 sequences from different AHs identified seven unique combinations of nucleotides at poly-

MHC						I17	I19	I19	I19	I19	I19	I19	I19	I20	I20	I21	I21	I21	I21	I23	I23	I24	I28	I28
C2	Bf	C4A	C4B	AH	Allele	5309	5656	5743	5763	5776	5842	5862	5869	6031	6093	6381	6440	6528	6567	6880	6884	7205	8249	8253
C	S	Q0	1	8 1	C4B1	C	T	G	A	G	G	A	T	G	G	C	A	T	C	G	C	G	G	-
C	S	3	1	7.1	C4B1	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T			
C	F	3	1	44.4	C4B1	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
C	S	6	1	57 1	C4B1	.	.				.	.	.	.	.	.	G	C	.	.	.	.	.	.
C	S	3	1	13 1	C4B1			.			.	.	.	.	.	.	G	C	.	.	.	.	.	.
	S	2	1	38 1	C4B1		.	.			A					.	G	C	.	.	.	.	.	.
	S	Q0	1	44 3	C4B1		C	.								T	G	C	.	.	.	.	C	C
C	F	3	1	44 2	C4B1	.	C	.	G	A		G	C		A		G	C	.		.	.	C	C
C	S	3+3	1	7.2	C4B1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	C	C
C	F1	3	Q0	18 2	C4A3	.	.	.	.	.	.	.	.	.	.		G	C				.	C	C
	S	3	Q0	58.1	C4A3	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.		.	C	C
C	F	3	1	44 2	C4A3		C	.		.			.			T	G	C	.	.	.	.	C	C
C	S	3	1	13.1	C4A3	.	.	.	.	.	.	.	.	.	.	.	G				.	.	C	C
C	S	3+3	Q0	44 1	C4A3_2	.		.				.	.	A	.	.	G	.	.	.	.	.	C	C
C	S	3	5	54.1	C4A3	.	.	.		.			.	A	.	.	G	.	.	.	.	.	C	C
C	S	3	1	7.1	C4A3	.	.	.	.	.	.	.	.	A	.	.		C	.		T	.	C	C
C	S	3+3	Q0	44.1	C4A3_1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.		.	C	C
C	F	2+3	Q0	35.2	C4A3	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	C	C
C	F	3	1	44.4	C4A3	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	C	C
C	S	3+3	1	7.2	C4A3_1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	C	C
C	S	3+3	1	7 2	C4A3_2	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.			C	C
C	S	3	3	62 1	C4A3	.	.	.	.	.	A	.		A	.	.	.	.	.	.	.	.	C	C

Table 11: Comparison of C4A3 and C4B1 sequences from different AHs. Only intronic polymorphisms are shown. Allotypes defined by serology can be split at the sequence level. There are at least 8 unique C4B1 and 7 unique C4A3 sequences.

morphic sites. It is notable, that some C4B1 genes are identical to C4A3 genes within the introns. For example, the C4B1 gene ot the 44.3 AH shares the intronic sequence with the C4A3 gene of the 44.2 AH. The intronic sequence of the 7.1 C4B1 gene is identical to the intronic sequences of the 7.2, 35.2, 44.1 and 44.4 C4A3 genes. In addition, the 13.1 and 57.1 C4B1 genes and the 18.2 and 58.1 C4A3 genes had identical intronic sequences. Sequences of all other C4 allotypes are listed in Table 14 and demonstrate that most C4 allotypes are not homogeneous at the DNA sequence level and may be split into subtypes. For example, sequence analysis of the C4B2 genes revealed two distinct C4B2 sequences. In addition, it was found that C4A2 genes are not homogeneous as DNA sequences revealed at least three unique combinations of alleles. Of the four C4A4 genes included in the study, one had a unique polymorphism that was not found on any of the other three C4A4 genes. Therefore, C4A4 genes could be split at the DNA level. Moreover, analysis of the C4A91 genes of the 42.1 and 47.1 AHs showed that both genes differed at at least seven polymorphic sites within 3.3 kb of sequence. Comparison of the C4B5 gene of the 54.1 AH and the C4B5 gene of the 55.1 AH revealed two unique sequences, indicating that C4B5 genes are not homogeneous.

In summary, characterisation of 3.3 kb of the C4 gene sequence revealed that C4 allotypes defined by serology can be split at the sequence level. In addition to the subtypes identified by analysis of the Rodgers and Chido specific residues, comparison of DNA sequences revealed an even greater level of polymorphism. For some C4 allotypes, as many as seven or eight unique DNA sequences were identified. It is notable, that all C4 allotypes that were included with at least two examples from different AHs, could be split into subtypes based on the variation seen at the sequence level.

4.6 Phylogenetic analysis

The polymorphic positions identified within the 3.3 kb of sequences were used to construct a phylogenetic tree. The isotype specific sequences encoding residues 1101, 1102, 1105 and 1106 were excluded from the analysis as they represent differences in functional activity between both C4 isotypes. Other polymorphic sites within the region included 12 exonic and 19 intronic sites. The phylogenetic tree generated using these polymorphic positions is presented in Figure 11 (showing distribution of both C4 isotypes) and Figure 12 (showing distribution of different C4 allotypes).

The phylogenetic tree shows two separate clusters (Figure 11). The first group contains only C4A genes and the second contains mostly C4B genes. Therefore, the tree represents the separation of C4 into C4A and C4B, although the isotypic residues determining the differences in functional activity between both C4 isotypes were not included. However, there are a number of additional sites that seem to contribute to the isotypic separation of C4, for example the sequences encoding the Rodgers and Chido antigenic determinants. In general, Rodger determinants are associated with C4A and Chido determinants with C4B, although reversed antigenicity has been observed for various C4 allotypes (see Section 4.4). In addition, a polymorphic motif in intron 28 of the C4 genes has been shown to consist of two cytosines (C----C) in C4A alleles, whereas C4B alleles have a guanine at the first position and a deletion at the second (G----D) [102]. The C----C motif is not entirely restricted to C4A alleles as several C4B genes included in the present analysis were found to have C----C. Moreover, two of the C4A genes included in this study had G----D. However, in general most C4A genes were characterised by the presence of C----C and most C4B genes had G----D, indicating that the polymorphic motif contributes to grouping of C4A and C4B genes into two separate clusters. This is further supported by the phylogenetic tree shown in Figure 22 (Appendix B.2) which was constructed using non-coding polymorphisms only. Even though this tree does not show two separate clusters representing both C4 isotypes, most C4A alleles are found together with other C4A alleles and most C4B alleles are found with other C4B alleles. However, the C4A12 and C4A91 alleles of the 42.1 AH, which carry the C4B specific motif (G----D), are found together with C4B alleles, whereas the C4B1 genes of the 7.2, 44.2 and

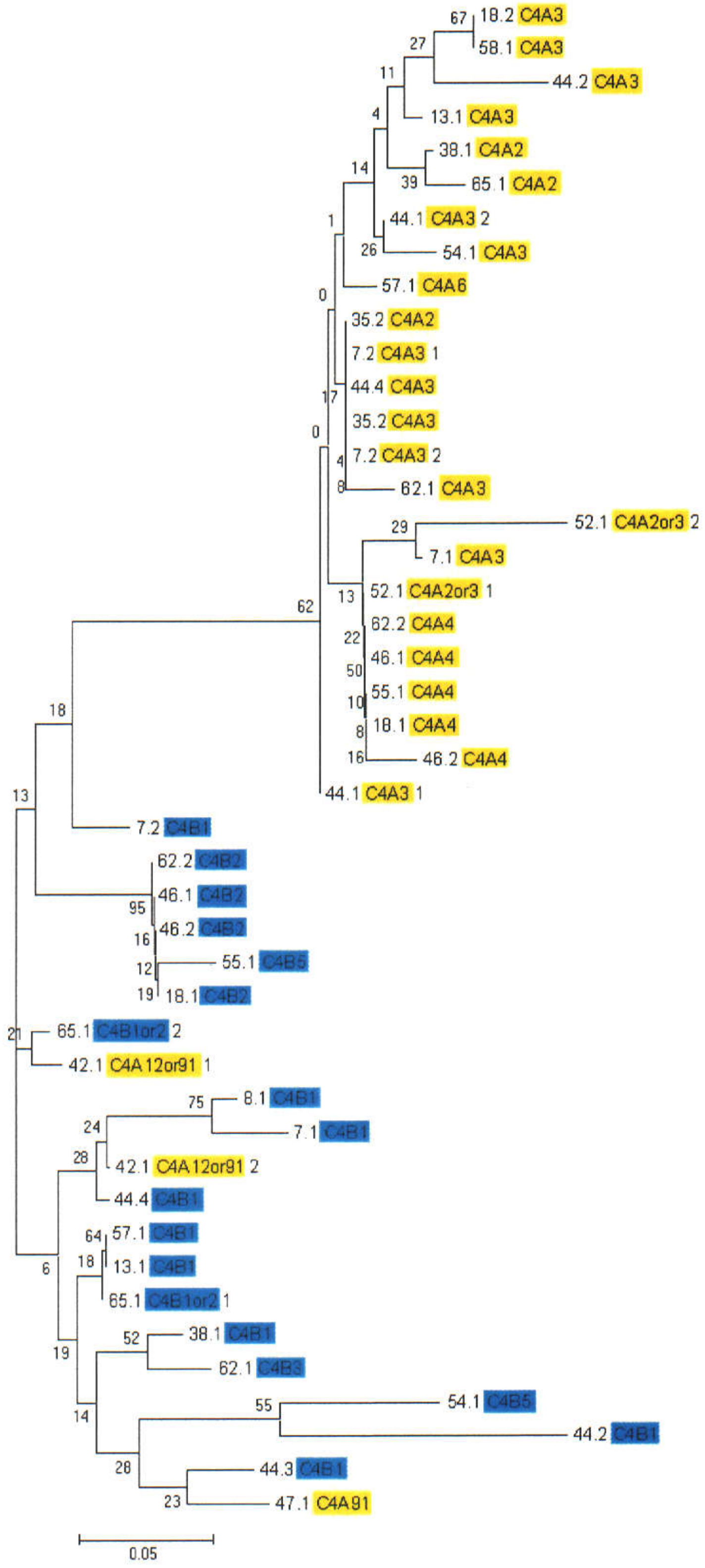

Figure 11: Phylogenetic sequence analysis. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Exonic and intronic polymorphic positions within 3.3 kb of sequence (spanning from intron 16 to 28) were used to construct the tree. The isotype specific sequences were excluded from the analysis as they represent differences in functional activity between both C4 isotypes. The bootstrap percentage from 1,000 replicates is indicated at each node. C4A allotypes are marked by a yellow box; C4B allotypes are highlighted by a blue box.

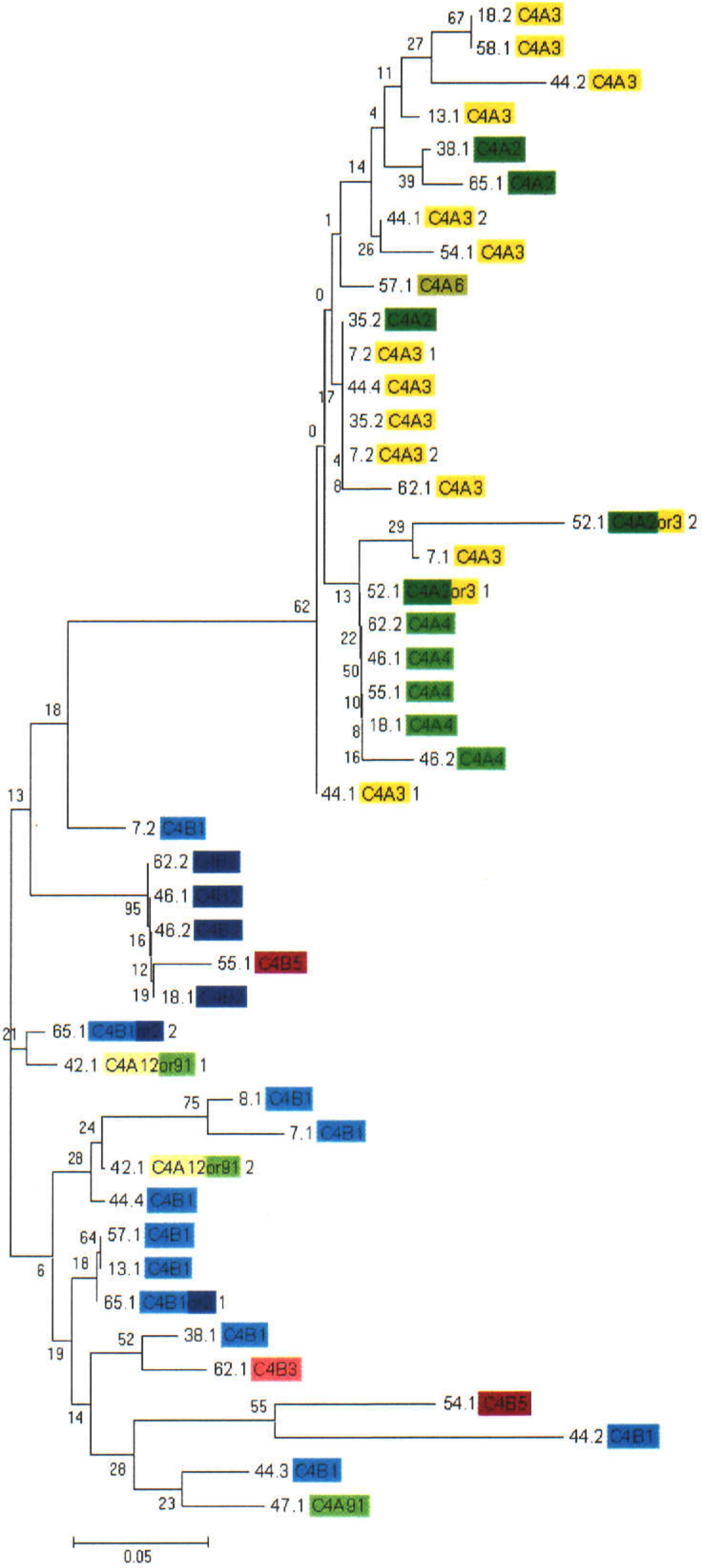

Figure 12: Phylogenetic sequence analysis. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Exonic and intronic polymorphic positions within 3.3 kb of sequence (spanning from intron 16 to 28) were used to construct the tree. The isotype specific sequences were excluded from the analysis as they represent differences in functional activity between both C4 isotypes. The bootstrap percentage from 1,000 replicates is indicated at each node. Each C4 allotype is marked by a different colour to show distribution of individual alleles within the tree.

However, it is notable, that the C4B1 genes of the 7.1 AH and the 8.1 AH haplotype cluster together as both haplotypes are associated with a number of autoimmune diseases. For example, the 7.1 AH is associated with susceptibility to multiples sclerosis (MS) and systemic lupus erythematosus (SLE). Individuals carrying the 8.1 AH have an increased susceptibility to insulin dependent diabetes mellitus (IDDM), SLE, IgA deficiency and several other autoimmune disorders. However, the 7.1 AH has been shown to be protective for diseases of which the 8.1 AH is susceptible (e.g. IDDM). All other C4 genes located within the second group of the phylogenetic tree did not show any significant subgrouping. C4B1 genes were found together with C4B3 and C4B5 genes, and also the C4A12 and C4A91 alleles contained within the second group seemed to be rather randomly distributed.

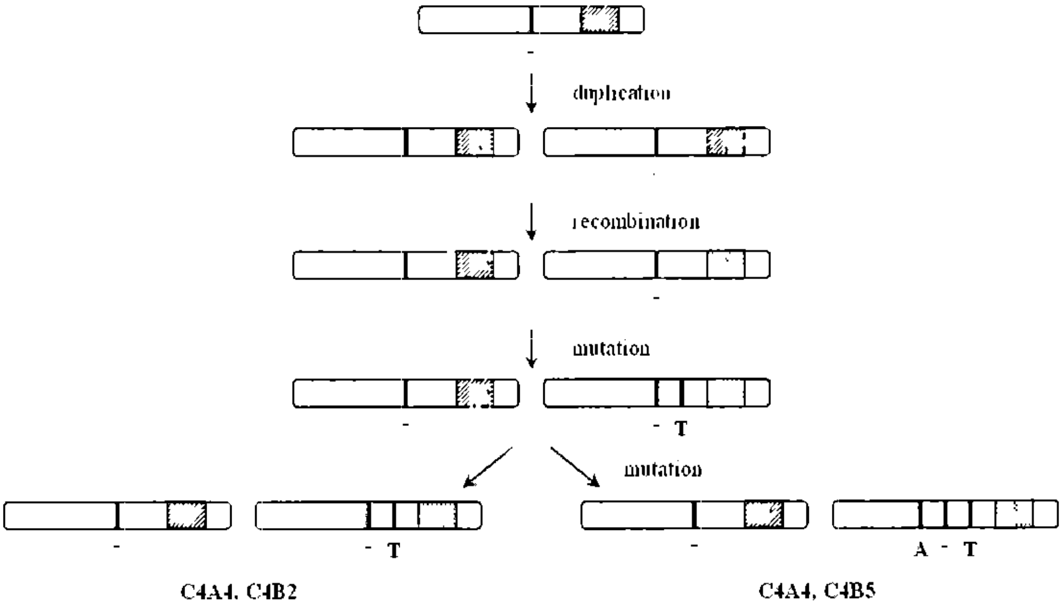

Figure 13: Proposed model of the evolution of the C4A4, C4B2 and C4B5 genes. The C4A4, C4B2 and C4B5 genes of the C4A4 containing haplotypes are characterised by the presence of a 1-bp deletion in intron 21 (indicated by a dash), which is unique to these genes. In addition, the C4B2 and C4B5 share a unique T nucleotide in intron 24. The C4B5 gene is characterised by a unique A nucleotide in exon 17. A model has been proposed to explain the evolution of these genes, suggesting that the C4A4, C4B2 and C4B5 genes have evolved from a common ancestor, which might have been C4A- or C4B-like (the isotype specific sequences are indicated by striped and dotted boxes)

It is of some interest that the C4A4, C4B2 and C4B5 genes of the C4A4 containing AHs (18.1, 46.1, 46.2, 55.1 and 62.2) share a unique 1-bp deletion in intron 21. This deletion is not found on the C4B2 and C4B5 genes of the 54.1 and 65.1 AHs. In addition, the C4B2 and C4B5 genes of the C4A4 containing AHs share a unique T nucleotide in intron 24, which is not present on other C4 genes. The C4B5 gene of the 55.1 AH is characterised by a unique A nucleotide in exon 17. The presence of the 1-bp deletion in intron 21 on all C4 genes of the C4A4 containing haplotypes suggests that these genes might have evolved from a common ancestor. A proposed model of the evolution of these genes is shown in Figure 13. The common ancestor might have been of the C4A or C4B isotype. Duplication and subsequent recombination would explain the presence of the 1-bp deletion in both C4A and C4B isotypes. Deletion, duplication and recombination events are frequently observed in the C4 gene region [114,115,175], and might therefore have also occurred during the evolution of these C4 genes. The recombination event must have comprised the isotype specific sequences, but might have also included the region from the isotypic site to the 5'-end of the C4 gene as the C4A4, C4B2 and C4B5 genes share at least the complete exonic sequence towards the 5'-end of the genes (see Table 20, Chapter 5). The unique T allele shared by the C4B2 and C4B5 genes is probably due to a single base substitution that occurred in the ancestral C4B gene but not in the C4A gene. The C4B genes must then have diverged to form the C4B2 and C4B5 allotypes. The only detected sequence difference between the C4B5 gene of the 55.1 AH and the C4B2 genes of the 18.1, 46.1, 46.2 and 62.2 AHs is a single base substitution in exon 17 that leads to an amino acid change. This amino acid change might account for the differences in electrophoretic mobility between both C4B allotypes.

It is notable that the 1-bp deletion is present on several haplotypes of a different ethnic origins. For example, the 18.1, 55.1 and 62.2 AHs are Caucasian haplotypes whereas the 46.1 and 46.2 AHs are predominantly found in Asia. Therefore, the changes that led to the generation of the C4A4, C4B2 and C4B5 genes must have occurred prior to the formation of these haplotypes, and hence must have occurred prior to freezing of the MHC blocks during primate evolution. The MHC is structurally organised in polymorphic frozen blocks (Figure 2, Section 2.1.3), within which recombination is rarely observed [34]. After the freezing, the blocks were shuffled to form ethnic specific haplotypes [15]. The ethnic specific haplotypes have been conserved over many generations and have therefore been called ancestral haplotypes [40,41]. They are characterised by a unique content of alleles and may carry specific polymorphisms (i.e. haplospecific polymorphisms). Haplospecific single nucleotide polymorphisms (SNPs) are also present in the C4 gene region of the 18.1, 46.1, 46.2, 55.1 and 62.2 AHs (see Table 22, Chapter 6).

In summary, the proposed model suggests that the C4A4, C4B2 and C4B5 genes have evolved from a common ancestor and have undergone several changes associated with duplication, recombination and mutation events prior to the formation of the conserved haplotypes.

4.7 Summary

The most polymorphic region of the C4 genes spanning from intron 16 to intron 28 was analysed using a panel of extensively characterised, homozygous workshop DNAs representing various C4 allotypes from a range of different ancestral haplotypes (AHs), spanning all racial groups. The region was screened for polymorphisms using a co-amplification approach. Polymorphisms were assigned to individual C4 genes by isotype specific PCR amplification.

The analysis revealed a number of interesting points:

1. Thirty-five sequence differences were identified within 3.3 kb of sequence, including 16 exonic and 19 intronic polymorphisms. Two of the intronic polymorphisms were 1-bp deletions, all other polymorphisms were single nucleotide substitutions.

2. The presence of the isotype specific sequences in exon 26 was confirmed. C4A genes encode Pro^{1101}, Cys^{1102}, Leu^{1105}, Asp^{1106}, whereas C4B genes encode Leu^{1101}, Ser^{1102}, Ile^{1105}, His^{1106}.

3. Introns 19 and 21 are highly polymorphic. In contrast, the introns spanning the polymorphic exons 25 to 28 were found to be conserved.

4. Quantitative sequences of the co-amplification approach reflect C4 copy numbers present on a haplotype. Sequence analysis revealed no additional C4 genes not included in the description of the AHs. C4 null alleles present on some AHs are likely to be due to gene deletion.

5. Rodger antigens are in general associated with C4A alleles, whereas Chido determinants are found on C4B alleles. However, some C4 allotypes show reversed antigenicity.

6. C4 allotypes defined by serology can be split according to different Rg/Ch combinations of some allotypes.

7. Characterisation of DNA sequences revealed an even greater level of polymorphism than has been observed at the protein level. Nucleotide polymorphisms do not correlate with serology. All C4 allotypes included with at least two examples from different AHs could be split at the sequence level.

8. Phylogenetic analysis demonstrates that C4 genes form two distinct clusters representing the separation of C4 isotypes, even if the isotypic sequences are excluded from the analysis.

9. Phylogenetic analysis also revealed that C4A4 genes are very similar. Interestingly, the C4A4 genes carry a unique 1-bp deletion in intron 21 that is not found on other C4A genes.

10. The C4B2 genes of the 18.2, 46.1, 46.2 and 62.2 AHs form a cluster with the C4B5 gene of the 54.1 AH within a phylogenetic tree. Sequence analysis showed that the five genes share two unique polymorphisms.

11. The C4B1 genes of the 7.1 and 8.1 AHs also formed a distinct subgroup within the phylogenetic tree. It is notable that both the 7.1 AH and 8.1 AH are associated with a number of autoimmune diseases.

12. C4A12 and C4A91 were found to differ from other C4A allotypes as both genes encode Chido specific residues instead of the Rodger antigens commonly found on C4A proteins. Both allotypes also resemble C4B allotypes with respected to their electrophoretic mobility. In addition, phylogenetic analysis revealed grouping of C4A12 and C4A91 with C4B alleles.

13. A model was proposed to explain the evolution of the C4A4, C4B2 and C4B5 genes found on the 18.1, 46.1, 46.2, 62.2 and 55.1 AHs.

MHC					E17	E17	I17	E19	I19	I19	I19	I19	I19	I19	I19	E20	I20	I20	E21	I21	I21	I21	I21	E22	I23	I23	I24	E25	E26	E26	E26	E26	E26	E28	E28	E28	E28	I28	I28
C2	Bf	C4A	C4B	AH	695	710	5309	786	5656	5743	5763	5776	5842	5862	5869	806	6031	6093	888	6381	6440	6528	6567	918	6880	6884	7205	1054	1076	1101	1102	1105	1106	1157	1186	1188	1191	8249	8253
C	S	Q0	1	8.1	CGT	CGG	C	ACG	T	G	A	G	G	A	T	GTT	G	G	ACC	C	A	T	C	GGG	G	C	G	GGC	GGC	CTC	TCT	ATA	CAT	AGC	GCC	GCG	CGG	G	—
	S	Q0	1	44.3	...	...	.	...	C	.	.	.	.	.	.	..C	.	.	G..	T	G	C	.	...	.	.	.	...	..A	...	...	...	...	...	..G	...	...	C	C
C	F1	3	Q0	18.2	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	G	C	.	...	.	.	.	.A.	...	.C.	.G.	T..	G.C	...	..G	.T.	.TC	C	C
	S	3	Q0	58.1	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	G	C	.	...	.	.	.	.A.	...	.C.	.G.	T..	G.C	...	..G	.T.	.TC	C	C
C	S	3+3	Q0	44.1	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	R	.	.	...	.	.	.	.A.	...	.C.	.G.	T..	G.C	.R.	..G	.T.	.TC	C	C
C	S	2+3	Q0	52.1	...	...	.	...	.	.	.	.	.	.	.	..Y	R	.	R..	.	.	C	.	...	.	Y	.	.R.	...	.C.	.G.	T..	G.C	.R.	..G	.T.	.TC	C	C
C	F	2+3	Q0	35.2	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	.	.	...	.	.	.	.A.	...	.C.	.G.	T..	G.C	.A.	..G	.T.	.TC	C	C
C	F	12+91	Q0	42.1	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	.	.	.	...	.	.	.	.R.	...	.C.	.G.	T..	G.C	...	...	...	...	.	.
C	F	91	Q0	47.1	...	...	.	..A	C	.	.	.	.	.	.	..C	.	.	G..	.	G	C	.	...	.	.	.	...	..A	.C.	.G.	T..	G.C	...	...	...	...	C	C
C	F	3	1	44.2	...	...	.	...	C	.	R	R	.	R	Y	..C	.	R	G..	Y	G	C	.	...	.	.	.	.R.	..M	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	C	C
C	F	3	1	44.4	...	...	.	...	.	.	.	.	.	.	.	..C	R	.	G..	.	.	Y	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	3	1	7.1	...	...	.	...	.	.	.	.	.	.	.	..Y	R	.	R..	.	.	C	.	...	.	T	.	.R.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	3	1	13.1	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	G	Y	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.A.	..S	.Y.	.KS	S	-/C
C	S	3+3	1	7.2	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	.	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	C	C
	S	2	1	38.1	...	...	.	...	.	.	.	.	R	.	.	..C	R	.	G..	.	G	Y	.	...	.	.	.	...	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	2	1+2	65.1	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	G	Y	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	6	1	57.1	...	...	.	...	.	.	.	.	.	.	.	..C	.	.	G..	.	R	Y	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.A.	..S	.Y.	.KS	S	-/C
C	S	3	3	62.1	M..	...	.	...	.	.	.	.	A	.	.	..C	R	.	G..	.	R	Y	.	...	.	.	.	.R.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	C	C
C	S	3	5	54.1	...	...	Y	...	.	R	.	R	.	R	Y	..C	R	.	G..	.	G	Y	.	..R	.	.	.	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
	S	4	5	55.1	...	.R.	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	C	—	...	.	.	K	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
Q0	S	4	2	18.1	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	C	—	...	.	.	K	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	4	2	46.1	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	C	—	...	.	.	K	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
C	S	4	2	46.2	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	C	—	...	R	.	K	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
	S	4	2	62.2	...	...	.	...	.	.	.	.	.	.	.	..C	A	.	G..	.	.	C	—	...	.	.	K	.A.	...	.Y.	.S.	W..	S.Y	.R.	..S	.Y.	.KS	S	-/C
Alternative					CGT	CGC		ACG								GTT			ACC					GGG				GGC	GGC	CTC	TCT	ATA	CAT	AGC	GCC	GCG	CGG		
codons					AGT	CAC		ACA								GTC			GCC					GGA				GAC	GGA	CCC	TGT	TTA	GAC	AAC	GCG	GTG	CTC		
Alternative aa					R / S	R / Q		T								V			T / A					G				G / D	G	L / P	S / C	I / L	H / D	S / N	A	A / V	R / L		

Table 12: Polymorphic positions within 3.3 kb of sequence. 24 different ancestral haplotypes (AHs) have been studied The composite DNA sequence of all C4 genes present on a haplotype compared to the C4B1 gene of the 8.1 AH is shown (sequence differences only). The AH designation and the complotype (C2, Bf, C4A and C4B allotypes) of each AH are indicated on the left. Nucleotide positions are relative to the sequence of the C4B1 sequence published in GenBank (accession number U24578) [176]. For exonic polymorphisms, codon number, alternative codon and alternative amino acid are shown. IUPAC codes have been used to indicate the presence of two nucleotides (R=A+G, Y=C+T, K=G+T, M=A+C, S=G+C, W=A+T). A deletion relative to the 8.1 AH is marked by a dash ("—")

MHC						E17	E17	I17	E19	I19	I19	I19	I19	I19	I19	I19	E20	I20	I20	E21	I21	I21	I21	I21	E22	I23	I23	I24	E25	E26	E26	E26	E26	E26	E28	E28	E28	E28	I28	I28
C2	Bf	C4A	C4B	AH		695	710	5309	786	5656	5743	5763	5776	5842	5862	5869	806	6031	6093	888	6381	6440	6528	6567	918	6880	6884	7205	1054	1076	1101	1102	1105	1106	1157	1186	1188	1191	8249	8253
C	S	Q0	1	8.1	C4B1	CGT	CGG	C	ACG	T	G	A	G	G	A	T	GTT	G	G	ACC	C	A	T	C	GGG	G	C	G	GGC	GGC	CTC	TCT	ATA	CAT	AGC	GCC	GCG	CGG	G	—
C	S	3	1	7.1	C4B1																		C				T													
	S	4	5	55.1	C4B5		A										C	A		G			C	—				T	A											
Q0	S	4	2	18.1	C4B2												C	A		G			C	—				T	A											
C	S	4	2	46.1	C4B2												C	A		G			C	—				T	A											
C	S	4	2	46.2	C4B2												C	A		G			C	—				T	A											
	S	4	2	62.2	C4B2												C	A		G			C	—				T	A											
C	F	3	1	44.4	C4B1												C			G			C																	
C	S	2	1+2	65.1	C4B1/2_2												C			G		G							A						R					
C	S	3	5	54.1	C4B5			T			A		A		G	C	C			G		G	C						A											
C	S	6	1	57.1	C4B1												C			G		G	C												A					
C	S	2	1+2	65.1	C4B1/2_1												C			G		G	C												R					
C	S	3	1	13.1	C4B1												C			G		G	C												A					
	S	2	1	38.1	C4B1									A			C			G		G	C																	
C	S	3	3	62.1	C4B3	A								A			C			G		G	C																C	C
C	S	3+3	1	7.2	C4B1												C	A		G																			C	C
	S	Q0	1	44.3	C4B1					C							C			G	T	G	C							A						G			C	C
C	F	3	1	44.2	C4B1					C		G	A		G	C	C		A	G		G	C							A									C	C
C	F	12+91	Q0	42.1	C4A12/91_1												C			G									A		C	G	T	G C						
C	F	12+91	Q0	42.1	C4A12/91_2												C			G											C	G	T	G C						
C	F	91	Q0	47.1	C4A91				A	C							C			G		G	C							A	C	G	T	G C					C	C
C	F1	3	Q0	18.2	C4A3												C			G		G	C						A		C	G	T	G C		G	T	TC	C	C
	S	3	Q0	58.1	C4A3												C			G		G	C						A		C	G	T	G C		G	T	TC	C	C
C	F	3	1	44.2	C4A3					C							C			G	T	G	C						A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	Q0	44.1	C4A3_2												C	A		G		G							A		C	G	T	G C	R	G	T	TC	C	C
C	S	3	1	13.1	C4A3												C			G		G							A		C	G	T	G C	A	G	T	TC	C	C
	S	2	1	38.1	C4A2												C	A		G		G									C	G	T	G C	A	G	T	TC	C	C
C	S	2	1+2	65.1	C4A2												C			G		G									C	G	T	G C	A	G	T	TC	C	C
C	S	3	5	54.1	C4A3												C	A		G		G			A				A		C	G	T	G C	A	G	T	TC	C	C
C	S	2+3	Q0	52.1	C4A2/3_1												C	A		G			C						A		C	G	T	G C	R	G	T	TC	C	C
C	S	2+3	Q0	52.1	C4A2/3_2																		C				T				C	G	T	G C	R	G	T	TC	C	C
C	S	3	1	7.1	C4A3												C	A		G			C				T		A		C	G	T	G C	A	G	T	TC	C	C
Q0	S	4	2	18.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	4	2	46.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	4	2	46.2	C4A4												C	A		G			C	—		A			A		C	G	T	G C	A	G	T	TC	C	C
	S	4	2	62.2	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
	S	4	5	55.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	Q0	44.1	C4A3_1												C	A		G									A		C	G	T	G C	R	G	T	TC	C	C
C	F	2+3	Q0	35.2	C4A3												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	F	2+3	Q0	35.2	C4A2												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	F	3	1	44.4	C4A3												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	1	7.2	C4A3_1												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	1	7.2	C4A3_2												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3	3	62.1	C4A3									A			C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	6	1	57.1	C4A6												C			G									A		C	G	T	G C	A	G	T	TC	C	C
Alternative codons						CGT AGT	CGG CAG		ACG ACA								GTT GTC			ACC GCC					GGG GGA				GGC GAC	GGC GGA	CTC CCC	TCT TGT	ATA TTA	CAT GAC	AGC AAC	GCC GCG	GCG GTG	CGG CTC		
Alternative aa						R / S	R / Q		T								V			T / A					G				G / D	G	L / P	S / C	I / L	H / D	S / N	A	A / V	R / L		

Table 13: Isotype specific DNA sequences spanning from intron 16 to intron 18. C4A and C4B genes have been separated by isotype specific amplification. Sequence differences compared to the 8.1 C4B1 gene are indicated. The AH designation, C2, Bf, C4A and C4B allotypes of each AH and the allele of the sequenced C4 gene are indicated on the left. A deletion relative to 8.1 is marked by a dash ("—"). "R" indicates the presence of either A or G "/" was used in the description of the C4 allele to indicate that the C4 allotype could not be assigned to either gene. To distinguish two genes of the same allotype on a haplotype, the genes were numbered with "_1" and "_2", respectively.

MHC						E17	E17	I17	E19	I19	I19	I19	I19	I19	I19	I19	E20	I20	I20	E21	I21	I21	I21	I21	E22	I23	I23	I24	E25	E26	E26	E26	E26	E26	E28	E28	E28	E28	I28	I28
C2	Bf	C4A	C4B	AH		695	710	5309	786	5656	5743	5763	5776	5842	5862	5869	806	6031	6093	888	6381	6440	6528	6567	918	6880	6884	7205	1054	1076	1101	1102	1105	1106	1157	1186	1188	1191	8249	8253
C	S	Q0	1	8.1	C4B1	CGT	CGG	C	ACG	T	G	A	G	G	A	T	GTT	G	G	ACC	C	A	T	C	GGG	G	C	G	GGC	GGC	CTC	TCT	ATA	CAT	AGC	GCC	GCG	CGG	G	—
C	S	3	1	7.1	C4B1																		C				T													
C	S	3+3	1	7.2	C4B1												C	A		G																			C	C
C	S	3	1	13.1	C4B1												C			G		G	C												A					
	S	2	1	38.1	C4B1									A			C			G		G	C																	
C	F	3	1	44.2	C4B1					C		G	A		G	C	C		A	G		G	C							A									C	C
	S	Q0	1	44.3	C4B1					C							C			G	T	G	C							A						G			C	C
C	F	3	1	44.4	C4B1												C			G			C																	
C	S	6	1	57.1	C4B1												C			G		G	C												A					
C	S	2	1+2	65.1	C4B1/2_1												C			G		G	C												R					
Q0	S	4	2	18.1	C4B2												C	A		G			C	—				T	A											
C	S	4	2	46.1	C4B2												C	A		G			C	—				T	A											
C	S	4	2	46.2	C4B												C	A		G			C	—				T	A											
	S	4	2	62.2	C4B												C	A		G			C	—				T	A											
C	S	2	1+2	65.1	C4B1_												C			G		G							A						R					
C	S	3	3	62.1	C4B3	A								A			C			G		G	C																C	C
C	S	3	5	54.1	C4B			T			A		A		G	C	C			G		G	C						A											
	S	4	5	55.1	C4B5		A										C	A		G			C	—				T	A											
C	F	2+3	Q0	35.2	C4A2												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
	S	2	1	38.1	C4A2												C	A		G		G									C	G	T	G C	A	G	T	TC	C	C
C	S	2+3	Q0	52.1	C4A2/3_2																		C				T				C	G	T	G C	R	G	T	TC	C	C
C	S	2	1+2	65.1	C4A2												C			G		G									C	G	T	G C	A	G	T	TC	C	C
C	S	3	1	7.1	C4A3												C	A		G			C				T		A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	1	7.2	C4A3_1												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	1	7.2	C4A3_2												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3	1	13.1	C4A3												C			G		G							A		C	G	T	G C	A	G	T	TC	C	C
C	FI	3	Q0	18.2	C4A3												C			G		G	C						A		C	G	T	G C		G	T	TC	C	C
C	F	2+3	Q0	35.2	C4A3												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	3+3	Q0	44.1	C4A3_1												C	A		G									A		C	G	T	G C	R	G	T	TC	C	C
C	S	3+3	Q0	44.1	C4A3_2												C	A		G		G							A		C	G	T	G C	R	G	T	TC	C	C
C	F	3	1	44.2	C4A3					C							C			G	T	G	C						A		C	G	T	G C	A	G	T	TC	C	C
C	F	3	1	44.4	C4A3												C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
C	S	2+3	Q0	52.1	C4A2/3_1												C	A		G			C						A		C	G	T	G C	R	G	T	TC	C	C
	S	3	5	54.1	C4A3												C	A		G		G			A				A		C	G	T	G C	A	G	T	TC	C	C
	S	3	Q0	58.1	C4A3												C			G		G	C						A		C	G	T	G C		G	T	TC	C	C
C	S	3	3	62.1	C4A3									A			C	A		G									A		C	G	T	G C	A	G	T	TC	C	C
Q0	S	4	2	18.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	4	2	46.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	4	2	46.2	C4A4												C	A		G			C	—		A			A		C	G	T	G C	A	G	T	TC	C	C
	S	4	5	55.1	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
	S	4	2	62.2	C4A4												C	A		G			C	—					A		C	G	T	G C	A	G	T	TC	C	C
C	S	6	1	57.1	C4A6												C			G									A		C	G	T	G C	A	G	T	TC	C	C
C	F	12+91	Q0	42.1	C4A1/91_1												C			G									A		C	G	T	G C						
C	F	12+91	Q0	42.1	C4A1/91_2												C			G											C	G	T	G C						
C	F	91	Q0	47.1	C4A91				A	C							C			G		G	C							A	C	G	T	G C					C	C
Alternative codons						CGT	CGG		ACG								GTT			ACC					GGG				GGC	GGC	CTC	TCT	ATA	CAT	AGC	GCC	GCG	CGG		
						AGT	CAG		ACA								GTC			GCC					GGA				GAC	GGA	CCC	TGT	TTA	GAC	AAC	GCG	GTG	CTC		
Alternative aa						R / S	R / Q		T								V			T / A					G				G / D	G	L / P	S / C	I / L	H / D	S / N	A	A / V	R / L		

Table 14: Isotype specific DNA sequences spanning from intron 16 to intron 18. Sequences were sorted according to their C4 allotypes. Sequence differences compared to the 8.1 C4B1 gene are indicated. The AH designation, C2, Bf, C4A and C4B allotypes of each AH and the allele of the sequenced C4 gene are indicated on the left A deletion relative to 8.1 is marked by a dash ("—"). "R" indicates the presence of either A or G. "/" was used in the description of the C4 allele to indicate that the C4 allotype could not be assigned to either gene To distinguish two genes of the same allotype on a haplotype, the genes were numbered with "_1" and "_2", respectively

5 Identification of coding polymorphisms

5.1 Introduction

Complement component C4 is highly polymorphic at the protein level. C4 typing based on differences in electrophoretic mobility and hemolytic activity has revealed more than 40 C4 allotypes. The allotypes can be separated by agarose gel electrophoresis of neuraminidase and carboxypeptidase treated EDTA plasma, resulting in single bands defining each C4 allotype. However, current C4 typing methods are labour intensive and unambiguous assignment of C4 allotypes may be difficult as some allotypes show very similar migration rates in agarose gels. In addition, assignment of C4 null alleles is not always possible by electrophoresis based typing, and therefore may require family studies.

The sequences which encode specific C4 allotypes have not been described. Most of the sequences published to date comprise the most polymorphic region of the C4 genes, the C4d region. However, as characterisation of 3.3 kb of sequence has shown (Chapter 4), this region does not explain the allotypic variation of the C4 proteins, suggesting that additional polymorphic sites contributing to protein polymorphism might be encoded in other parts of the C4 genes. A limited number of complete C4A and C4B gene sequences is available at GenBank. However, these sequences only cover few different C4 allotypes, and therefore do not allow extensive analysis of the allotypic variation.

The aim of the present study was to systematically characterise the DNA polymorphisms accounting for the variation observed at the protein level. All exonic regions that were not included in the 3.3 kb of sequence described in Chapter 4, were screened for sequence variations by denaturing HPLC (dHPLC) and the variations were characterised by DNA sequencing. The DNAs used in this study were chosen from an international workshop panel, representing extensively characterised homozygous DNAs. A range of C4 allotypes from different AHs spanning all racial groups was included in this study, allowing comparative analysis of C4 gene sequences from various C4 allotypes.

5.2 Screening for polymorphisms by denaturing HPLC

Denaturing high performance liquid chromatography (dHPLC) is a method that can be used to screen PCR products for unknown single base substitutions and small insertions or deletions [177]. DNA fragments are separated on the HPLC by ion-paired reverse phase chromatography using TEAA (Triethyl ammonium acetate) as bridging molecule between the hydrophobic matrix of the column and the DNA. TEAA interacts with the negatively charged phosphate ions of the DNA, and therefore, the number of TEAA molecules bound to the DNA is directly proportional to the length of the DNA fragment. Due to the alkyl chains of the TEAA molecules, the DNA is absorbed to the hydrophobic matrix of the column. As more TEAA molecules can interact with longer DNA fragments, these DNA fragments have a higher affinity to the column. The DNA is eluted from the column with increasing amounts of acetonitrile in the mobile phase. Under non-denaturing conditions, DNA fragments are separated solely depending on the size of the fragment.

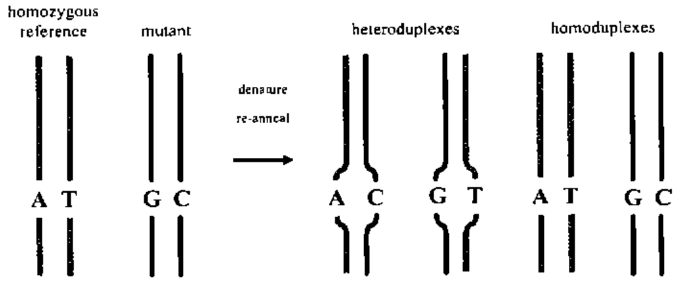

Figure 14: Schematic representation of heteroduplex formation for dHPLC analysis. Screening of PCR products for unknown polymorphisms requires mixing of samples with a known homozygous reference. Mixtures are denatured by heating and re-annealed by slow cooling to facilitate heteroduplex formation. HPLC analysis under partially denaturing conditions resolves heteroduplexes from homoduplexes, and therefore allows detection of sequence variations. Reproduced from Ref. [173]

Under partially denaturing conditions, however, DNA sequence variations can be detected. To screen PCR products for unknown polymorphisms, each sample needs to be mixed with a known homozygous reference. To facilitate heteroduplex formation between the reference DNA and the sample DNA, mixtures need to be denatured at 95°C followed by slow renaturation (Figure 14). If mixtures are analysed under partially denaturing conditions on the dHPLC, heteroduplexes melt at the mismatch site, whereas homoduplexes require higher temperatures to be melted. Partially single-stranded heteroduplexes have a lower affinity to the hydrophobic matrix of the column, which results in shorter retention times of heteroduplexes compared to homoduplexes. Thus, DNA fragments containing polymorphisms show a distinct peak pattern on the dHPLC and can therefore be distinguished from homozygous DNAs.

Figure 15 shows heteroduplex resolution in four different mixtures of PCR fragments and known homozygous reference compared to the peak pattern of the homozygous reference (shown at the top). Depending on the nature and the position of the polymorphism in the PCR fragment, different heteroduplex peak patterns were observed. However, to characterise the detected sequence variations, DNA sequencing of PCR products was required.

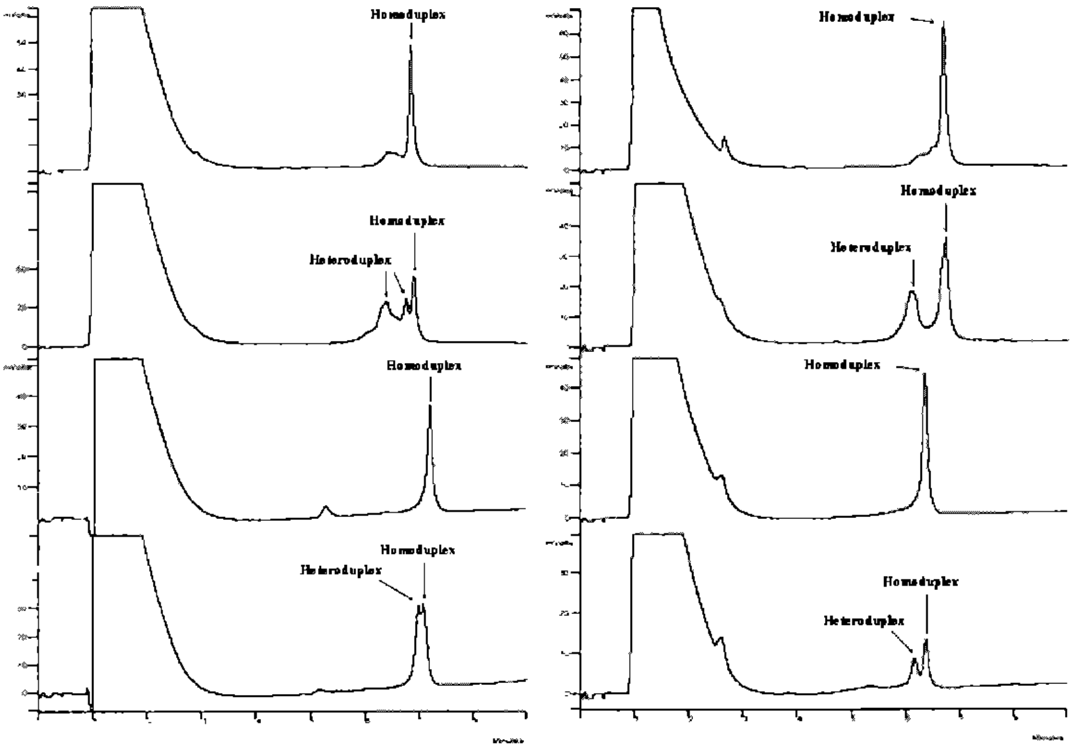

Figure 15: Screening for polymorphisms by denaturing HPLC. Under partially denaturing conditions, DNA sequence variations can be detected on the HPLC due to different retention times of heteroduplex and homoduplex DNAs. Homoduplex peaks and heteroduplex peaks are presented for four different fragments. Homoduplex peaks are shown at the top, respective heteroduplex peaks are shown below.

Top left C to T and C to G substitution at positions 99 and 377 of a 404 bp fragment; bottom left: A to G substitution at position 39 of a 375 bp fragment; top right T to C substitution at position 62 of a 302 bp fragment; bottom right: G to A substitution at position 204 of a 268 bp fragment.

5.2.1 Sample preparation and optimisation of method

To amplify the exonic regions not characterised in Section 4 (i.e. exons 1 to 16 and exons 29 to 41), 23 sets of conserved primers were designed spanning one or two exons each. Conserved regions within the polymorphic C4 genes were identified by comparison of complete C4 sequences available at GenBank. All primers were tested and PCR reactions optimised using the 8.1 AH. This haplotype was chosen as reference sample as it was well characterised, known to be homozygous and carried only one C4 gene. Size, yield and purity of PCR products was determined by non-denaturing analysis of amplicons on the HPLC. Unpurified PCR products were used as post PCR purification is not recommended [174]. The results of the non-denaturing analysis are shown in Figures 24 to 27 (Appendix C.1). All amplicons gave a sharp, single peak at 50°C, indicating that PCR amplification produced one specific product for all sets of primers. Fragment sizes were determined by comparison to the HaeIII digested pUC18 size marker (Figure 34, Appendix C.4) and were found to correlate with expected fragment sizes (listed in Table 7). The intensity of the peaks ranged between 30 mV and 100 mV, and was therefore sufficient for all amplicons. To confirm homozygosity of the 8.1 reference, all 23 amplicons of this sample were sequenced forward and reverse. Sequence analysis revealed no heterozygous positions within any of the 23 amplicons, and hence, the 8.1 AH could be used as known homozygous reference. PCR products of each sample were mixed with PCR products of the 8.1 reference, denatured at 95°C and re-annealed by slow cooling to facilitate heteroduplex formation.

5.2.2 Temperature selection

The temperature at which the PCR products are analysed on the dHPLC has to be carefully chosen as the sensitivity of the method is strongly dependent on the correct melting temperature. The melting temperature is affected by the location of the polymorphism within the fragment and also depends on the melting character of the surrounding nucleotides. Therefore, different polymorphisms within a DNA fragment may be detected at different temperatures as several melting domains may be present within a fragment. Temperatures for optimal resolution of heteroduplex and homoduplex peaks can be determined empirically by injecting each PCR amplicon at increasing temperatures until a significant decrease in the retention time is observed. However, this method should be used in combination with melting temperature prediction software.

In the present study, melting temperatures of each amplicon were determined with the DHPLCMelt software. For a given sequence, the DHPLCMelt software predicts the melting temperature and gradient conditions that will resolve heteroduplexes from homoduplexes on the dHPLC. Using this software, optimal temperatures for dHPLC analysis were found to range between 61°C and 63°C for all amplicons. However, to assess the accuracy of the melting temperature prediction software, melting temperatures were also determined empirically for some known homozygous and heterozygous amplicons. A 586 bp homozygous fragment was analysed at 50°C, 59°C, 60°C, 61°C, 62°C, 63°C and 64°C. The results of the temperature titration are presented in Figure 28 (Appendix C.1). Analysis at 59°C revealed a single, sharp peak similar to the peak observed under non-denaturing conditions (50°C). At 61°C however, a change in the peak pattern could be observed indicating partial melting of the homoduplex DNA. With increasing temperatures, resolution of the homoduplex peaks was reduced and peaks were shifted to shorter retention times as partially melted DNA is retained less strongly on the hydrophobic column. Analysis at 64°C resulted in no peak, indicating that the DNA was completely melted at this temperature. Thus, the optimal temperature for the 586 bp amplicon determined by empirical temperature titration was 61°C, which correlated exactly with the temperature predicted by the melting temperature software. In addition, a 546 bp amplicon spanning exons 27 and 28, that was known to be heterozygous, was analysed at multiple temperatures to determine the range of temperatures at which heterozygosity could be detected. The results are shown in Figure 29 (Appendix C.2) and demonstrate resolution of heteroduplex peaks from homoduplex peaks at at least three temperatures. The temperature that showed best resolution was 63°C, however, heteroduplex peaks were also observed at 62°C and 64°C. The optimal temperature for dHPLC analysis predicted by the melting temperature software was 62°C, and therefore would have detected the polymorphism present in this sample. Melting temperatures were determined empirically for two other amplicons and were found to correlate with the melting temperatures predicted by the software. Hence, melting temperatures determined by empirical temperature titration and temperatures predicted by the DHPLCMelt software correlated for all amplicons tested, and therefore optimal temperatures for dHPLC analysis of all other amplicons were determined using the DHPLCMelt software.

In a previous study by Jones et al. [172] examining the sensitivity and specificity of denaturing HPLC, 96% of heterozygotes could be detected at the temperature predicted by the DHPLCMelt software. However, to achieve sensitivities of >96%, Johns et al. recommend to analyse samples at two different temperatures: the predicted melting temperature (T_m) and the T_m plus 2°C. As high sensitivity was required in the present study, all samples were analysed at both recommended temperatures. For some amplicons however, no peak was detectable at the T_m plus 2°C due to complete melting of the DNA. These amplicons were analysed at the T_m plus 1°C as this combination of temperatures also detected most heterozygotes in the study performed by Jones et al. However, six amplicons could only be analysed at one temperature (the T_m). Every further increase of the temperature resulted in no peak, and therefore analysis was not possible at any further temperature above the T_m.

5.2.3 Reproducibility of results

To ensure that results obtained by dHPLC analysis were reproducible, a number of samples that showed heteroduplex peaks on the first analysis were repeatedly injected on different runs of the HPLC. The chromatograms obtained from a 302 bp amplicon are presented in Figure 30 (Appendix C.3). Although retention times varied slightly on different runs, all peak patterns of the 302 bp heterozygote sample were significantly different from the homozygous reference shown at the top. Sequencing of the heterozygous sample revealed a T to C substitution at position 62 of the fragment, indicating that sequence variations as little as a single base pair can be easily detected by dHPLC. Reproducibility was tested for a second sample, which was 404 bp in size. The results are shown in Figure 31 (Appendix C.3) and demonstrate resolution of heteroduplex peaks from homoduplex peaks on all different runs of this sample. As observed for the 302 bp fragment, retention times and peak morphology varied slightly, but did not affect the results of the analysis. Heterozygosity of the 404 bp fragment was confirmed by sequencing and revealed three single base pair substitutions (C to T, G to A and C to G at positions 99, 178 and 377, respectively).

However, there was one instance where heteroduplex peaks could not be reproduced on a second run of the same sample. Sequence analysis of a 525 bp fragment did not reveal any sequence variations, and therefore, the chromatogram observed on the dHPLC was inspected more closely. As shown in Figure 32 (Appendix C.3), the peak pattern observed for the unknown sample was significantly different from the homoduplex peak of the homozygous reference, which is shown at the top. Interestingly, both samples also differed in their retention times, which is not usually observed for homoduplex and heteroduplex peaks of the same amplicon, but is characteristic for partial melting of the DNA. To further investigate the melting behaviour of this sample, the sample was re-injected at multiple temperatures ranging from 63°C (T_m) to 65°C. The results of the second run are also shown in Figure 32 (Appendix C.3) and demonstrate a homoduplex peak pattern of the unknown sample at the T_m, confirming the results of the DNA sequencing analysis. At 64°C (T_m plus 1°C), the chromatogram showed a small shoulder at the edge of the homoduplex peak and the retention time was slightly decreased, both indications of partial melting of the DNA. At 65°C (T_m plus 2°C), the retention time was further decreased and the peak pattern was very similar to the pattern observed on the initial run of this sample at 63°C. These results suggest that the heteroduplex peak of the initial analysis might in fact be due to partial melting of the homoduplex DNA. Instabilities of the column temperature might cause homoduplexes to melt at temperatures adjusted to resolve heteroduplex DNAs. Differences between oven temperatures and actual column temperatures have been observed previously for dHPLC instruments [172]. Similar chromatograms were observed for another sample of the same 525 bp amplicon analysed at the same run of the HPLC. As found for the first sample, DNA sequencing did not reveal any sequence variations. Re-injection on a second run resulted in a homoduplex peak pattern at 63°C, but the peak pattern observed at 65°C resembled the pattern of the initial analysis. The hypothesis, that the heteroduplex peaks observed for both samples were due to temperature instabilities of the instrument, was further supported by the fact that apparent heteroduplex peak patterns were significantly different from peaks observed for fragments confirmed to be heterozygous by DNA sequencing.

5.2.4 Sensitivity of the dHPLC method

To assess the sensitivity of the dHPLC method, 69 samples were screened for polymorphisms by dHPLC and in addition analysed by DNA sequencing. The 69 samples represented three different amplicons spanning exons 10+11, exon 12 and exon 29. These amplicons were chosen as comparison of complete C4 sequences available at GenBank had shown that exons 12 and 29 and the intronic regions flanking exon 10 and 11 included several polymorphic sites.

Of the 69 samples analysed by dHPLC and sequencing, 34 were heterozygotes. Characterisation of the polymorphisms by sequence analysis revealed several single base substitutions, including C to T, C to A, G to A, G to T and G to C. Of the 34 heterozygotes, 23 contained a single polymorphic site, 8 samples had two polymorphic sites and 3 had three polymorphic sites. All heterozygotes could be detected by dHPLC using two different temperatures (T_m and T_m plus 2°C for E29, T_m and T_m plus 1°C for E10+11 and E12). At the T_m, 18 of the 34 heterozygotes could be detected, which corresponds to a sensitivity of 53%. No polymorphisms were identified by DNA sequencing among 35 samples that showed peak patterns characteristic for homoduplexes. In summary, our results suggest that a high sensitivity of the method (100%) can only be achieved using a combination of two temperatures as recommended by Jones et al. [172]. Thus, for all other samples, two temperatures were used.

Examination of the sensitivity also included two AHs with three C4 genes (7.1 and 65.1 AHs). Analysis of quantitative nucleotide peaks in sequence electropherograms of these AHs indicated that nucleotide substitutions in exon 29 (codon 1226) of the 7.1 AH and in introns 11 (nucleotide position 3234) and 29 (nucleotide position 8720) of the 65.1 AH were only present on one of the three C4 genes. Therefore, mixing of these samples with the homozygous reference resulted in a 3:1 ratio of DNA carrying the reference allele and DNA carrying the substitution. However, all three substitutions were detected by dHPLC, indicating that one allele can still be detected in the presence of a 3-fold excess of another allele.

Another critical factor affecting the sensitivity of the dHPLC method is the quality of the column. After a certain number of injections the column shows degradation of performance characterised by low resolution of heteroduplex peaks from homoduplex peaks. An example is presented in Figure 33 (Appendix C.4). Heteroduplex peaks were resolved in the chromatogram shown at the top, whereas low resolution of peaks was observed in the chromatogram shown below. Therefore, the performance of the HPLC had to be evaluated prior to each run. Under non-denaturing conditions (50°C), the HaeIII digested pUC18 size marker was used to evaluate column quality. A chromatogram of the pUC18/HaeIII marker is shown in Figure 34 (Appendix C.4). Resolution of the 257 bp and 267 bp peaks indicated adequate performance of the column. Under partially denaturing condition (61°C to 65°C), samples that were found to be heterozygous in previous runs were re-injected to confirm resolution of heteroduplex peaks under current column conditions.

5.2.5 Detection of polymorphisms by dHPLC

Using the denaturing HPLC approach, 529 samples representing 23 different amplicons were screened for sequence variations. The results of the dHPLC analysis are presented in Table 19. Of the 529 samples, 103 showed a peak pattern that was significantly distinct from the pattern observed for the homozygous reference. Heterozygotes were detected in 17 different amplicons spanning exons 1, 2, 3, 9, 10+11, 12, 13, 14, 15+16, 29, 30, 31, 32+33, 34+35, 36+37, 38 and 40, whereas no heterozygotes were detected in the 6 amplicons spanning exons 4+5, 6, 7, 8, 39 and 41. For each amplicon, 23 different samples representing different AHs were analysed. The number of samples that were detected to be heterozygous varied between different amplicons. For example, more than 50% of the samples were found to be heterozygotes when screening amplicons 9, 12, 29 and 30. In contrast, screening of amplicons 1, 2, 14, 32+33, 34+35 and 36+37 revealed only one or two heterozygous samples, suggesting that these exons may contain rare mutations.

Of the 103 heterozygotes, 69 could be detected at the T_m, which corresponds to an overall sensitivity of <67% at this temperature. In a previous study by Jones et al. [172] examining the utility of the DHPLCMelt software for the prediction of the optimal melting temperature, 96% (99/103) of heterozygotes could be detected at the temperature recommended by the software. However, the authors also recommend analysis of samples at a second temperature (T_m plus 2°C) if higher sensitivity is required. Following these recommendations, another 34 heterozygotes could be identified, which were not detectable at the T_m. Samples that showed heteroduplex peaks at the T_m were not analysed systematically at the second temperature, and therefore the sensitivity of this temperature could not be assessed. However, there were at least two examples where a heterozygote was detected at the T_m, but escaped detection at the second temperature. Moreover, resolution of heteroduplex peaks from homoduplex peaks was much lower at the second temperature, making the interpretation of the results more difficult. Thus, analysis of samples at the second temperature instead of the T_m would not have resolved all heterozygotes. In conclusion, dHPLC analysis should be carried out using a combination of two temperatures as this results in the highest sensitivity of the method.

5.3 Characterisation of detected polymorphisms

PCR products that were found to be heterozygous by dHPLC were analysed by DNA sequencing to characterise the nature of the sequence variations. The results are presented in Table 20. PCR amplicons covered the exonic regions from exon 1 to 16 and from exon 29 to 41 of the C4 genes. However, all primers were designed to bind in intronic regions, and therefore amplicons also included intronic sequence (i.e. intronic sequences flanking the exons). Some of the amplicons spanned more than one exon as exons may be less than 100 bp in size, and hence, these amplicons included the complete intronic region between two exons. As a result, sequence variations listed in Table 20 include both intronic and exonic polymorphisms.

Primers for amplification of exonic regions were designed to bind within conserved regions of the C4 genes. Therefore, PCR reactions resulted in amplification of all C4 genes present on a haplotype (i.e. all C4A and C4B genes) and hence, DNA sequences were composite sequences. Quantitative analysis of sequence electropherograms has been shown to reflect C4 copy numbers present on a haplotype (see Section 4.2). To confirm the previous findings, relative nucleotide quantities were determined for all polymorphic positions identified in the this part of the study. No heterozygous positions were found on haplotypes predicted to carry only one C4 gene. Quantitative sequence analysis of AHs carrying two C4 genes revealed homozygous as well as heterozygous nucleotide peak patterns at polymorphic sites, indicating that some changes were present on both C4 genes of the haplotype whereas others were only found on one of the two genes. Nucleotide peaks at heterozygous positions were found to have comparable intensities, as would be expected for haplotypes carrying two C4 genes. Analysis of the 7.2 and 65.1 AHs, which both had three C4 genes, revealed a 2 1 ratio of relative nucleotide intensities at most heterozygous positions. Relative nucleotide quantities at heterozygous position are listed in Table 15 for both haplotypes.

5.3.1 Exonic polymorphisms

Characterisation of exonic regions not included in the 3.3 kb of sequence described in Chapter 4 revealed 15 additional polymorphic sites. The polymorphic sites were located in exons 3, 9, 11, 12, 13, 29, 33, 34, 36 and 40. All changes were single nucleotide substitutions, no insertion or deletions were found. Of the 15 nucleotide substitutions, six resulted in an amino acid change. However, only three of the six changes resulted in a change in the property of the amino acid. The amino acid change from an arginine to a

AH	C4 genes	Position		Nucleotide	Relative quantity
7.2	C4A3, C4A3, C4B1	E12	476	Y	C < T
		I28	8441	Y	C < T
		E29	1226	R	G > A
		I29	8720	S	C < G
65.1	C4A1, C4B1, C4B2	E3	122	S	C > G
		I11	3234	Y	C > T
		I15	4572	R	A > G
		I29	8720	S	C G
		E34	1473	Y	Cannot determine

Table 15: Quantitative sequence analysis of 7.2 and 65.1 AHs. Both AHs carry three C4 gens. Quantitative sequence analysis revealed a 2:1 ratio of relative nucleotide intensities, representing the number of C4 genes present on both haplotypes. Relative nucleotide quantities at heterozygous positions are presented. IUPAC codes have been used to indicate the presence of two nucleotides (R=A+G, Y=C+T, S=G+C).

tryptophan in exon 12 (codon 458) and the change from a histidine to a proline in exon 13 (codon 530) resulted in a charged amino acid being replaced with a non-polar one. The nucleotide substitution in exon 29 (codon 1267) resulted in an change from the non-polar amino acid alanine to the polar amino acid serine. The three other amino acid changes in exons 3 (codon 122), 9 (codon 328) and 33 (codon 1395) resulted in no change in the property of the amino acid.

5.3.2 Intronic polymorphisms

The sequence differences in the introns consisted mostly of single base pair substitutions. However, there was one 4 bp deletion (relative to the 8.1 C4B1) in intron 2, that was only present on one of the two C4 genes of the 46.1 AH and was not found on any other haplotype. Six of the 23 amplicons used to characterise the exonic regions spanned two exons, and therefore also covered the complete intronic region between the two exons. However, only one nucleotide substitution was found within the six introns that were analysed completely. This nucleotide change was located in intron 15 at nucleotide position 4572. Introns 4, 10, 32, 34 and 36 were found to be conserved between all C4 genes included in this study. In a previous study comparing a C4B1 gene with a C4A3 gene, Ulgiati et al. [104] identified several sequence differences in introns 4, 10 and 15, whereas no nucleotide changes were identified in introns 32, 34 and 36. Therefore, both studies are consistent in that the small introns towards the 3' end of the C4 gene are conserved between different C4 genes. The additional changes found in introns 4, 10 and 15 by Ulgiati et al. might be due to the comparison of their C4B1 gene with a C4A3 gene of a different AHs not included in our study. In addition to the sequence variations in introns 2 and 15, eleven nucleotide substitutions were identified in the intronic regions flanking the exons. Most of the changes were present on several different haplotypes, however, four nucleotide substitutions were specific for individual haplotypes. For example, the C to T substitution in intron 12 (nucleotide position 3678) was unique to one of the C4 genes of the 54.1 AH. Both C to T substitutions in intron 13 (positions 4001 and 4010) were specific for one particular haplotype. The substitution at position 4001 was only found on the 18.1 AH and the nucleotide change at position 4010 was unique to the 44.4 AH. The A to G substitution in intron 30 (position 9240) was only present on one of the C4 genes of the 7.1 AH.

5.3.3 Promoter region

The amplicon spanning exon 1 also included about 180 bp of upstream promoter region. Within this region, one single base substitution was found at position -146 of the C4 promoter. The A to G substitution was only present on one of the two C4 genes of the 44.2 AH, but was not found on any other C4 gene included in this study. A previous study by Vaishnaw et al. [178] investigating C4 promoter polymorphism in SLE patients (n=52) and healthy controls (n=51) had failed to identify any polymorphism within 500 bp upstream of the transcriptional start site. Their study also comprised the comparison of different AHs, demonstrating extensive conservation of the upstream promoter region. Ulgiati et al. [103] compared the C4B1 gene of the 8.1 AH with a C4A3 gene and found no sequence differences within 1.8 kb upstream of the transcriptional start site, suggesting that the wide range of C4 concentrations observed in the blood plasma of different AHs is not explained by promoter polymorphisms. However, our study identified one polymorphic site at position -146 of the C4 promoter. Vaishnaw et al. [105] analysed the upstream promoter region using reporter gene assays and demonstrated that the sequence contained within the -178 to -39 region is associated with maximal reporter gene expression. Studies by Ulgiati et al. [106] had shown that the region important for transcriptional activity included Sp1 and E box sites. The -140 region contained four closely positioned GT boxes that were found to be important as mutations introduced at -140 resulted in almost no reporter gene expression. The polymorphism identified in the present study is located adjacent to the region that is important in the regulation of transcriptional activity. Therefore, the polymorphism identified at -146 might affect C4 expression levels. The only haplotype that showed the polymorphism was the 44.2 AH, which was heterozygous at -146, indicating that the A to T substitution was only present on either the C4A3 or the C4B1 gene of this haplotype. None of the other C4A3 or C4B1 genes included in the present study showed the -146 polymorphism, suggesting that the A to T substitution might be unique to the 44.2 AH. The 44.2 AH has been shown to be associated with IgA deficiency and celiac disease [40]. Whether the polymorphism at -146 of the C4 upstream promoter region might contribute to disease susceptibility needs to be investigated.

5.4 Comparison of coding regions

The complete exonic sequence of various C4 genes from different AHs was analysed in the present study. The most polymorphic region spanning from exon 17 to 28 was characterised by DNA sequencing as described in Chapter 4. All other exonic regions were screened for polymorphisms by dHPLC and the nature of the identified polymorphisms was characterised by DNA sequencing. The sequence differences seen when comparing the exonic sequences of all C4 genes included in this study were plotted as shown in Figure 16. Exons 26 and 28, known to encode the isotype specific residues and the Rodger and Chido antigenic determinants, were found to be most polymorphic. Other sequence differences in the C4d region were located in exons 25 and 29. The C4d fragment of the C4 α-chain is known as the most polymorphic domain of C4. However, a number of sequence differences were present outside the C4d region. Towards the 3' end of the C4 gene, exons 33, 34, 36 and 40 were found to be polymorphic. Upstream of the C4d region, exons 3, 9, 11 to 13, 17 and 19 to 22 contained several sequence differences. In contrast, comparison of a C4B1 and a C4A3 gene by Ulgiati et al. [104] had shown that exonic sequence differences were located only in the exons clustered around the C4d region. However, the present study included a wider range of C4 allotypes and ancestral haplotypes, and hence additional allotype and haplotype specific polymorphisms were identified.

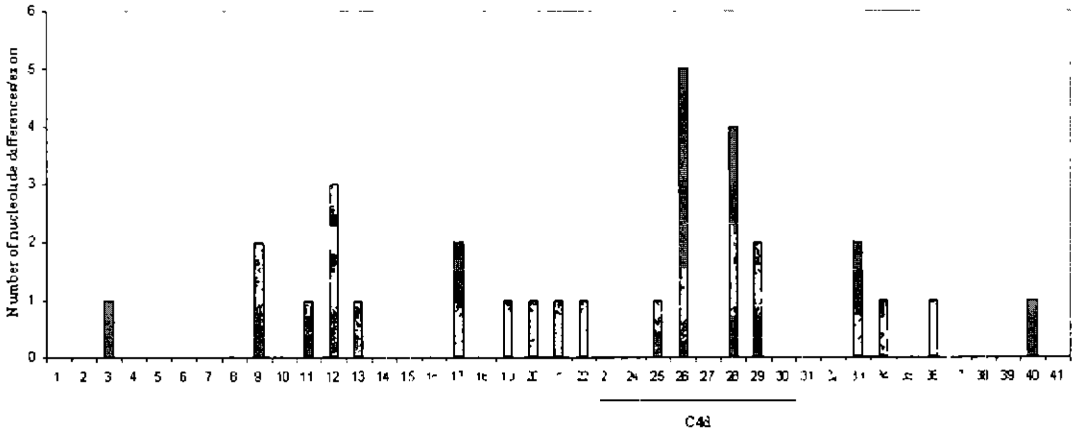

Figure 16: Exonic sequence differences. The complete exonic sequence of C4 genes from various AHs was compared and the number of nucleotide differences per exon was plotted

In the present study, a number of novel exonic sequence differences were identified. A list of changes that have to our knowledge not been previously described in the literature is presented in Table 16. Four of the nucleotide substitutions located in exons 13, 17 and 33 resulted in an amino acid change, whereas all other substitutions were synonymous changes. Most of the changes were only present on one of the 24 haplotypes studied, suggesting that these changes might be haplotype specific or might be characteristic for one of the less frequent C4 allotypes. For example, three nucleotide substitutions (located in exons 9, 33 and 36) were only found on the 42.1 AH. This haplotype carries a C4A12 and a C4A91 gene, both alleles that have a low frequency in most populations. Other changes were identified on more common C4 allotypes, but seemed to be unique to individual AHs. For example, the C to G substitution in exon 22 was present on the C4A3 gene of the 54.1 AH, but was not found on any of the other C4A3 genes. In contrast, the amino acid change in exon 13 (codon 530) was present on five different AHs. However, all five AHs had the C4A4 allele in common, suggesting that the amino acid change might be specific for the C4A4 allotype.

Exon	codon	Amino acid	Nucleotide	AHs	C4 genes
9	320	Tyr	C/A	42.1	C4A12 or C4A91
11	408	Asp	C/T	54.1	C4A3 or C4B5
12	469	Thr	T/C	47.1	C4A91
13	530	His/Pro	A/C	18.1, 46.1, 46.2, 55.1, 62.2	C4A4
17	695	Arg/Ser	C/A	62.1	C4B3
17	710	Arg/Gln	G/A	55.1	C4B5
19	786	Thr	G/A	47.1	C4A91
22	918	Gly	G/A	54.1	C4A3
33	1395	Ala/Pro	G/C	42.1	C4A12 or C4A91
33	1403	Asp	T/C	13.1	C4A3, C4B1
34	1473	His	C/T	65.1	C4A2, C4B1 or C4B2
36	1507	Val	C/T	42.1	C4A12, C4A91

Table 16: **Novel exonic sequence differences described in this study.** A number of nucleotide substitutions were identified in exonic regions that have not been previously described in the literature. Four of the nucleotide substitutions result in an amino acid change, whereas all other substitutions are synonymous changes.

5.4.1 C4A and C4B null alleles

All sequence differences found in exonic regions were single base substitutions. No insertions or deletions were identified within the exons, indicating that all C4 genes had the same reading frame. Previous studies investigating the molecular basis of C4 pseudogenes have identified various small insertions and deletions, that cause frame-shifts and generate premature stop codons. Two 1-bp deletions located in exons 13 and 20 were identified as the cause of C4BQ0 and C4AQ0 pseudogenes, respectively [119, 121]. A 2-bp insertion was found in exon 29 of both C4AQ0 and C4BQ0 pseudogenes [120]. None of these insertions or deletions was found in any of the C4 genes included in our study, suggesting that the C4A and C4B null alleles present on some AHs were most likely due to gene deletions rather that point mutations. Further evidence supporting the presence of gene deletions comes from the homozygous nucleotide peaks observed in the sequence electropherograms at the isotype specific positions (see Section 4.2).

5.4.2 Amino acid changes

Comparison of the complete exonic regions of all C4 genes included in this study revealed 17 amino acid changes (see Table 17). Four of the changes, at amino acid positions 1101, 1102, 1105 and 1106 in exon 26, were known to encode the isotype specific residues. The Rodger- and Chido-specific residues were encoded within exons 25 and 28. In addition to this, only 9 polymorphic amino acid residues were identified. Of the 9 changes, 4 were located in the β-chain of the C4 protein (in exons 3, 9, 12 and 13). No amino acid change was found in the γ-chain (exons 33 to 41). Thus, all other changes were located within the α-chain. Two changes were identified in exon 17, which encodes the anaphylotoxin C4a [87]. The two amino acid substitutions in exons 22 and 33 were located in the α-chain, but outside the C4d region. Therefore, only 9 of the 17 amino acid changes were present in the C4d region, suggesting that other regions of the C4 genes contribute to the polymorphism observed at the protein level.

The amino acid change at position 458 in exon 12 has been previously shown to be characteristic for the C4A6 allotype [123]. This finding could be confirmed by our study as the C4A6 gene had TGG at position 458 coding for tryptophan, whereas all other C4A and C4B genes had CGG at this position, resulting in an arginine. The arginine to tryptophan substitution has been suggested to be responsible for the hemolytic inactivity of the C4A6 allotype as it results in the disruption of the C5 binding site in the C4 β-chain [123]. C5 binding is important for the assembly of the C5 convertase, and thus the activation of the classical and lectin complement pathways.

Amino acid position 1267 has been previously reported to be the only amino acid difference between C4A3a and C4A3b, with an alanine residue in C4A3a and a serine residue in C4A3b [179]. Our analysis revealed an alanine at position 1267 for most C4 genes. The serine residue was only present in three genes. Both C4A2 and C4A3 of the 52.1 AH had a serine, indicating that the C4A3 of this haplotype was in fact a C4A3b gene. In addition, the 7.1 AH was found to be heterozygous at the first nucleotide position of codon 1267. Thus, either the C4A3 or the C4B1 gene of this haplotype had a serine residue at amino acid position 1267. If the serine residue was present on C4A3 than this gene

MHC					E3	E9	E12	E13	E17	E17	E21	E25	E26	E26	E26	E26	E28	E28	E28	E29	E33
C2	Bf	C4A	C4B	AH	122	328	458	530	695	710	888	1054	1101	1102	1105	1106	1157	1188	1191	1267	1395
C	S	Q0	1	8.1	L	S	R	H	R	R	T	G	L	S	I	H	S	A	R	A	A
	S	Q0	1	44.3		Y					A										
C	F1	3	Q0	18.2							A	D	P	C	L	D		V	L		
	S	3	Q0	58.1							A	D	P	C	L	D		V	L		
C	S	3-3	Q0	44.1		Y					A	D	P	C	L	D	S/N	V	L		
C	S	2-3	Q0	52.1							T/A	G/D	P	C	L	D	S/N	V	L	S	
C	F	2+3	Q0	35.2		Y					A	D	P	C	L	D	N	V	L		
C	F	12+91	Q0	42.1							A	G/D	P	C	L	D					A/P
C	F	91	Q0	47.1							A		P	C	L	D					
C	F	3	1	44.2		S/Y					A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	F	3	1	44.4		S/Y					A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	3	1	7.1							T/A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L	A/S	
C	S	3	1	13.1		S/Y					A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	3+3	1	7.2							A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
	S	2	1	38.1		Y					A		L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	2	1+2	65.1	L/V						A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	6	1	57.1		S/Y	R/W				A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	3	3	62.1		Y			R/S		A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	3	5	54.1		Y					A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
	S	4	5	55.1		S/Y		H/P		R/Q	A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
Q0	S	4	2	18.1	L/V	S/Y		H/P			A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	4	2	46.1	V	S/Y		H/P			A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
C	S	4	2	46.2	V	S/Y		H/P			A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
	S	4	2	62.2	L/V	S/Y		H/P			A	D	L/P	S/C	I/L	H/D	S/N	A/V	R/L		
Alternative codons					CTC	TCT	CGG	CAT	CGT	CGG	ACC	GGC	CTC	TCT	ATA	CAT	AGC	GCG	CGG	GCG	GCA
					GTC	TAT	TGG	CCT	AGT	CAG	GCC	GAC	CCC	TGT	TTA	GAC	AAC	GTG	CTC	TCG	CCA
Alternative aa					L/V	S/Y	R/W	H/P	R/S	R/Q	T/A	G/D	L/P	S/C	I/L	H/D	S/N	A/V	R/L	A/S	A/P

Table 17: Amino acid differences between various C4 genes representing different AHs. The complete exonic regions of the C4 genes have been compared. Only differences that resulted in an amino acid change are shown (relative to the C4B1 gene of the 8.1 AH). Codon positions, alternative codons and alternative amino acids are indicated.

would have been a C4A3b gene. All other C4A3 genes included in this study had GCG at position 1267 resulting in an alanine, and hence were C4A3a genes.

5.5 C4 allotypic variation

C4 allotyping based on differences in electrophoretic mobility and hemolytic activity has revealed more than 40 C4 allotypes [83]. The C4 allotypes can be separated by agarose gel electrophoresis of carboxypeptidase B and neuraminidase treated EDTA plasma [180]. Neuraminidase treatment is required to remove the glycosylation present in human C4. Glycosylation has been shown to be heterogeneous, and therefore contributes to the structural variation of C4 proteins [181]. Carboxypeptidase B removes C-terminal basic amino acids which is required to produce single, distinct bands for each C4 allotype [182]. After separation of C4 allotypes by agarose gel electrophoresis, C4 proteins are immobilised by immunofixation and bands are visualised by protein staining. Each C4 allotype is characterised by a single band, although some C4 allotypes show similar migration rates and are therefore difficult to distinguish.

To examine whether the allotypic variation observed at the protein level is encoded within the C4 gene sequence, the amino acid changes listed in Table 17 were examined in more detail.

5.5.1 Glycosylation and sulphation sites

Post-translational modifications of C4 involve glycosylation at four N-linked glycosylation sites [87]. Three of the glycosylation sites are located within the α-chain at residues 843, 1309 and 1372, respectively. The other glycosylation site is present in the β-chain at position 207. All four residues have been found to be conserved between the C4 genes analysed in the present study. Therefore, the heterogeneity of the glycosylation observed at the protein level is not caused by amino acid changes at the four N-linked glycosylation sites.

Post-translational modification also includes sulphation [183], which is known to be a modulator of extracellular protein-protein interactions [184]. The sulphation sites are clustered at residues 1198, 1199 and 1200 of the α-chain. Analysis of the C4 genes revealed no polymorphism at any of the three sulphation sites, indicating that these residues were conserved and do not contribute to the allotypic variation.

5.5.2 Changes in property of amino acids

The 17 amino acid changes listed in Table 17 were further examined to see whether the changes affected the property of the amino acid. Of the 17 changes, 9 were found to result in no change in the property of the amino acid, including 6 non-polar amino acids that were replaced with other non-polar amino acids and 3 polar amino acids that were replaced by other polar ones. One of the 17 changes resulted in a non-polar amino acid being replaced with a polar one (alanine to serine substitution in exon 29). Five other changes involved charged amino acids that were replaced by uncharged (i.e. polar or non-polar) amino acids. For example, the four positively charged arginine residues in exons 12, 17 and 29 were replaced by uncharged amino acids (tryptophan, serine, glutamine and leucine, respectively). The histidine to proline substitution in exon 13 also resulted in a charged amino acid being replaced by an uncharged one. In addition, the substitution of a glycine residue with an aspartic acid residue in exon 25 resulted in a change in the charge of the amino acid, as aspartic acid is a negatively charged amino acid and glycine is an uncharged one. The histidine to aspartic acid substitution in exon 26 involved a change from a positively charged amino acid to a negatively charged one.

As differences in charge affect the electrophoretic mobility of proteins in agarose gels, the amino acid substitutions involving charged residues might contribute to the allotypic variation of the C4 proteins. To further explore this possibility, all C4 genes included in this study were examined for changes encoding charged amino acid residues.

5.5.3 Contribution of charged amino acids to the allotypic variation

The amino acid changes involving charged amino acids comprised seven polymorphic sites located in exons 12, 13, 17, 21, 25, 26 and 28. The region spanning from exon 16 to 28 had been characterised by isotype specific amplification and subsequent sequencing (see Section 4.3). Therefore, C4 sequences spanning this region had been separated to represent individual C4 genes (Table 13). The sequences spanning exons 12 and 13 were composite sequences representing all C4 genes present on a haplotype. To split these sequences into C4A and C4B specific sequences, a isotype specific PCR was developed spanning from exon 12 to the isotypic site in exon 26. Using this method, heterozygous sequences from 18.1, 46.1, 46.2, 55.1, 57.1 and 62.2 AHs (see Table 17) could be assigned to either C4 isotype.

The amino acids present at these seven polymorphic sites are listed in Table 18 for all C4 genes studied. To further explore the role of these residues, the amino acids of each individual C4 gene were compared to the 8.1 C4B1 gene. Where amino acid differences were identified, the resulting change in the charge was calculated For example, the arginine to tryptophan substitution at position 458 of the 57.1 C4A6 gene resulted in a positively charged amino acid being replaced with an uncharged one Therefore, the C4A6 gene encoded one positive charge less than the 8.1 C4B1 gene, or in other words, had a net charge of -1 compared to the 8 1 C4B1. However, there were a number of additional substitutions present on the C4A6 gene. The glycine to aspartic acid substitution resulted in the introduction of another negative charge on the C4A6 allotype. The amino acid change in exon 26 involved the substitution of a positively charged amino acid with a negatively charged one. Hence, this change added another two negative charges to C4A6. One further negative charge was introduced to C4A6 due to the amino acid substitution in exon 28. Adding up all changes present on the C4A6 gene, the C4A6 gene encoded a net charge of -5 compared to the 8.1 C4B1 gene. Using this approach, cumulative charges were calculated for all other C4 genes. The results are presented in Table 18.

Cumulative charge differences of C4 genes representing the same C4 allotype were compared and plotted as shown in Figure 17. Cumulative charges of C4 genes representing the same C4 allotype were generally found to correlate. For example, the ten C4B1 genes included in this study had the same amino acids at all seven polymorphic positions encoding charge differences, and hence, all ten C4B1 genes had identical calculated net

		E12	E13	E17	E17	E25	E26	E26	Charge difference
C4 gene	AH	458	530	695	710	1054	1106	1191	(compared to 8.1 C4B1)
C4B1	7.1	Arg	His	Arg	Arg	Gly	His	Arg	0
	7.2	Arg	His	Arg	Arg	Gly	His	Arg	0
	8.1	Arg	His	Arg	Arg	Gly	His	Arg	0
	13.1	Arg	His	Arg	Arg	Gly	His	Arg	0
	38.1	Arg	His	Arg	Arg	Gly	His	Arg	0
	44.2	Arg	His	Arg	Arg	Gly	His	Arg	0
	44.3	Arg	His	Arg	Arg	Gly	His	Arg	0
	44.4	Arg	His	Arg	Arg	Gly	His	Arg	0
	57.1	Arg	His	Arg	Arg	Gly	His	Arg	0
	65.1	Arg	His	Arg	Arg	Gly	His	Arg	0
C4B2	18.1	Arg	His	Arg	Arg	Asp	His	Arg	-1
	46.1	Arg	His	Arg	Arg	Asp	His	Arg	-1
	46.2	Arg	His	Arg	Arg	Asp	His	Arg	-1
	62.2	Arg	His	Arg	Arg	Asp	His	Arg	-1
	65.1	Arg	His	Arg	Arg	Asp	His	Arg	-1
C4B3	62.1	Arg	His	Ser	Arg	Asp	His	Arg	-2
C4B5	54.1	Arg	His	Arg	Arg	Asp	His	Arg	-1
	55.1	Arg	His	Arg	Gln	Asp	His	Arg	-2
C4A91	42.1	Arg	His	Arg	Arg	Gly	Asp	Arg	-2
	47.1	Arg	His	Arg	Arg	Gly	Asp	Arg	-2
C4A12	42.1	Arg	His	Arg	Arg	Asp	Asp	Arg	-3
C4A2	35.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	38.1	Arg	His	Arg	Arg	Gly	Asp	Leu	-3
	52.1	Arg	His	Arg	Arg	Gly	Asp	Leu	-3
	65.1	Arg	His	Arg	Arg	Gly	Asp	Leu	-3
C4A3	7.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	7.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	7.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	13.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	18.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	35.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	44.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	44.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	44.2	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	44.4	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	52.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	54.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	58.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
	62.1	Arg	His	Arg	Arg	Asp	Asp	Leu	-4
C4A4	18.1	Arg	Pro	Arg	Arg	Asp	Asp	Leu	-5
	46.1	Arg	Pro	Arg	Arg	Asp	Asp	Leu	-5
	46.2	Arg	Pro	Arg	Arg	Asp	Asp	Leu	-5
	55.1	Arg	Pro	Arg	Arg	Asp	Asp	Leu	-5
	62.2	Arg	Pro	Arg	Arg	Asp	Asp	Leu	-5
C4A6	57.1	Trp	His	Arg	Arg	Asp	Asp	Leu	-5

Table 18: Amino acid changes involving charged residues. Seven polymorphic amino acid residues have been identified that result in a change in the charge of the amino acid. These residues are encoded within exons 12, 13, 17, 25, 26 and 28 The amino acid residues at these seven positions are indicated for all C4 genes included in our study The resulting charge differences encoded by the seven residues have been calculated (using the 8.1 C4B1 as reference) and are indicated in the column at the right

charges. Similarly, the calculated net charge differences of all C4B2, C4A3 and C4A4 genes (relative to the 8.1 C4B1 gene) were found to be identical within the same allotype. There were only two exceptions where calculated charges differed between C4 genes of the same allotype. One of the four C4A2 genes was found to encode a net charge of -4 while the other C4A2 genes had a net charge of -3. In addition, both C4B5 genes differed by -1 in their calculated charge.

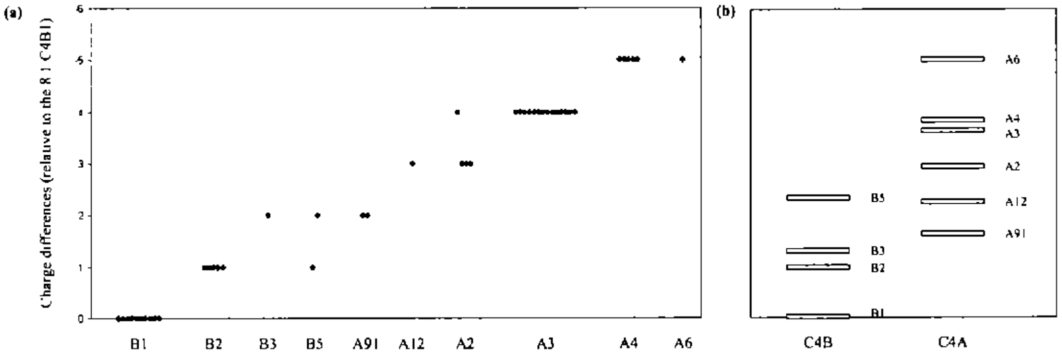

Figure 17: Electrophoretic migration of C4 allotypes. (a) Cumulative charge differences between various C4 allotypes Charge differences relative to the 8.1 C4B1 were calculated based on the analysis of amino acid substitutions involving charged residues. The resulting net charges were compared between C4 genes representing the same allotype and plotted as shown here. (b) Map of C4 allotypes as seen during the VIth Complement Genetics Workshop reference typing

It is notable, that the pattern presented in Figure 17 (a) is remarkably similar to the band pattern observed by C4 allotyping (Figure 17 (b)). To see whether the allotypic variation seen at the protein level could be explained using the cumulative charge model, migration rates of all individual C4 allotypes included in the analysis were examined in more detail. Calculated charge differences within the C4B allotypes indicated the highest negative charge for the C4B5 allotype whereas the C4B1 allotype had the lowest negative charge. Hence, the C4B5 allotype would be expected to migrate fastest in an anodal gel while the C4B1 allotype would be slowest. These theoretical expectations exactly correlate with the electrophoretic mobility observed in an agarose gel for both allotypes. In addition, the C4B2 and C4B3 allotypes were found to have calculated net charges of -1 and -2, respectively. Electrophoretic separation of the C4B allotypes results in intermediate band positions of C4B2 and C4B3 (i.e. both bands are located between the bands of the C4B1 and C4B5 allotypes). These migration rates are reflected by the calculated charges. C4B2 and C4B3 had a higher negative charge than C4B1, and C4B2 had a lower negative charge than C4B5. The calculated charges of C4B3 and C4B5 where found to be identical.

Similarly, electrophoretic mobilities of the C4A allotypes could be explained using the cumulative charge model. For example, the C4A6 allotype had been found to have the highest negative net charge (compared to the 8.1 C4B1) and correspondingly, migrates fastest in an agarose gel. In contrast, the net charge of the C4A91 allotype had been calculated to be -2, and thus was lowest compared to all other C4A allotypes. In an agarose gel, the C4A91 allotype migrates slowest of all C4A allotypes, but shows a relative electrophoretic mobility similar to the C4B3 and C4B5 allotypes. Correspondingly, the calculated net charges of the C4A91, C4B3 and C4B5 allotypes were found to be identical (except for the 54.1 C4B5). Migration rates observed for the C4A allotypes range from C4A91 as the slowest variant and sequentially increase towards C4A6. Calculated charges demonstrated a similar sequential increase in the negative charge, ranging from -2 for C4A91 to -6 for C4A4 and C4A6.

In addition, this model could also explain the differences observed in electrophoretic mobility between both C4 isotypes. In general, C4A allotypes show faster migration rates in agarose gels than C4B allotypes It is known that both isotypes differ by only four amino acids (located at positions 1101, 1102, 1105 and 1106). C4A allotypes have Pro-Cys-Leu-Asp, whereas Leu-Ser-Ile-His are found on C4B allotypes. The amino acid substitution at position 1106 results in a change from a negatively charged amino acid to positively charged one. This corresponds to a net charge difference of -2 between both isotypes, and therefore explains the different migration rates observed by agarose gel electrophoresis.

Minor differences observed between band positions of C4 allotypes with similar electrophoretic mobilities (e.g. C4A91, C4B3 and C4B5) might be explained by structural effects of the native protein. For example, some charged amino acids may be more exposed than others, and may therefore have a greater contribution to the net charge of the protein. In addition, other amino acid substitutions that do not directly contribute to the charge of the protein may also be involved, as they might affect protein folding which in turns may affect the location of charged amino acids within the protein.

To sum up, the suggested model, based on the calculation of cumulative charge differences, could explain most of the variation seen by C4 allotyping. Of the 45 C4 genes included in the present study, only two could not be explained using the suggested model. The C4A2 gene from the 35.2 AH had a aspartic acid at position 1054 that was not seen in any of the three other C4A2 genes. This substitution adds one additional negative charge to this C4A2 allotype resulting in a distinct theoretical migration rate. In addition, the 54 1 C4B5 lacked one negative charge, which was found in exon 17 of the other C4B5 gene However, the 54.1 C4B5 might have another amino acid change somewhere else in the gene

The electrophoretic mobilities of all other C4 alleles representing 10 different C4 allotypes from various ancestral haplotypes could be explained using the suggested model.

The model is based on only seven polymorphic amino acid residues, indicating that only a small fraction of the nucleotide substitutions observed at the DNA level contributes to the allotypic variation seen at the protein level.

5.6 Summary

The exonic regions of the C4 genes, that were not included in the 3.3 kb of sequence described in Chapter 4, were screened for polymorphic sites by denaturing HPLC. The identified polymorphisms were characterised by DNA sequencing and sequences representing various C4 allotypes from different AHs were compared. The results obtained from the analysis are summarised here:

1. Validation of the dHPLC method revealed that it is a highly sensitive method for the detection of unknown polymorphisms.
2. Characterisation of exonic regions revealed 31 single nucleotide substitutions. Seventeen of these substitutions resulted in an amino acid change.
3. No insertions or deletions were present within the exons, suggesting that all C4 genes had the same reading frame and that C4A and C4B null alleles were probably due to gene deletions rather than pseudogenes.
4. The most polymorphic exons are located in the C4d region of the C4 α-chain. However, several exonic polymorphisms are present outside this region.
5. Polymorphic amino acid residues are located in the α- and β-chain of the C4 protein. The γ-chain was found to be conserved between different C4 allotypes.
6. A number of polymorphisms were identified in exonic regions that have not been previously reported in the literature.
7. An A to G substitution was found at position -146 of the C4 promoter. This polymorphism was unique to the 44.2 AH.
8. The presence of a previously described C4A6 specific residue in exon 12 was confirmed. C4A6 carries a tryptophan at residue 458 while all other C4 allotypes have an arginine at this position.
9. A model was suggested explaining the allotypic variation observed for C4 proteins. The model is based on only seven polymorphic amino acid residues. These residues contribute to the charge of the C4 proteins, and hence affect electrophoretic mobilities used to identify C4 allotypes.

MHC					E1			E2			E3			E4+5			E6			E7		E8		E9		E10+11			E12			E13			E14		
C2	Bf	C4A	C4B	AH	62°C	64°C	Sequ.	61°C	62°C	Sequ	62°C	63°C	Sequ.	62°C	63°C	Sequ	61°C	62°C	Sequ	61°C	Sequ.	61°C	Sequ.	61°C	Sequ.	61°C	62°C	Sequ	62°C	63°C	Sequ	62°C	63°C	Sequ.	63°C	64°C	Sequ.
	S	Q0	1	44.3	-	-		-	-		-	-		-	-		-	-	N	-		-		x	Y	x	x	Y	x		Y	-	-		-	-	
C	Fl	3	Q0	18.2	-	-		-	-		-		N	-	-		-	-		-		-		-		-	-	N	-	-	N	-	-		-	-	
	S	3	Q0	58.1	-	-		-	-		-	-		-	-		-	-		-		-		-		-	-	N	-	-	N	-	-	N	-	-	
C	S	3+3	Q0	44.1	-	-	N	-	-		-		N	-	-		-	-		-		-		x	Y	-	-	N	-	x	Y	-	-		-	-	
C	S	2+3	Q0	52.1	-	-		-	-	N	-	-		-	-		-	-		-		-		-		-	-	N	-	-	N	-	-		-	-	
C	F	2+3	Q0	35.2	-	-		-	-		-		N	-	-		-	-		-		-		x	Y	-	-	N	-	x	Y	-	-		-	-	
C	F	12+91	Q0	42.1	-	-		-	-		-		N	-	-		-	-		-		-		x	Y	-	-	N	x		Y	-	-		-	-	
C	F	91	Q0	47.1	-	-		-	-		-		N	-	-		-	-		-		-		-		x	x	Y	x		Y	-	-		-	-	
C	F	3	1	44.2	x	-	Y	-	-		-		N	-	-		-	-		-		-		x	Y	-	x	Y	x		Y	-	-	N	-	-	
C	F	3	1	44.4	-	-		-	-		-	-		-	-		-	-	N	-		-		x	Y	-	-	N	x		Y	-	-		x	x	Y
C	S	3	1	7.1	-	-		-	-		-	-		-	-		-	-		-		-		-		-	-	N	-	-	N	-	-		-	-	
C	S	3	1	13.1	-	-		-	-		-		N	-	-		-	-		-		-		x	Y	-	-	N	-	-	N	-	-		-	-	
C	S	3+3	1	7.2	-	-		-	-		-	-		-	-		-	-		-		-		x	Y	-	-	N	-	x	Y	-	-		-	-	
	S	2	1	38.1	-	-		-	-		-	-		-	-		-	-		-		-		x	Y	-	-	N	-	-	N	-	-		-	-	
C	S	2	1+2	65.1	-	-		-	-		-	x	Y	-	-		-	-		-		-		-		-	x	Y	-	-	N	-	-		-	-	
C	S	6	1	57.1	-	-		-	-		-	-		-	-		-	-		-		-		x	Y	-	x	Y	-	x	Y	-	-		-	-	
C	S	3	3	62.1	-	-		-	-	N	-	-		-	-	N	-	-		-		-		x	Y	-	-	N	-	-	N	-	-	N	-	-	
C	S	3	5	54.1	-	-		-	-	N	-	-		-	-		-	-		-		-		x	Y	x	x	Y	-	-	N	x		Y	-	-	N
	S	4	5	55.1	-	-		-	-		-	-		-	-		-	-		-		-		x	Y	-	-	N	-	x	Y	x		Y	-	-	
Q0	S	4	2	18.1	-	-		-	-		x		Y	-	-		-	-		-		-		x	Y	-	-	N	x		Y	x		Y	-	x	Y
C	S	4	2	46.1	-	-		x		Y	x		Y	-	-		-	-		-		-		x	Y	-	-	N	x		Y	x		Y	-	-	
C	S	4	2	46.2	-	-	N	-	-	N	x		Y	-	-		-	-		-	N	-		x	Y	-	-	N	x		Y	x		Y	-	-	
	S	4	2	62.2	-	-	N	-	-	N	-	x	Y	-	-		-	-		-		-		x	Y	-	-	N	-	x	Y	-	x	Y	-	-	N

Table 19: Analysis of PCR products by dHPLC. PCR products spanning from exon 1 to 16 and from exon 29 to 41 are shown. The AH designation and the complotype (C2, Bf, C4A and C4B allotypes) of each DNA are indicated on the left. Temperatures used for denaturing analysis of samples on the HPLC are shown at the top. Heteroduplex peaks observed on the dHPLC are marked by "x", whereas "-" indicates the presence of homoduplex peaks. Blanks indicate that the sample has not been tested at the respective temperature. The results of the sequencing analysis are indicated with Y (polymorphism present) and N (no polymorphism identified).

MHC					E15+16			E29			E30			E31			E32+33			E34+35			E36+37			E38			E39		E40		E41	
C2	Bf	C4A	C4B	AH	62°C	63°C	Sequ.	63°C	65°C	Sequ.	62°C	63°C	Sequ	62°C	64°C	Sequ.	63°C	65°C	Sequ	63°C	64°C	Sequ.	63°C	64°C	Sequ.	63°C	64°C	Sequ.	61°C	Sequ	62°C	Sequ	63°C	Sequ.
	S	Q0	1	44 3	-	-	N	-	-	N	-	-		×	×	Y	-	-		-	-		-	-		-	-		-		-		-	
C	Fl	3	Q0	18.2	-	-	N	×		Y	-	×	Y	-	-		-	-	N	-	-		-	-	N	-	-		-		-		-	
	S	3	Q0	58.1	-	-		×		Y	-	×	Y	-	-		-	-		-	-		-	-		-	-		-		-		-	
C	S	3+3	Q0	44.1	-	-		×		Y	-	×	Y	-	-		-	-		-	-		-	-		×		Y	-		×	Y	-	
C	S	2+3	Q0	52.1	-	-		-	×	Y	-	-		-	-		-	-		-	-		-	-		×		Y	-		-		-	
C	F	2+3	Q0	35 2	-	-		×		Y	-	×	Y	-	-		-	-	N	-	-		-	-		-	-		-		-		-	
C	F	12+91	Q0	42.1	-	-		×		Y	-	×	Y	-	-		×	×	Y	-	-		×	-	Y	-	-		-		-		-	
C	F	91	Q0	47.1	-	-		-	-	N	-	-		×	-	Y	-	-		-	-		-	-		-	-		-		-		-	
C	F	3	1	44.2	-	-		-	-	N	-	-		×	×	Y	-	-		-	-		-	-		-	-		-		-		-	
C	F	3	1	44 4	-	-		×	×	Y	-	×	Y	-	-		-	-		-	-	N	-	-	N	-	-	N	-		-	N	-	
C	S	3	1	7.1	-	-		-	×	Y	-	-		×	-	Y	-	-		-	-		-	-		×		Y	-		-		-	
C	S	3	1	13 1	×	×	Y	-	×	Y	-	×	Y	-	-		-	×	Y	-	-		-	-		-	-		-		-		-	
C	S	3+3	1	7 2	-	-		×	×	Y	-	×	Y	-	-		-	-	N	-	-		-	-		×		Y	-		×	Y	-	
	S	2	1	38.1	-	-		-	×	Y	-	×	Y	-	-		-	-		-	-		-	-		×		Y	-		-		-	
C	S	2	1+2	65.1	×	×	Y	-	×	Y	-	×	Y	-	-		-	-		-	×	Y	-	-		-	-	N	-		-		-	
C	S	6	1	57.1	×	×	Y	-	×	Y	-	×	Y	-	-		-	-		-	-		-	-		-	-		-		-		-	
C	S	3	3	62.1	-	-		-	×	Y	-	×	Y	-	-		-	-		-	-		-	-		×		Y	-		×	Y	-	
C	S	3	5	54.1	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-		-	-		-		-		-	
	S	4	5	55.1	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-		×		Y	-		-		-	
Q0	S	4	2	18.1	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-	N	×		Y	-		-		-	
C	S	4	2	46.1	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-		×		Y	-		-		-	
C	S	4	2	46.2	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-		×		Y	-		-	N	-	
	S	4	2	62.2	-	-		-	-	N	-	-		-	-		-	-		-	-		-	-		×		Y	-		-		-	

MHC					UTR	I2 I2 I2 I2	E3	E9	E9	E11	I11	I11	E12	E12	E12	I12	I12	E13	I13	I13	I15	I28	E29	E29	I29	I30	I30	E33	E33	E34	E36	I38	E40
C2	Bf	C4A	C4B	AH	-146	[illegible]	122	320	328	408	3229	3234	458	469	476	3573	3678	530	4001	4010	4572	8441	1226	1267	8720	9240	9273	1395	1403	1473	1507	13278	1669
C	S	Q0	1	8.1	A	TGTT	CTC	TAC	TCT	GAC	C	C	CGG	ACT	GCC	C	C	CAT	C	C	A	C	CCG	GCG	C	A	T	GCA	GAT	CAC	GTC	T	CTG
	S	Q0	1	44.3	.	....	...	...	.A.	...	T	.	...	...	...	A	.	...	.	.	.	.	...	...	.	.	C	...	...	...	...	.	...
C	FI	3	Q0	18.2	.	....	...	...	...	...	.	.	...	...	...	.	.	...	.	.	.	T	...	...	G	.	.	...	...	...	...	.	...
	S	3	Q0	58.1	.	....	...	...	...	...	.	.	...	...	...	.	.	...	.	.	.	T	...	...	G	.	.	...	...	...	...	.	...
C	S	3+3	Q0	44.1	.	....	...	...	.A.	...	.	.	...	...	..Y	.	.	...	.	.	.	T	...	...	G	.	.	...	...	...	...	C	..A
C	S	2+3	Q0	52.1	.	....	...	...	...	...	.	.	...	...	...	.	.	...	.	.	.	.	...	T..	.	.	.	...	...	...	...	C	...
C	F	2+3	Q0	35.2	.	....	...	...	.A.	...	.	.	...	...	..T	.	.	...	.	.	.	T	...	...	G	.	.	...	...	...	...	.	...
C	F	12+91	Q0	42.1	.	....	...	..Y	...	...	.	.	...	...	...	M	.	...	.	.	.	T	...	...	G	.	.	S..	...	...	..T	.	...
C	F	91	Q0	47.1	.	....	...	...	...	...	T	.	...	..C	...	A	.	...	.	.	.	.	...	...	.	.	C	...	...	...	...	.	...
C	F	3	1	44.2	R	....	...	...	.M.	...	T	.	...	...	...	A	.	...	.	.	.	.	...	...	.	.	Y	...	...	...	...	.	...
C	F	3	1	44.4	.	....	...	...	.M.	...	.	.	...	...	...	M	.	...	Y	.	.	T	...	...	G	.	.	...	...	...	...	.	...
C	S	3	1	7.1	.	....	...	...	...	...	.	.	...	...	...	.	.	...	.	.	.	.	...	K..	.	R	.	...	...	...	...	C	...
C	S	3	1	13.1	.	....	...	...	.M.	...	.	.	...	...	...	.	.	...	.	.	R	.	...	...	S	.	.	...	..C	...	...	.	...
C	S	3+3	1	7.2	.	....	...	...	...	...	.	.	...	...	..Y	.	.	...	.	.	.	Y	..R	...	S	.	.	...	...	...	...	C	..A
	S	2	1	38.1	.	....	...	...	.A.	...	.	.	...	...	...	.	.	...	.	.	.	Y	..R	...	S	.	.	...	...	...	...	C	...
C	S	2	1+2	65.1	.	....	S..	...	...	...	.	Y	...	...	...	.	.	...	.	.	R	.	...	...	S	.	.	...	...	..Y	...	.	...
C	S	6	1	57.1	.	....	...	...	.M.	...	.	Y	Y..	...	...	.	.	...	.	.	R	.	...	...	S	.	.	...	...	...	...	.	...
C	S	3	3	62.1	.	....	...	...	.A.	...	.	.	...	...	...	.	.	...	.	.	.	Y	..R	...	S	.	.	...	...	...	...	C	..A
C	S	3	5	54.1	.	....	...	...	.A.	..T	Y	.	...	...	...	.	Y	...	.	.	.	.	...	...	.	.	.	...	...	...	...	.	...
	S	4	5	55.1	.	....	...	...	.M.	...	.	.	...	...	...	M	.	.M.	.	.	.	.	...	...	.	.	.	...	...	...	...	C	...
Q0	S	4	2	18.1	.	....	S..	...	.M.	...	.	.	...	...	...	M	.	.M.	.	Y	.	.	...	...	.	.	.	...	...	...	...	C	...
C	S	4	2	46.1	.	indel	G..	...	.M.	...	.	.	...	...	...	M	.	.M.	.	.	.	.	...	...	.	.	.	...	...	...	...	C	...
C	S	4	2	46.2	.	....	G..	...	.M.	...	.	.	...	...	...	M	.	.M.	.	.	.	.	...	...	.	.	.	...	...	...	...	C	...
	S	4	2	62.2	.	....	S..	...	.M.	...	.	.	...	...	...	M	.	.M.	.	.	.	.	...	...	.	.	.	...	...	...	...	C	...
Alternative							CTC	TAC	TCT	GAC			CGG	ACT	GCC			CAT					CCG	GCG				GCA	GAT	CAC	GTC		CTG
codons							GTC	TAT	TAT	GAT			TGG	ACC	GCT			CCT					CCA	TCG				CCA	GAC	CAT	GTT		CTA
Alternative aa							L / V	Y	S / Y	D			R / W	T	A			H / P					P	A / S				A / P	D	H	V		L

Table 20: Polymorphisms detected by dHPLC and characterised by DNA sequencing. The approach covers all exonic regions that are not included in the 3.3 kb of sequence described in Section 4 (i e exon 1 to 16 and exon 29 to 41) and several flanking intronic regions. Only differences compared to the C4B1 gene of the 8 1 AH are indicated Composite sequences for all C4 genes present on the haplotype are shown. Nucleotide numbering is relative to the C4B1 gene published in GenBank (accession number U24578) [176]. The codon position is indicated for exonic polymorphisms. IUPAC codes have been used to indicate the presence of two nucleotides (R=A+G, Y=C+T, K=G+T, M=A+C, S=G+C, W=A+T). "indel" indicates that a deletion relative to 8.1 is present on one of the C4 genes, but not on the other.

6 C4 and MHC associated diseases

6.1 Introduction

The MHC is well known for its association with numerous diseases, particular autoimmune diseases. Although the MHC has been extensively studied for many years, very few practical benefits in the treatment of these disease have been achieved. The tight linkage disequilibrium across the MHC makes it extremely difficult to identify individual disease susceptibility genes. Further studies are required to characterise the highly polymorphic genes located in the MHC region.

Various studies suggest a role of the complement C4 genes in susceptibility to disease. Their complex organisation and great genetic diversity render C4 an excellent candidate gene for MHC associated disease studies. However, current studies are limited by the lack of molecular markers that would facilitate the determination of allelic frequencies in normal and disease populations. Previous studies have shown that complete or partial C4 deficiencies have an increased frequency in patients with autoimmune or immune complex diseases, such as systemic lupus erythematosus (SLE), insulin-dependent diabetes mellitus (IDDM), IgA deficiency, rapid progression to HIV infection and multiples sclerosis. However, these disease association do not necessarily imply a direct role of C4 in susceptibility to disease but could also reflect the effect of closely linked genes. Elucidation of the distribution of all C4A and C4B alleles in different ethnic groups as well as in different disease populations would yield further insights into the role of C4 polymorphism in disease.

6.2 Haplospecific polymorphisms

Comparison of DNA sequences revealed a number of haplospecific polymorphisms, i.e. polymorphisms that were unique to one particular AH and that were not found on any other haplotype. The AHs that carry unique single nucleotide polymorphisms (SNPs) and their alleles are listed in Table 22. The Table includes haplospecific polymorphisms that were located within the 3.3 kb of sequence (spanning from intron 16 to 28, see Chapter 4) as well as haplotype specific polymorphisms that were identified in other regions of the C4 genes (see Chapter 5).

Of the 24 different AHs studied, fourteen carried unique SNPs within the C4 gene region and in addition to this, two had unique combinations of SNPs that allow identification of these AHs. For example, the 8.1 AH had an A at nucleotide position 6259 and a T at position 6528. This combination of SNPs was not present on any other haplotype. As both polymorphic positions were only 269 bp apart, these SNPs may be used as markers to identify the 8.1 AH, for example using an PCR-SSP (Sequence-specific primer) assay. In addition, the 44.3 AH had a unique combination of SNPs at nucleotide positions 7641 and 8174, which would allow identification of this haplotype using an PCR-SSP assay. The other AHs that were found to carry unique SNPs might be identified using one of the various available SNP typing techniques. Haplospecific SNPs are useful markers in disease mapping studies as typing of recombinant AHs for these markers allows the identification of disease susceptibility loci [142]. It has been previously shown, that a number of the AHs included in the present study are associated with various autoimmune diseases. These disease associations are also shown in Table 22.

In addition to the unique SNPs found on various AHs, a number of polymorphisms have been identified that were only present on two or three different AHs. For example, the 44.2, 44.3 and 47.1 AHs share three unique SNPs at nucleotide positions 5656, 7641 and 9273 that were not found on any of the other AHs. In addition, the 44.2 AH also shares three unique SNPs with the 54.1 AH (at positions 5776, 5862 and 5869). The 7.1 and the 52.1 AHs share a T at nucleotide position 6885 and another T at position 8642. Additional polymorphisms, that were only present on a limited number of AHs, have been identified at at least five other polymorphic sites (see Tables 13 and 20). These polymorphisms might be haplotypic, i.e. might be present on all examples of a particular AH. Haplotypic polymorphisms that are found on a number of AHs rather than one particular AH do not allow direct identification of an AH, however, they might be useful in combination with other markers. For example, the 7.1, 8.1 and 52.1 AHs share a unique T at nucleotide position 5902 and a unique A at position 6259. In addition to this, the 7.1 AH and the 8.1 AH carry haplospecific SNPs or combination of SNPs that allow identification of both haplotypes. No such haplospecific marker was found on the 52.1 AH. However, the 52.1 AH might be identified by negative selection versus the 7.1 and 8.1 AHs by typing for one of the SNPs shared by all three AHs and subsequent typing for 7.1 and 8.1 haplospecific markers.

Of the 45 C4 genes studied, remarkably few genes have been found to be identical over the whole region included in the analysis. Apart from the C4A4 and C4B2 genes that were shown to be highly similar (see Section 4.6), only two other pairs of genes were found to be identical. The C4A3 genes of the 18.2 and 58.1 AHs share the complete exonic sequence (see Chapter 5) as well as all intronic regions that were analysed (introns 10, 15, 17 to 27, 32, 34, 36 and intronic sequences flanking the exons, see Chapters 4 and 5). In addition, both haplotypes share a C4BQ0 allele, however, are different elsewhere in the gamma block (i.e. have different alleles at the Bf locus). Interestingly, both haplotypes are associated with insulin-dependent diabetes mellitus (IDDM), although in different populations. The 18.2 AH is found in Caucasians, whereas the 58.1 AH is a Mongoloid haplotype that has been shown to be associated with IDDM in the Chinese population. In addition to this, both C4A genes of the 35.2 AH were found to be identical at the sequence level, even though one of the C4A alleles has been typed as C4A2 and the other one as C4A3 using current C4 allotyping methods [180]. Therefore, the DNA sequences do not explain the differences observed at the protein level between both C4A alleles.

6.3 A haplospecific marker of the 62.1 AH

Sequence analysis of the C4 genes of the 62.1 AH revealed an A nucleotide at the first base of codon 695 (exon 17) that was unique to the C4B3 gene of this haplotype. All other C4 genes included in the present study were found to have a T nucleotide at this position. The A to T substitution in codon 695 results in an amino acid change from a serine residue on the C4B3 allotype to an arginine residue on all other C4 allotypes.

Allele at the C4B locus (codon 695, first base)	C4B3 positive (n=9)	C4B3 negative (n=10)
A	3	0
A/T	5	0
T	1	10

Table 21: Carriage of a 62.1 specific SNP among C4B3 positive and negative individuals. The frequency of the A and T allele at the first base of codon 695 of the C4B genes is shown.

To examine whether all C4B3 genes carry the A allele in exon 17, 9 C4B3 positive and 10 C4B3 negative individuals were tested for the presence of the SNP by sequence based typing. The results of the SNP typing are shown in Table 21. The A allele was present in 8 of the 9 C4B3 positive individuals and in none of the C4B3 negative individuals. The only individual that did not carry the A allele although C4B3 was present, did not have any other markers of the 62.1 AH (i.e. was negative for HLA-B15, HLA-DR4 and HLA-DQ8), raising the possibility that the C4B allele of this individual might have been misreported. Some C4 allotypes are difficult to distinguish using electrophoresis based typing methods since some allotypes show very similar migration rates. For example, the C4B3 and C4A91 allotypes migrate at very similar rates, suggesting that the C4B3 allele could also have been a C4A91 allele which would explain the absence of the A allele in this individual.

Apart from this one exception, the data indicates that the A allele at the first base of codon

695 of the C4 genes is always present when C4B3 is present and is not found on other C4 genes. The C4B3 allotype is unique to the 62.1 AH. The presence of a C4B3 specific marker in exon 17 of the C4 gene therefore allows identification of this AH. The 62.1 AH is a particularly interesting haplotype as several studies have shown that the 62.1 AH is increased in patients with insulin-dependent diabetes mellitus (IDDM) [49–51]. The presence of a 62.1 specific SNP marker in the C4 gene region facilitates studies into the role of the complement genes and other gamma block genes in IDDM.

To facilitate SNP typing of large sample sets, a PCR-SSP assay was developed. Typing by PCR-SSP is an inexpensive, convenient and highly sensitive method, which allows detection of one allele in the presence of an excess of another allele. This is very important as an individual can have up to six C4 genes, of which only one might have the SNP. Sequence specific primers were designed that bind to the A nucleotide allele in exon 17 of the C4B3 gene, and thus, will only amplify in the presence of the A nucleotide at the first base of codon 695. Typing of the T nucleotide allele is not informative as all other C4 genes carry this allele, including the C4A3 gene of the 62.1 AH. Therefore, the T allele would be expected to be present in all individuals. The specificity of the SSP primers was tested using a number of C4B3 positive and negative samples. The results are shown in Figure 18, and demonstrate that the A nucleotide specific primers only amplify in the presence of the C4B3 allele.

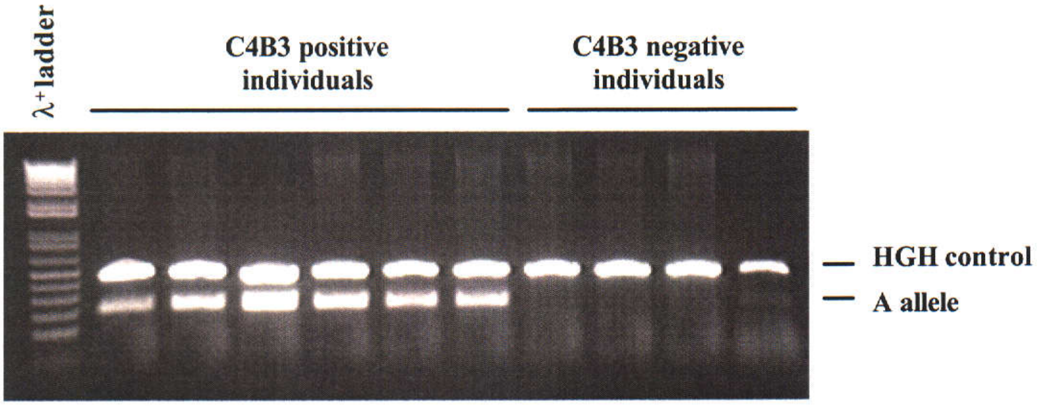

Figure 18: SSP-PCR typing for a 62.1 specific SNP marker. Typing for the 62.1 specific SNP (A allele at the first base of codon 695 of the C4B3 gene) was performed using sequence specific primers (SSP). HGH (Human growth hormone) primers were used as PCR control. Typing of C4B3 positive and negative individuals shows that the A specific band is only present when C4B3 is present.

6.4 Frequency of a 62.1 specific marker in IDDM patients and controls

The 62.1 AH is marked by the presence of HLA-DR4/DQ8, which has been previously shown to confer a high risk of diabetes in Caucasians. Although the MHC class II genes are known as the strongest genetic determinants contributing to disease susceptibility [155, 160, 166], they do not explain all MHC associations with IDDM. Several studies have suggested a role of central MHC genes in susceptibility to disease [3, 51, 154]. To examine the possible role of the complement C4 genes in diabetes, we have typed a control group (n=65) and two IDDM patient groups (n=58) for a 62.1 specific SNP marker in the C4 gene region. The results are shown in Figure 19.

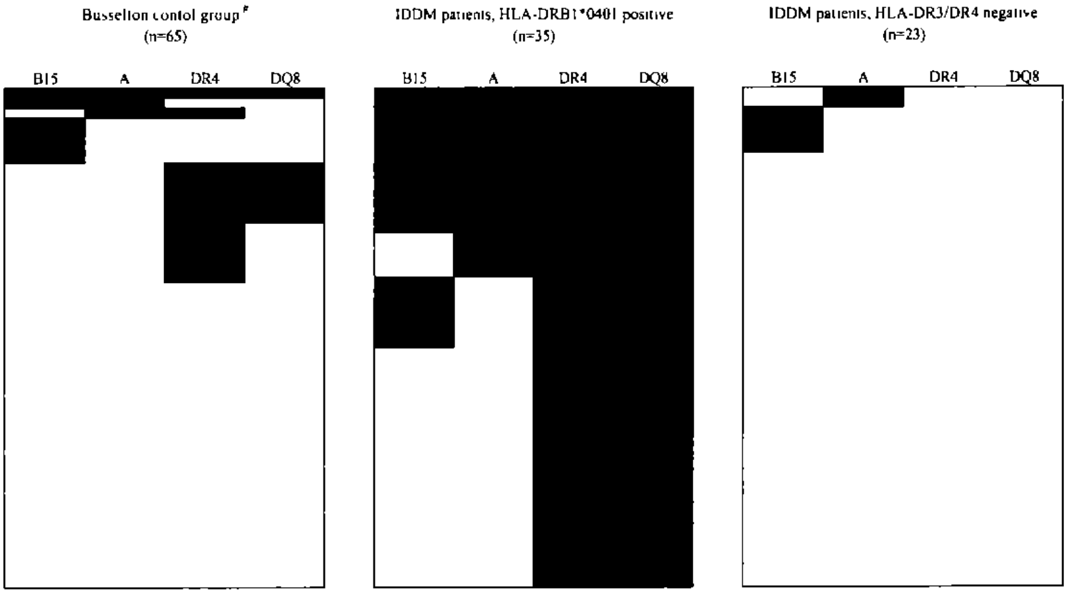

Figure 19: Markers of the 62.1 AH in diabetes patients and the Busselton control population. The presence of various markers of the 62.1 AH is indicated by shading. The vertical axis shows the percentage of patients that carry components of the 62. AH. "A" indicates the presence of an adenine in codon 695 (first base) of the C4 genes.

The presence of the HLA-B15 allele could not be excluded in one HLA-DR4-DQ8 positive patient and in 33 HLA-DR4-DQ8 negative patients.

SNP typing of the control group (n=65) revealed that the 62.1 specific C4 allele was present in 3 individuals (4.6%). Of the 3 individuals, one carried the complete 62.1 AH (including HLA-B15, DR4, DQ8), one carried HLA-B15 in addition to the 62.1 specific C4 marker and one carried HLA-DR4 in addition to the 62.1 SNP. Of the 62 individuals that did not have the 62.1 specific SNP, 6 individuals carried the telomeric end of the 62.1 AH (characterised by HLA-B15) and 8 individuals carried the centromeric end of the haplotype (marked by the presence of HLA-DR4, DQ8).

Typing of the 62.1 specific SNP was also performed in two IDDM patient groups. The first group of IDDM patients (n=35) was selected on the presence of HLA-DRB1*0401, which is a marker of the 62.1 AH but in addition is also found on the 44.1 and 62.2 AHs. The 62.1 specific SNP was found in 13 patients (37.1%), of which 10 carried the complete 62.1 AH from HLA-B to HLA-DQ and 3 carried the centromeric end of the 62.1 AH. HLA-B15, DR4, DQ8 was present in 5 patients that did not have the 62.1 specific SNP. These 5 patients are most likely to carry the 62.2 AH which shares the HLA-B, DR and DQ alleles with the 62.1 AH but has different alleles in the central region of the MHC (e.g. has C4A4 and C4B2 as compared to C4A3 and C4B3 on the 62.1 AH), however, could also be recombinant 62.1 AHs that lack the central MHC of this haplotype. Seven of the patients that were negative for the 62.1 specific SNP had HLA-B44, suggesting that these patients carry the 44.1 AH which is also characterised by HLA-DRB1*0401. In addition, 23 IDDM patients that were HLA-DR3/DR4 negative were typed for the 62.1 specific SNP. Typing revealed that the SNP was present in one patient, who did not carry any other markers of the 62.1 AH. Moreover, HLA-B15 was present in two patients. All other patients carried different AHs.

The typing results of the HLA-DR3/DR4 negative IDDM patient group suggest that the complement region of the 62.1 AH does not contribute to disease susceptibility independently of HLA-DR4 as the 62.1 AH specific SNP was only found in one patient. The patients were selected on the absence of HLA-DR3/DR4 as the HLA-DR4, DQ8 and HLA-DR3, DQ2 class II genes are known to confer the highest risk of diabetes in Caucasians [155,160,166]. However, these genes were absent in the HLA-DR3/DR4 negative IDDM patients, indicating that other factors, possibly located in the central region of the MHC, must contribute to disease in these patients. From the present data it appears that the complement region of the 62.1 AH does not contribute to disease independently from HLA-DR4, however, larger numbers of patients and a control group selected on the same criteria are required to confirm these results. The typing results of the HLA-DR4 positive IDDM group are highly relevant as they show the frequency of the 62.1 AH specific SNP, and with that, the frequency of the 62.1 complement region in a diabetic group. However, further studies, including the characterisation of a HLA-DR4 positive control population, are required to show whether the complement region of 62.1 AH contributes to disease in HLA-DR4 positive patients. Identification of recombinant 62.1 AHs haplotypes would facilitate disease mapping studies aimed at the characterisation of disease susceptibility loci outside the MHC class II region. In the present study, a 62.1 specific SNP marker located in the C4 gene region was characterised. The results obtained from the present work enable future studies to include the C4 gene region in disease mapping approaches, which would elucidate the role of the polymorphic C4 genes and other gamma block genes in susceptibility to diabetes.

6.5 Summary

Characterisation of the complement C4 genes revealed a number of points that may be highly relevant for further investigations of MHC disease associations:

1. A number of AHs carry haplospecific polymorphisms within the C4 genes. Many of these AHs have been shown to be associated with autoimmune diseases. Therefore, haplospecific polymorphisms located in the C4 gene region may serve as markers in disease mapping studies.
2. Apart from the C4A4 and C4B2 genes, very few C4 genes were found to be identical at the sequence level. The only other examples were the C4A3 genes of two diabetogenic AHs (18.2 and 58.1) and the two C4A genes of the 35.2 AH.
3. A 62.1 haplospecific SNP marker was identified in exon 17 of the C4B3 gene. Typing of C4B3 positive and negative individuals confirmed that the SNP is haplospecific.
4. The 62.1 AH has been previously shown to be associated with IDDM. The characterisation of a haplospecific marker in the C4 gene region of the 62.1 AH enables future studies to include this region into disease mapping approaches.
5. The frequency of the 62.1 specific SNP was shown for a control population and two IDDM patient groups. Further studies are required to elucidate the role of C4 in susceptibility to diabetes.

MHC					UTR	I2	E9	E11	E12	E12	I12	I13	I13	E17	E17	I17	E19	I19	I19	I20	E21	I21	E22	I23	E26	E28	I30	E33	E33	E34	E36		
C2	Bf	C4A	C4B	AH	-165	296	2383	3159	3455	3490	3678	4001	4010	5129	5175	5309	5585	5743	5763	6093	6259	6528	6595	6880	7641	8174	9240	10644	10676	11006	12689	Disease association of the AH	Reference
C	S	3	1	7.1																							G					SLE, MS, CD, HM	[40]
C	S	3+3	1	7.2																													
C	S	Q0	1	8.1																	A	T										IDDM, SLE, MG, IgAD...	[46,47]
C	S	3	1	13.1																									C			PV	[48]
Q0	S	4	2	18.1									T																			SLE, C2D	[40]
C	F1	3	Q0	18.2																												IDDM (Sar)	[40,49–51]
C	F	2+3	Q0	35.2																												HIV rapid progression	[40,52]
	S	2	1	38.1																													
C	F	12+91	Q0	42.1			T																					C			T		
C	S	3+3	Q0	44.1																												RA	[40]
C	F	3	1	44.2	G														G	A												IgAD, CD	[40]
	S	Q0	1	44.3																					A	G							
C	F	3	1	44.4								T																					
C	S	4	2	46.1		-																										MG (Th,S)	[41]
C	S	4	2	46.2																				A									
C	S	2+3	Q0	52.1																													
C	F	91	Q0	47.1						C							A															21D	[40,53]
C	S	3	5	54.1				T			T					T		A					A									IDDM (J)	[51,54]
	S	4	5	55.1											A																	21D	[53]
C	S	6	1	57.1					T																							IgAD, PV	[40,41,48]
	S	3	Q0	58.1																												IDDM (Ch)	[40,51]
C	S	3	3	62.1										A																		IDDM, RA	[40,49–51]
	S	4	2	62.2																												IDDM	[53]
C	S	2	1+2	65.1																										T		21D, IgAD	[40,53]

Table 22: Haplospecific single nucleotide polymorphisms (SNPs). Various ancestral haplotypes carry haplospecific markers within the C4 genes. As a number of these AHs are associated with immune related diseases, the haplospecific SNPs are useful markers in the mapping of disease susceptibility loci within the MHC. Nucleotide positions are indicated at the top, the AH designation is shown on the left, disease associations of the AH are listed on the right. Note that unique polymorphisms might be present on all C4 genes of the AH or only on part of the C4 genes.

IDDM, insulin-dependent diabetes mellitus, SLE, systemic lupus erythematosus, 21D, 21-hydroxylase deficiency; MG, myasthenia gravis; MS, multiple sclerosis; IgAD, IgA deficiency; RA, rheumatoid arthritis; PV, psoriasis vulgaris, CD, celiac disease; HM, hemochromatosis, C2D, C2 deficiency. Ch, Chinese; J, Japanese, S, Singapore: Sar, Sardinians; Th, Thai.

□ Haplospecific SNPs or combinations of SNPs.

7 Discussion

Human complement component C4 is known as one of the most complex and polymorphic molecules of the complement system. Two isotypic forms with different chemical reactivities have been described. In addition, typing of C4 proteins revealed more than 40 C4 allotypes. Although the characteristics of the different C4 allotypes are well established at the protein level, the sequences encoding the allotypic variants have not been described. The present study was aimed at the elucidation of the molecular heterogeneity of the complement C4 genes and the identification of the sequences which encode the protein variants. The C4 genes from a range of different ancestral haplotypes (AHs), representing various C4A and C4B allotypes, were examined. Although only one example of each AH was studied, previous studies have shown that all examples of a haplotype carry identical sequences at all loci tested [40,42].

Different AHs vary in the number of C4 genes, with most AHs carrying two but many carrying one or three C4 genes [44]. The AHs included in the present study were chosen to represent a range of different arrangements of the C4 genes. Most of the AHs studied contained two C4 genes, however, two AHs carried three C4 genes (7.2 and 65.1 AHs) and a number of AHs carried only one C4 gene. Several of these AHs contained C4A or C4B null alleles, characterised by the absence of gene product. C4 null alleles may be due to the presence of single C4A or C4B genes (as on the 18.2 and 8.1 AHs, respectively), or may result from the expression of identical C4 isotypes or allotypes on AHs with two C4 loci (as on the 42.1 and 52.1 AHs) [83]. Previous studies have shown that non-expressed genes (pseudogenes) have a frequency of less than 1% [94]. Analysis of our samples suggested that the C4 gene copy number of a haplotype can be estimated by quantitative sequence analysis. Quantitative sequences were found to correlate with the number of C4 genes indicated in the description of every individual AH. Quantitative sequences of AHs with C4 null alleles provided no evidence for the presence of non-expressed C4 genes at the second locus. In addition, none of the previously described insertions or deletions detected in C4 pseudogenes [119–121] were identified in our samples. Thus, the C4 null alleles present on some of the AHs studied were most likely due to gene deletions rather than the presence of pseudogenes.

Analysis of the isotypic sequences in exon 26 confirmed that C4A genes encode Pro^{1101}, Cys^{1102}, Leu^{1105}, Asp^{1106} while C4B genes encode Leu^{1101}, Ser^{1102}, Ile^{1105}, His^{1106}. These four amino acid residues have been shown to account for the differences in functional activity of both C4 isotypes [91]. The activated form of C4A has a high binding affinity towards amino group containing substrates, whereas C4B preferentially reacts with hydroxyl group containing substrates. Further variation comes from the Rodgers and Chido antigenic determinants detected on red blood cells [99]. All together, two Rodgers, six Chido and one rare determinant called WH have been described. The sequences encoding these determinants are located in exons 25, 26 and 28. Analysis of our C4 genes confirmed that Rodger antigens are found on C4A isotypes and Chido antigens on C4B. However, there were a number of C4 allotypes that showed reversed antigenicity (C4A12 and C4A91) or partially reversed antigenicity (C4A2, C4A3 and C4B1). In addition, Rodgers and Chido analysis revealed that some C4 allotypes such as C4A2, C4A3 and C4B1 are not homogeneous and may be split into subtypes, as has been previously shown for other C4 allotypes [95].

The C4d region of the C4 α-chain, which encodes the residues accounting for the isotypic and Rg/Ch variation, is known as the most polymorphic region of C4. However, a number of polymorphic amino acid residues are encoded outside this region. Our analysis identified four additional amino acid changes in the α-chain (outside the C4d region) and another 4 amino acid changes within the β-chain. However, no polymorphic amino acid residues were identified within the γ-chain, indicating that this region is conserved between different AHs.

Comparison of C4 gene sequences (including exonic and intronic regions) revealed an even higher degree of polymorphism than has been observed at the protein level. All C4 allotypes that were included with more than one example from different AHs could be split at the sequence level due to the presence of different alleles or combinations of alleles at polymorphic sites. For some C4 allotypes as many as 7 or 8 unique sequences were identified. All together, analysis of 3.3 kb of sequence revealed 35 sequence variations. The MHC is well known for its extreme polymorphism. Especially the coding regions of the MHC class I and II genes show unusually high levels of nucleotide variability [35, 185]. However, studies of the central region of the MHC have also indicated nucleotide variabilities greater than 1% [35], which is much higher than observed elsewhere in the genome [186, 187].

The structural and functional diversity of C4A and C4B proteins and the molecular heterogeneity of the C4 genes can be largely attributed to gene rearrangements, duplications, deletions and conversions which occur frequently in the C4 gene region [114, 115, 175, 188]. The prevalence of different RCCX length variants increases the frequency of genetic recombination or unequal crossover, contributing to genetic instability and homogenisation of RCCX genes [83, 94, 114, 189]. Misalignments of RCCX modules give rise to duplications and deletions of the RP1, C4A, C4B, CYP21 and TNX genes [130]. However, recombination may also occur within the C4 genes as signal sequences enhancing recombination were identified in the C4d region which may promote the exchange of genetic information among different C4 variants [102]. Gene conversion has been suggested as mechanism generating identical mutations in both C4A and C4B genes [120 188] A number of studies have shown that rearrangements contributing to the C4 polymorphism are ongoing processes. The various polymorphic forms may have created selection advantage during evolution allowing recognition of a wide range of different microbial antigens, however, the rearrangements have also caused deficiency states leading to disease [94, 114].

Phylogenetic analysis, based on both intronic and exonic polymorphisms within 3.3 kb of sequence (roughly the C4d region), revealed that sequences representing the same C4 allotype did not cluster together in most cases. The only exception were the C4A4 allotypes which formed a separate subgroup within the phylogenetic tree. All other C4 alleles were found together with C4 alleles representing different C4 allotypes. These results indicate that polymorphisms observed at the DNA level do not correlate with C4 allotypes defined by serology. However, the phylogenetic analysis revealed that C4A alleles from different AHs fell into one group and the C4B alleles fell into another, even when the residues determining the functional differences between both isotypes were excluded from the analysis. A similar tree showing separation of C4A and C4B alleles was obtained by Kawaguchi et al. [190], comparing exonic C4d sequences from three primate species with human C4d. In addition, comparison of exonic C4d sequences from nine primate species in a study by Paz-Artal et al. [99] suggested trans-species evolution for the C4d region, because alleles belonging to different species were found to cluster together. However, the same studies have shown that phylogenetic trees based on both intronic and exonic sequences demonstrate clustering of alleles from each primate species [99, 190]. These results were interpreted as supporting the theory of extensive homogenisation of the C4 genes in each species.

The sequences were further analysed for the presence of polymorphisms that encode the allotypic variation seen by electrophoresis based C4 typing methods. As differences in electrophoretic mobility may be caused by charge differences of the protein, the C4 gene sequences were analysed for the presence of polymorphisms that may affect the charge. It was found that seven nucleotide polymorphisms located in exons 12, 13, 17, 25, 26 and 28 resulted in an amino acid change involving a charged residue. The amino acids encoded at these positions were compared and the resulting charge differences were calculated. It could be shown that these seven polymorphic residues account for the allotypic variation observed at the protein level. The calculated charges were found to correlate with the electrophoretic mobilities observed by C4 allotyping methods, i.e. the C4 allotypes that migrate fastest in an agarose gel had the highest calculated charges. Using this model, the separation of the most common C4A allotypes (including C4A91, C4A12, C4A2, C4A3, C4A4 and C4A6) and C4B allotypes (including C4B1, C4B2, C4B3 and C4B5) could be explained. It is notable, that only a very small fraction of the polymorphisms observed at the sequence level contributes to the allotypic variation of C4 as the model is based on only seven polymorphic residues. Whether the model can also be applied for the other less frequent C4 allotypes needs to be investigated. The elucidation of the sequences encoding the C4B92, C4B94 or C4B96 allotypes would be of special interest as these allotypes show very slow electrophoretic migration rates (i.e. migrate slower than C4B1).

One of the objectives of the present study was to develop DNA based C4 typing methods as current C4 allotyping methods are technically difficult and do not allow a high throughput of samples. In addition, unambiguous assignment of C4 allotypes may be difficult using current typing methods as some allotypes show very similar electrophoretic mobilities However, characterisation of the polymorphisms encoding the allotypic variation revealed that there are no allotype specific polymorphisms that would allow identification of all individual C4 allotypes. Although some C4 allotypes including C4A4, C4A6, C4B3 and possibly C4B5 carry unique amino acid residues, other C4 allotypes cannot be identified by a single amino acid residue but are characterised by cumulative charge differences resulting in their distinct electrophoretic mobilities. Therefore, DNA based typing methods for the identification of the known allotypic variants would have to include several polymorphic sites, and thus would be rather complex.

However, the present study has provided some insights into the molecular heterogeneity of the complement C4 genes. Most of the polymorphisms described by DNA sequence analysis are not detectable using current C4 allotyping methods, and hence are not included in the current C4 nomenclature. The current nomenclature of C4 is based on the distinction of C4 into two isotypes (C4A and C4B) based on their hemolytic activity and into several C4 allotypes defined by their electrophoretic mobility [191]. It has been previously acknowledged that this nomenclature has many limitations [56]. For example, the association of C4A and C4B allotypes with different Rodger and Chido antigens is not indicated. Therefore, Schneider et al. [192] have proposed a revised nomenclature including several subtypes based on Rg/Ch typing results in addition to the allotype designation defined by electrophoretic typing methods. However, there are still several limitations to this revised nomenclature. The variation in the functional activity of the C4 allotypes is not included. The C4 genes differ in size due to the presence of the HERV-K(C4) insertion in some C4 genes, and may therefore may be long or short, which is not indicated by the current nomenclature. In addition, our analysis has shown that almost all C4 allotypes are heterogeneous and that the polymorphisms observed at the DNA level do not correlate with the C4 allotypes defined by electrophoretic mobility. Taking all this into account, a systematic revision of the C4 nomenclature should be considered.

Sequence comparison of the C4 genes from different AHs also revealed a number of haplospecific single nucleotide polymorphisms (SNPs). SNPs are useful markers in disease mapping studies as typing of recombinant AHs for these markers allows the identification of disease susceptibility loci [142]. It has been previously shown, that a number of the AHs included in the present study are associated with various autoimmune diseases. The 8.1 AH is a particular interesting haplotype as it is associated with several immunopathological disorders, including insulin-dependent diabetes mellitus (IDDM), systemic lupus erythematosus, myasthenia gravis, rapid progression of HIV infection and IgA deficiency [46, 47]. Using recombinant mapping techniques, it has been shown that a number of disease susceptibility genes are located within the central region of the MHC [47, 151–153]. However, this region contains more than 50 genes with immunological and non-immunological functions, and to date very few individual susceptibility genes have been identified. A number of studies have examined the telomeric region of the central MHC, but few studies have included the complement region, despite the high degree of polymorphism found for some complement proteins. It is well established that the MHC is organised in evolutionary "frozen" blocks within which recombination is inhibited, whereas the boundaries of each block have been shown to be hotspots of recombination (see Figure 2, Section 2.1.3). The central region of the MHC is organised in at least two different blocks, between which recombination has been observed [38]. The genes located at the centromeric end of the central MHC (including the complement genes C2, Bf, C4A and C4B) belong to the gamma block, whereas the genes located at the telomeric end of the central MHC (including BAT1 and TNF) belong to a different block. Therefore, haplospecific SNPs present within the C4 genes are important markers as they represent a separate block of the MHC. Recombinant mapping studies using haplospecific SNPs present within the C4 genes would yield insights into the role of C4, but also into the role of other gamma block genes in susceptibility to MHC associated diseases.

A SNP marker has also been identified on the C4B3 gene of the 62.1 AH. Previous studies have shown that the 62.1 AH has an increased frequency in patients with insulin-dependent diabetes mellitus (IDDM) [49–51]. Although this haplotype carries the HLA-DR4, DQ8 genes, which are known to confer a high risk of diabetes in Caucasians [155, 160, 166], several studies suggest that central MHC genes may contribute to disease susceptibility [3, 51, 154]. The identification of a SNP marker in the complement C4 gene region of a diabetogenic AH enables further studies to investigate the role of the gamma block genes in susceptibility to IDDM. In the present study, a PCR-SSP assay was developed to facilitate SNP typing. Using this assay, the frequency of the 62.1 specific SNP marker was shown for a control population and two IDDM patient groups. Although further studies, using a larger patient cohort and an appropriate number of control subjects, are required to examine the significance of the gamma block genes in diabetes, the present study demonstrates a possible approach for future MHC disease mapping studies.

Similar approaches, using respective haplospecific SNP markers present in the C4 gene region, may also be applied for disease mapping studies investigating other MHC associated diseases. Traditionally, MHC disease mapping studies were based on typing for alleles at the classical HLA loci, including HLA-A, HLA-B and HLA-DR. However, these classical markers only represent the alpha (HLA-A), beta (HLA-B) and delta (HLA-DR) block of the MHC, but do not allow disease mapping studies to elucidate the role of the central MHC in susceptibility to disease. Therefore, more recently, polymorphic microsatellites and haplospecific SNP markers have been used in disease mapping studies [39, 151, 153, 168, 170]. SNP markers in the central region of the MHC have, for example, been identified in the BAT1 and TNF genes [27, 149, 150]. However, these genes are located at the telomeric end of the central MHC, outside the gamma block. Few of the genes located in the gamma block have been characterised at the molecular level, resulting in a lack of well characterised molecular markers for this region. The identification of haplospecific SNP markers in the complement C4 gene region enables future studies to include the gamma block genes in MHC disease mapping.

SNPs may also be useful as genetic markers for the identification of high risk haplotypes in clinical settings. For example, it has been shown that the 57.1 AH is associated with hypersensitivity to abacavir [39, 193]. Abacavir is a potent nucleoside reverse transcriptase inhibitor used in combination with other anti-retroviral drugs in the treatment of HIV infected patients [194]. However, in about 5% of patients treated with abacavir a potentially life-threatening hypersensitivity reaction occurs [195, 196]. The 57.1 AH is marked by HLA-B*5701, C4A6, HLA-DRB1*0701 (DR7) and HLA-DQB1*0303 (DQ3). In a study by Mallal et al. [39], the presence of the 57.1 AH had positive and negative predictive values for abacavir hypersensitivity of 100% and 72%, respectively. Recombinant mapping studies have identified the region contained between C4A6 and HLA-Cw6 on the 57.1 AH to confer susceptibility to the hypersensitivity reaction [39]. Therefore, markers of the 57.1 AH would be useful in the development of a highly predictive test for abacavir hypersensitivity, which would allow for a safer use of the drug. One such marker of the 57.1 AH is the C4A6 allotype, which is not found on other AHs. The C4A6 allotype is marked by fast electrophoretic mobility in gel based C4 typing methods, but can also be identified due to the presence of a unique SNP in exon 12 of the gene.

Some of the nucleotide variations identified in the present study result in amino acid changes, and may therefore affect the functional activity of the C4 proteins. For example, the arginine to tryptophan substitution at position 458 of the C4A6 allotype has been previously shown to disrupt the C5 binding site of the C4 β-chain, resulting in hemolytic inactivity of the C4A6 allotype [123]. A similar effect has been found for the proline to leucine substitution at position 459 of a hemolytic inactive C4B1 allotype [197]. Previous studies suggest an important role of C4 in the immune response as C4 deficiency is almost invariably associated with autoimmune or immune complex diseases. For example, it has been shown that C4A null alleles have an increased frequency in patients with systemic lupus erythematosus-like diseases [125, 126, 128]. The C4A isotype functions in immunoclearance through binding to IgG in antibody-antigen aggregates or to antigens of immune complexes. Thus, C4A deficiency may result in impaired clearance of these complexes, which might cause inflammatory damage and lead to autoantibody production [198]. Complete deficiency of C4B is associated with recurrent bacterial and viral infections [130]. C4B is highly reactive towards hydroxyl group containing antigens which are found on enveloped viruses and bacteria covered with capsular polysaccharides [90]. The C4 proteins are part of a cascade of reactions leading to complement activation at sites of infection. Polymorphisms that affect C4 expression levels or alter the functional activity of C4 may therefore have complex effects on the immune response. In the present study, a C4 promoter polymorphism and a number of coding polymorphisms have been described. However, further studies are required to elucidate their possible roles in susceptibility to MHC associated diseases.

8 Summary and conclusions

Complement component C4 is encoded by two highly polymorphic loci, C4A and C4B, located in the central region of the MHC. C4 allotyping based on differences in electrophoretic mobility revealed several different C4 allotypes. The present study suggests a model, based on seven polymorphic amino acid residues, that explains the allotypic variation observed at the protein level. These seven amino acid residues contribute to the charge of the C4 proteins, and hence affect the electrophoretic mobility of the protein variants. Comparison of DNA sequences representing the same C4 allotype but different AHs revealed that most C4 allotypes are heterogeneous and may be split into several subtypes due to the presence of unique polymorphisms or combinations of polymorphisms at the DNA level. The present study has for the first time systematically characterised the extent of nucleotide polymorphism of most common C4 allotypes. The results show that sequence variations do not correlate with the C4 allotypes defined by serology, and that the C4 genes are far more polymorphic than is indicated by the present C4 nomenclature. These findings as well as previous studies suggest that a revised C4 nomenclature will be necessary, which is based on our current knowledge of the structural and functional diversity of the C4 protein variants, the extent of C4 gene polymorphism and the complexity of the C4 gene arrangements.

The localisation of the C4 genes in the MHC and their extensive genetic diversity render C4 an excellent candidate gene for MHC associated disease studies. C4 acts as part of a cascade of reactions leading to complement activation at sites of infection. Genetic deficiency or insufficient regulation of one of the complement components can lead to a defect of the whole system. Partial or complete deficiencies of C4 are almost invariably associated with autoimmune or immune complex diseases. Very few studies have examined the contribution of other C4 alleles to disease susceptibility, probably due to the technical difficulties associated with current C4 allotyping methods. However, the present study has revealed a number of SNPs which allow identification of various C4 allotypes and of several disease associated AHs. Hence, future studies aimed at the elucidation of the physiological role of C4 in susceptibility to disease, are facilitated by the availability of a number of molecular markers in the C4 gene region. Haplospecific SNPs in the C4 gene region are also important markers for MHC disease mapping studies as they represent a separate block of the MHC (i.e. the gamma block). The gamma block contains a number of genes with immunological and non-immunological functions. However, few studies have investigated the significance of these genes in susceptibility to disease, which may in part be related to the lack of characterised molecular markers located in the gamma block. Thus, recombinant disease mapping studies using SNP markers located in the C4 gene region would not only elucidate the physiological significance of different C4 alleles, but would also contribute to the identification of other closely linked genes that may be important in susceptibility to MHC associated diseases.

In summary, this study, aimed at the characterisation of the molecular heterogeneity of the human complement C4 genes, revealed the following:

1. Mutation screening by DNA sequencing allows quantitative sequence analysis. Quantitative sequences reflect C4 gene copy numbers.
2. Denaturing HPLC is a highly sensitive mutation screening method which allows a high throughput of samples.
3. The C4d region of the C4 α-chain is most polymorphic. However, polymorphic amino acid residues are also present in other regions of the α-chain and in the β-chain, whereas the γ-chain is conserved between different C4 allotypes. The highest degree of polymorphism was seen in the introns.
4. Nucleotide polymorphisms do not correlate with the C4 allotypes defined by serology.
5. Most C4 allotypes are heterogeneous at the DNA level and may be split into several subtypes.
6. The C4A4, C4B2 and C4B5 genes of the C4A4 containing haplotypes have probably evolved from a common ancestor.
7. The different electrophoretic mobilities of the C4 allotypes can be explained by cumulative charge differences. Seven polymorphic amino acids have been shown to account for the allotypic variation of the most common C4 allotypes.
8. A number of haplospecific SNP markers are present in the C4 gene region that allow identification of various disease associated ancestral haplotypes.
9. Haplospecific markers within the C4 genes facilitate studies into the role of the gamma block genes in susceptibility to MHC associated diseases. A possible approach was demonstrated for a SNP marker present of a diabetogenic ancestral haplotype.

A Reagents, buffers and solutions

A.1 Reagents

Reagent	Source
Ampicillin	Promega
Bacto Agar	DIFCO Laboratories
Bacto Tryptone	BD Biosciences
Bacto Yeast Extract	DIFCO Laboratories
Big Dye Terminator Cycle Sequencing Kit	Applied Biosystems
Boric Acid	BDH AnalaR
Bromophenol Blue	Bio-Rad
dNTP Set, PCR Grade	Invitrogen Life Technologies
EDTA	BDH AnalaR
Ethidium bromide	Sigma
Expand Long Template PCR System	Roche
Gelatin	BDH AnalaR
Glucose	BDH AnalaR
Hydrochloric Acid (HCl)	BDH AnalaR
IPTG	Promega
JM109 competent cells	Promega
Lambda plus DNA ladder	GIPCO BRL
Magnesium chloride ($MgCl_2$)	Ajax Chemicals
QIAamp DNA Mini Kit	Quiagen
pGEM-T Easy Vector	Promega
Platinum Taq DNA polymerase	Invitrogen Life Technologies
Potassium chloride (KCl)	Sigma
Protease K	Quiagen
Sodium acetate	Applied Biosystems
Sodium chloride (NaCl)	APS Finechem
Sucrose	Boehringer Mannheim
TempliPhi DNA Amplification Kit	Amersham Biosciences
Trizma Base	Sigma
UltraClean PCR Clean-up Kit	MO BIO Laboratories
Ultra Pure Agarose	Invitrogen Life Technologies
X-Gal	Promega

A.2 Buffers and solutions

1% Agarose gel

4.5 g	Ultra Pure Agarose
450 ml	0.5 × TBE buffer
170 µl	1 mg/ml ethidium bromide

The agarose was dissolved in 0.5 × TBE buffer in the microwave and 170 µl of ethidium bromide were added. Agarose gels were stored at room temperature or at 60°C for immediate use.

40 mM dNTP

200 µl	100 mM dATP solution
200 µl	100 mM dCTP solution
200 µl	100 mM dGTP solution
200 µl	100 mM dTTP solution
1200 µl	CSL water

200 µl of each dNTP were combined and added to 1200 µl of CSL water. Aliquots of 20 µl and 40 µl were stored at -20°C.

1 mM EDTA

37.2 g EDTA were dissolved in 100 ml Milli-Q water, autoclaved and stored at room temperature.

Elution buffer (1:2)

Elution buffer was diluted 1:2 with CSL water and stored at room temperature.

100 mM IPTG

24 mg IPTG were dissolved in 1 ml of CSL water and stored at -20°C.

Lambda plus DNA ladder

1 ml	Lambda plus DNA Ladder (1 µg/µl)
2 ml	loading buffer type IV
7 ml	CSL water

Lambda plus DNA ladder, loading buffer type IV and water were mixed. 1 ml aliquots were stored at -20°C for long term storage or 4°C for immediate use.

LB medium

4 g	Bacto Tryptone

2 g	Bacto Yeast Extract
2 g	NaCl

Bacto Tryptone, Bacto Yeast Extract and NaCl were dissolved in 300 ml Milli-Q water, pH was adjusted to 7.0 with NaOH and made up to 400 ml with Milli-Q water. The medium was autoclaved and stored at room temperature.

LB plates with ampicillin/IPTG/X-Gal

6 g	Bacto Agar
450 ml	LB medium
400 µl	ampicillin (1 mg/ml)
12-15	85 mm petri dishes

The agar was added to the LB medium and dissolved in a microwave. The medium was autoclaved and allowed to cool to 50°C before ampicillin was added. 30-35 ml of the LB agar medium were poured into each petri dish and left until agar was hardened. Plates were stored at room temperature for up to one week or at 4°C for up to one month. Prior to use 100 µl of 100mM IPTG and 20 µl of 50 mg/ml X-Gal were spread over the surface of each LB-ampicillin plate and allowed to absorb for 30 min at 37°C.

Loading buffer type IV

8 g	sucrose
0.05 g	Bromophenol Blue

Sucrose and Bromophenol Blue were added to 20 ml of Milli-Q water and stirred on a magnetic stirrer. Once dissolved, solution was made up to 200 ml with Milli-Q water. 1 ml aliquots were stored at -20°C for long term storage or at 4°C for immediate use.

1 M Magnesium chloride ($MgCl_2$)

20.3 g $MgCl_2 \cdot H_2O$ were dissolved in 100 ml Milli-Q water, autoclaved and stored at room temperature.

2 M Magnesium chloride ($MgCl_2$)

40.6 g $MgCl_2 \cdot H_2O$ were dissolved in 100 ml Milli-Q water, autoclaved and stored at room temperature.

10 × PCR buffer

0.625 ml	1 M $MgCl_2$
2.5 ml	1 M Tris-HCl pH 8.3
12.5 ml	1 M KCl
0.125 ml	20 mg/ml Gelatin

9.250 ml	CSL water

$MgCl_2$, Tris-HCl pH 8.3, KCl and Gelatin were mixed and filled up to 25 ml with CSL water. 1 ml aliquots were stored at -20°C.

PCR primers (25 pmol/µl)

27 nmol	dry, desalted primer
108 µl	TE buffer pH 8.0

Dry, desalted primers were reconstituted in TE buffer pH 8.0 to a final concentration of 250 pmol/µl and put on a rotor for 30 min. For use in PCR reactions, stock solutions were diluted 1:10 with CSL water. Stock solutions of primers (250 pmol/µl) and PCR primer solutions (25 pmol/µl) were stored at -20°C.

1 M Potassium chloride (KCl)

18.65 g KCl were dissolved in 250 ml Milli-Q water, autoclaved and stored at room temperature.

Sequencing primers (1 pmol/µl)

PCR primers (25 pmol/µl) were diluted 1:25 with CSL water for use as sequencing primers and stored at -20°C.

SOC medium

2 g	Bacto Tryptone
0.5 g	Bacto Yeast Extract
1 ml	1 M NaCl
0.25 ml	1 M KCl
1 ml	2 M $MgCl_2$
1 ml	2 M glucose

Bacto Tryptone, Bacto Yeast Extract, NaCl and KCl were added to 97 ml of Milli-Q water, stirred until dissolved and autoclaved. After cooling to room temperature, $MgCl_2$ and glucose were added, solution was made up to 100 ml with sterile Milli-Q water, filtered and stored at room temperature.

1 M Sodium chloride (NaCl)

14.1 g NaCl were dissolved in 100 ml Milli-Q water, autoclaved and stored at room temperature.

0.5 × TBE buffer

0.5 l of 10 × TBE buffer were mixed with 9.5 l Milli-Q water and stored at room temperature.

10 × TBE buffer

215.6 g	Trizma Base
110 g	boric acid
16.4 g	EDTA

Trizma Base, boric acid and EDTA were weighed into a sterile flask and dissolved in 1.2 l of Milli-Q water with continuous mixing on a magnetic stirrer. Once dissolved, the solution was made up to 2 l with Milli-Q water, autoclaved and stored in the dark.

TE buffer pH 7.5

2 ml	10 ml Tris-HCl pH 7.5
0.4 ml	1 mM EDTA

Tris-HCl and EDTA were mixed with 150 ml Milli-Q water. pH was adjusted to 7.5 with HCl/NaOH. Solution was made up to 200 ml with Milli-Q water, filtered and stored at room temperature.

TE buffer pH 8.0

2 ml	10 ml Tris-HCl pH 7.5
0.4 ml	1 mM EDTA

Tris-HCl and EDTA were mixed with 150 ml Milli-Q water. pH was adjusted to 8.0 with HCl/NaOH. Solution was made up to 200 ml with Milli-Q water, filtered and stored at room temperature.

10 mM Tris-HCl pH 7.5

10 ml of 1 M Tris-HCl were made up to 70 ml with Milli-Q water, pH was adjusted and solution filled up to 100 ml. Storage was at room temperature.

1 M Tris-HCl pH 8.3

30.3 g Trizma Base were dissolved in 250 ml Milli-Q water. pH was adjusted to 8.3 with HCl, buffer was autoclaved and stored at room temperature.

B Phylogenetic analysis

B.1 Phylogenetic tree based on coding polymorphisms

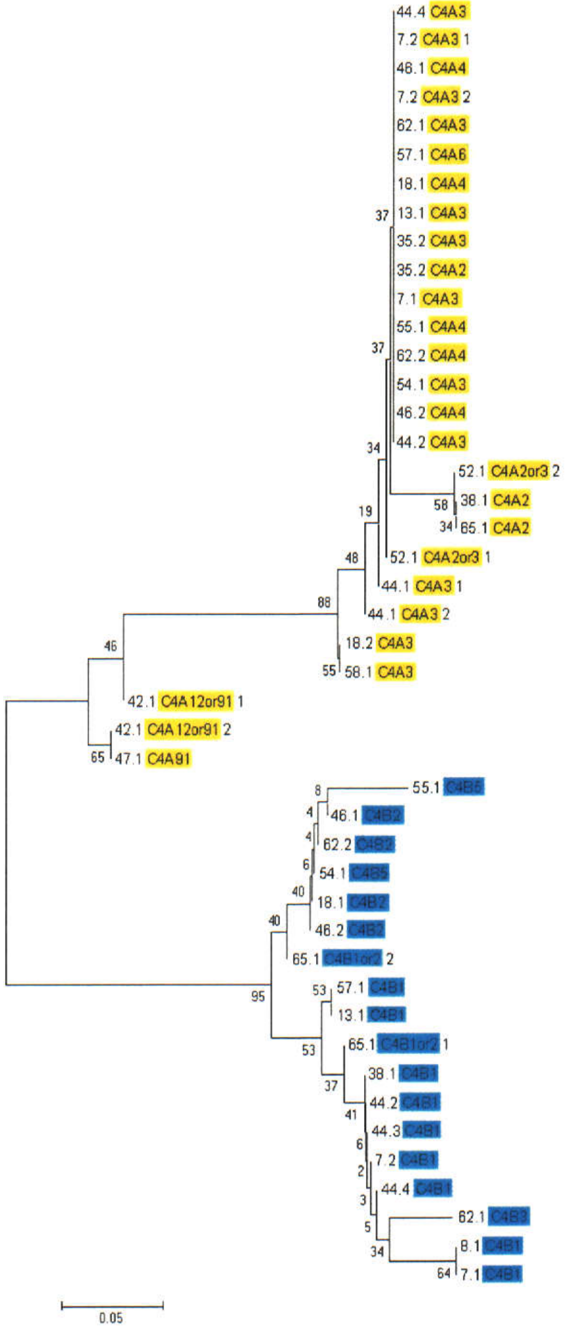

Figure 20: Phylogenetic sequence analysis based on coding polymorphisms. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Coding polymorphisms within 3.3 kb of sequence (spanning from intron 16 to 28) were used to construct the tree. These include the isotype specific sequences and the sequences encoding the Rodger and Chido determinants. The bootstrap percentage from 1,000 replicates is indicated at each node. C4A allotypes are marked by a yellow box; C4B allotypes are highlighted by a blue box.

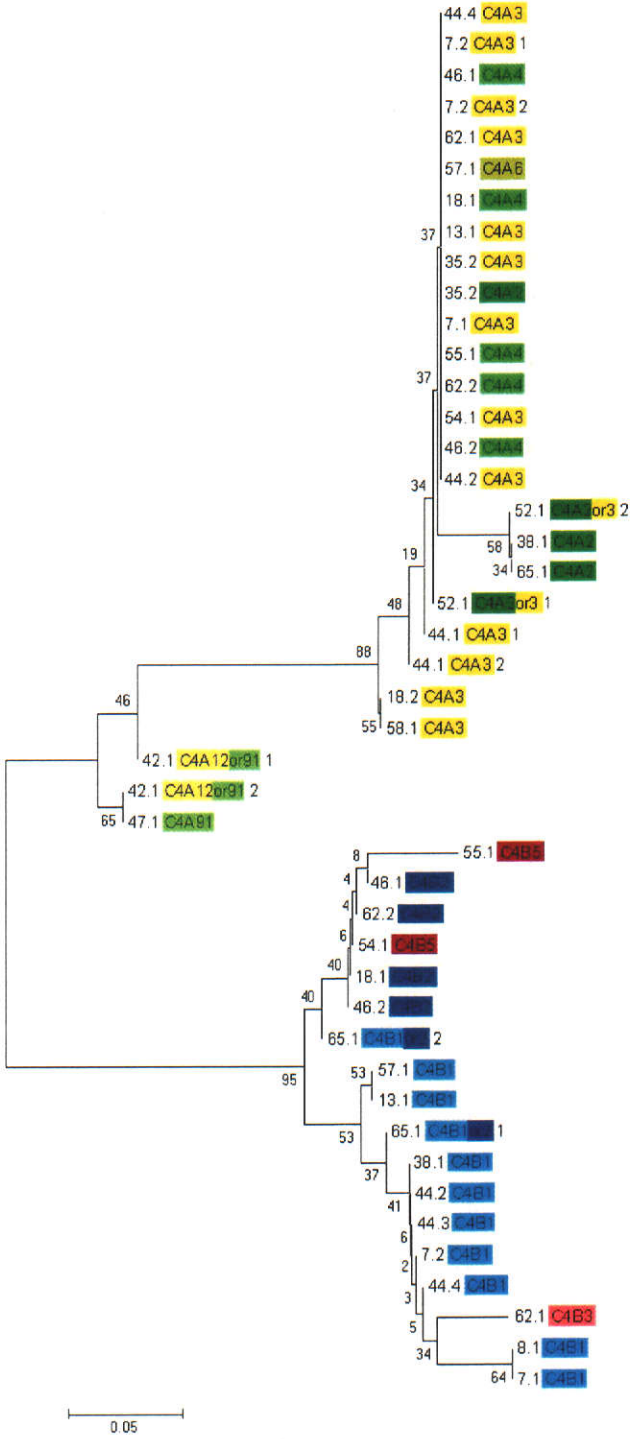

Figure 21: Phylogenetic sequence analysis based on coding polymorphisms. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Coding polymorphisms within 3.3 kb of sequence (spanning from intron 16 to 28) were used to construct the tree. These include the isotype specific sequences and the sequences encoding the Rodger and Chido determinants. The bootstrap percentage from 1,000 replicates is indicated at each node. Each C4 allotype is marked by a different colour to show distribution of individual alleles within the tree.

B.2 Phylogenetic tree based on non-coding polymorphisms

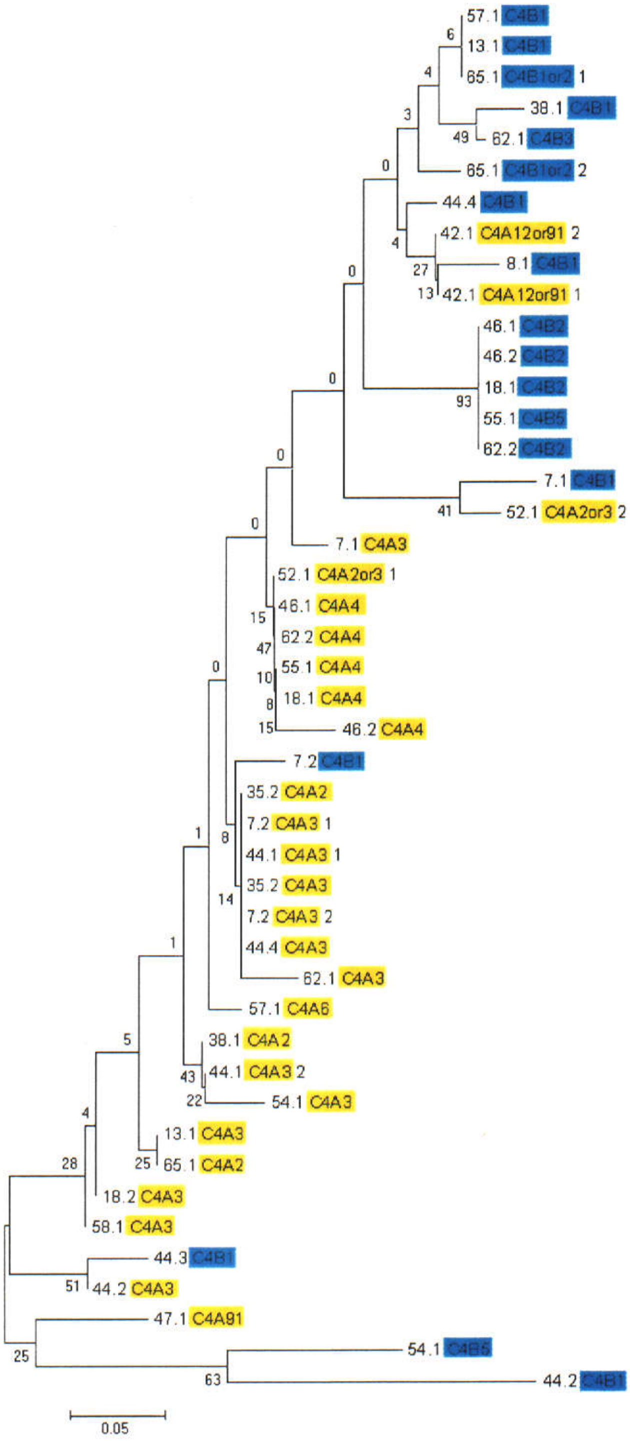

Figure 22: Phylogenetic sequence analysis based on non-coding polymorphisms. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Intronic and synonymous polymorphisms within 3.3 kb of sequence (including 22 single nucleotide substitutions and two 1-bp deletions) were used to construct the tree. The bootstrap percentage from 1,000 replicates is indicated at each node. C4A allotypes are marked by a yellow box; C4B allotypes are highlighted by a blue box.

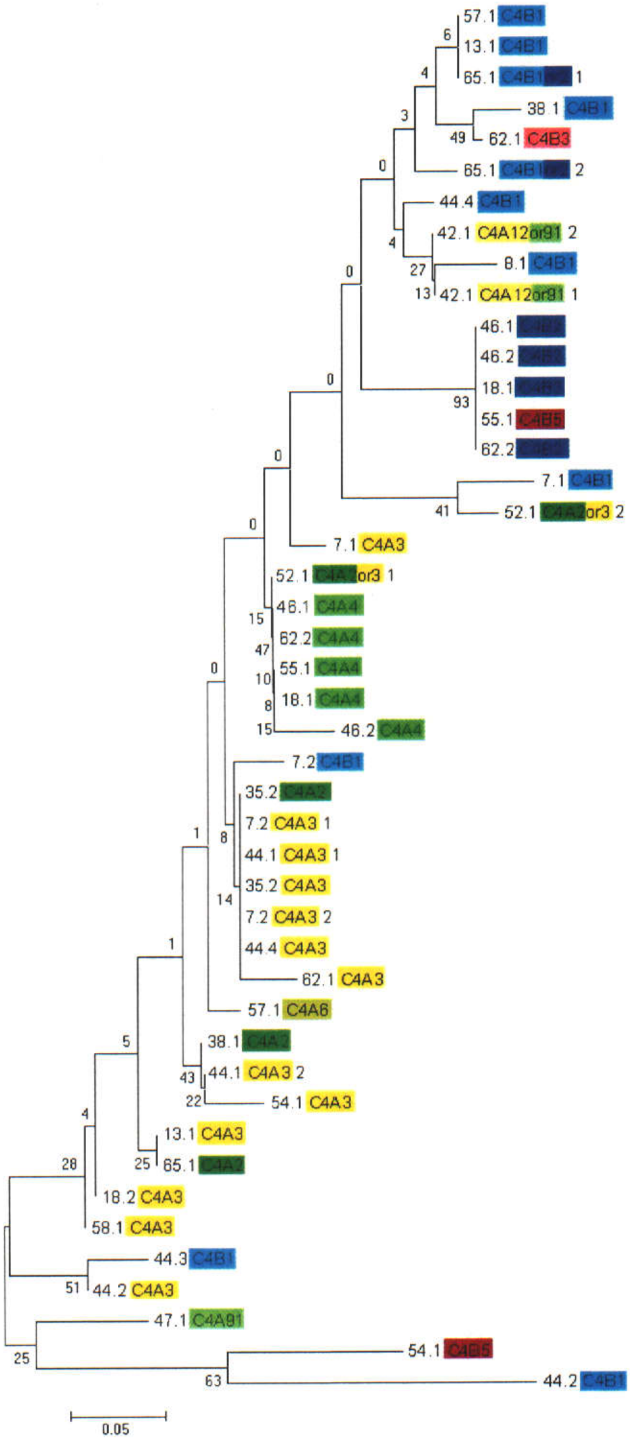

Figure 23: Phylogenetic sequence analysis based on non-coding polymorphisms. Unrooted neighbour-joining tree of human C4A and C4B genes from different AHs. Intronic and synonymous polymorphisms within 3.3 kb of sequence (including 22 single nucleotide substitutions and two 1-bp deletions) were used to construct the tree. The bootstrap percentage from 1,000 replicates is indicated at each node. Each C4 allotype is marked by a different colour to show distribution of individual alleles within the tree.

C Denaturing HPLC

C.1 Non-denaturing analysis

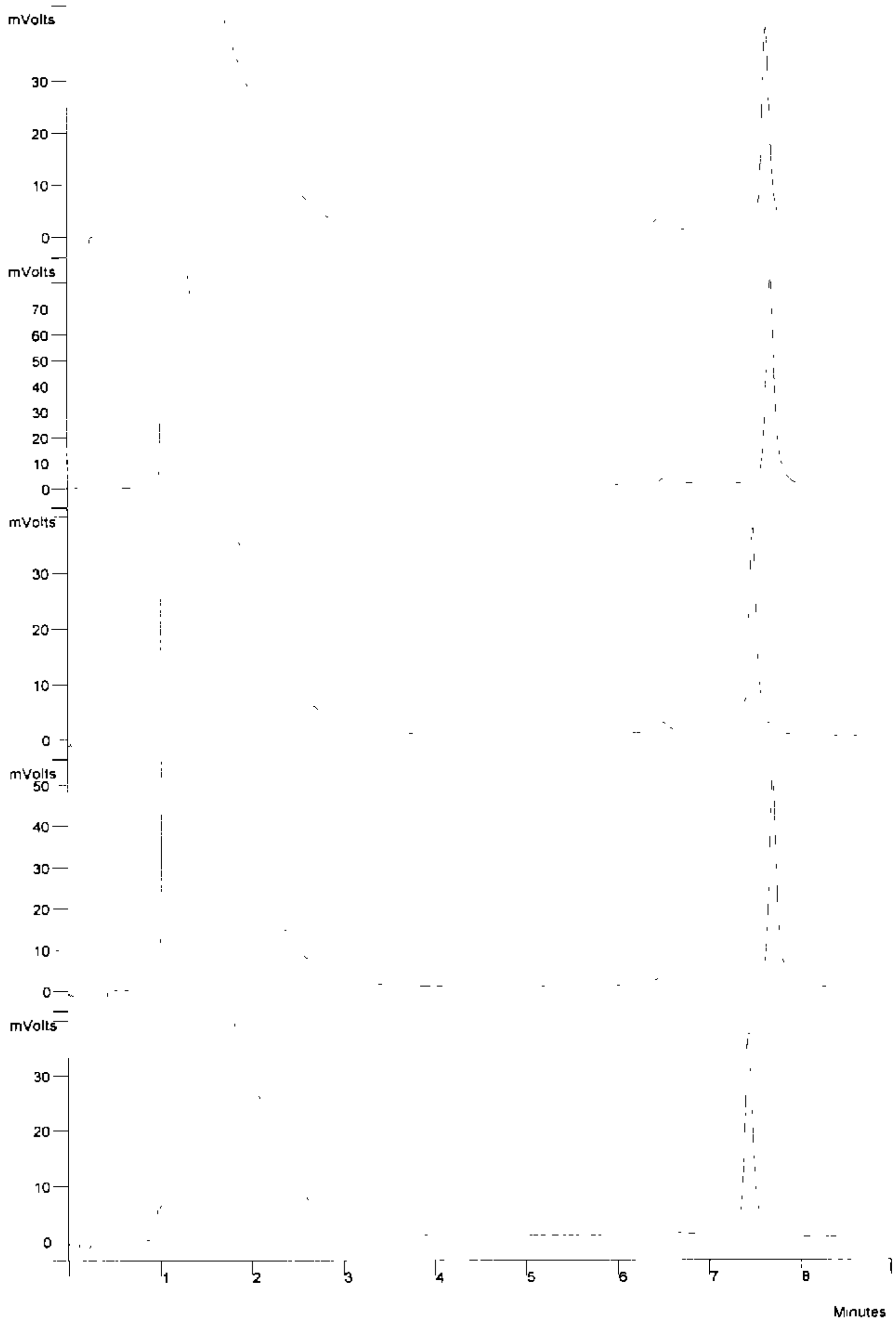

Figure 24: Non-denaturing analysis of samples on the HPLC. To determine size, yield and purity of PCR products, all 23 amplicons of the reference sample (8.1 AH) were analysed under non-denaturing conditions (50°C)
From top to bottom: E1 (375 bp), E2 (349 bp), E3 (330 bp), E4+5 (376 bp) and E6 (319 bp).

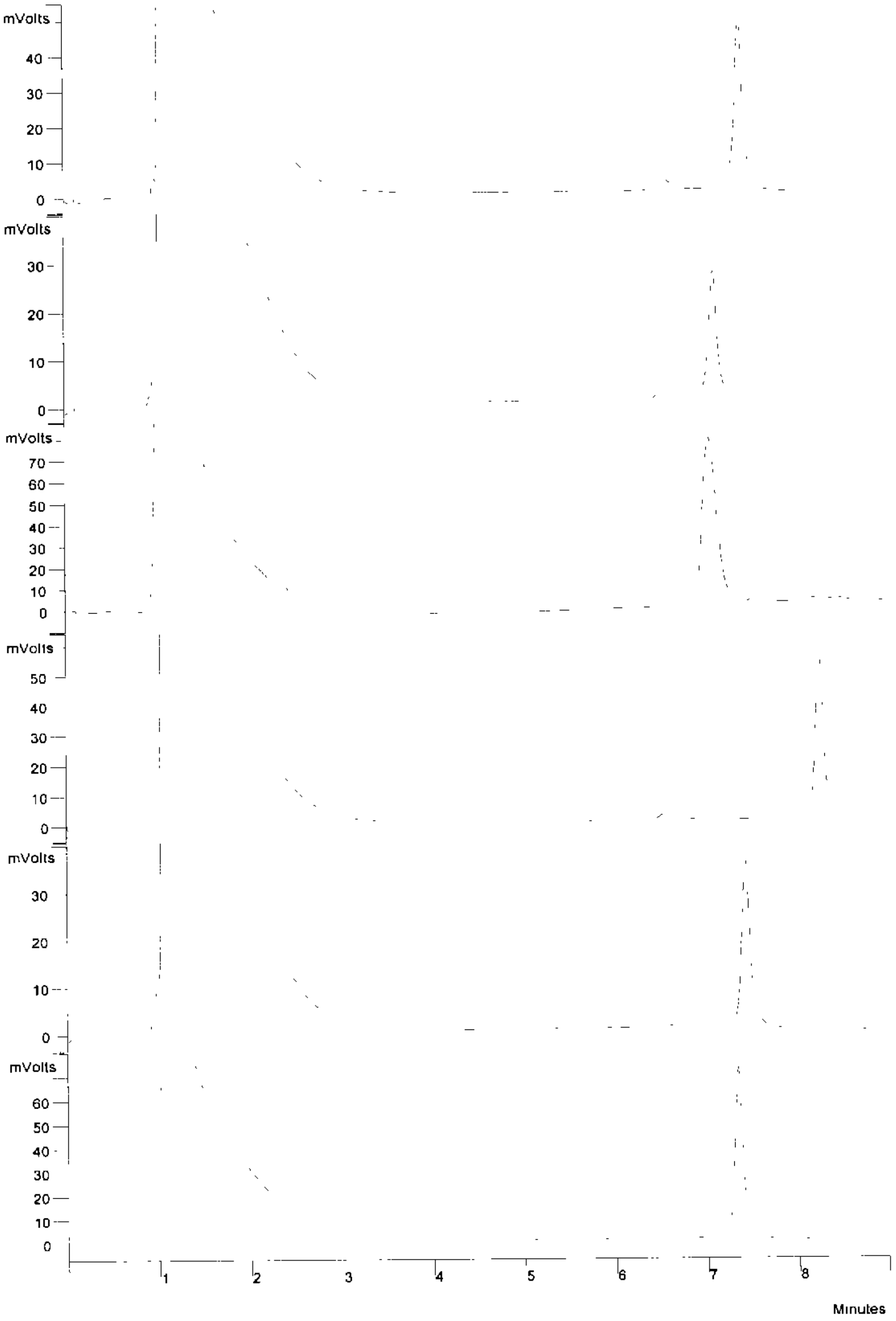

Figure 25: Non-denaturing analysis of samples on the HPLC. To determine size, yield and purity of PCR products, all 23 amplicons of the reference sample (8.1 AH) were analysed under non-denaturing conditions (50° C).
From top to bottom: E7 (297 bp), E8 (254 bp), E9 (342 bp), E10–11 (586 bp), E12 (323 bp) and E13 (303 bp).

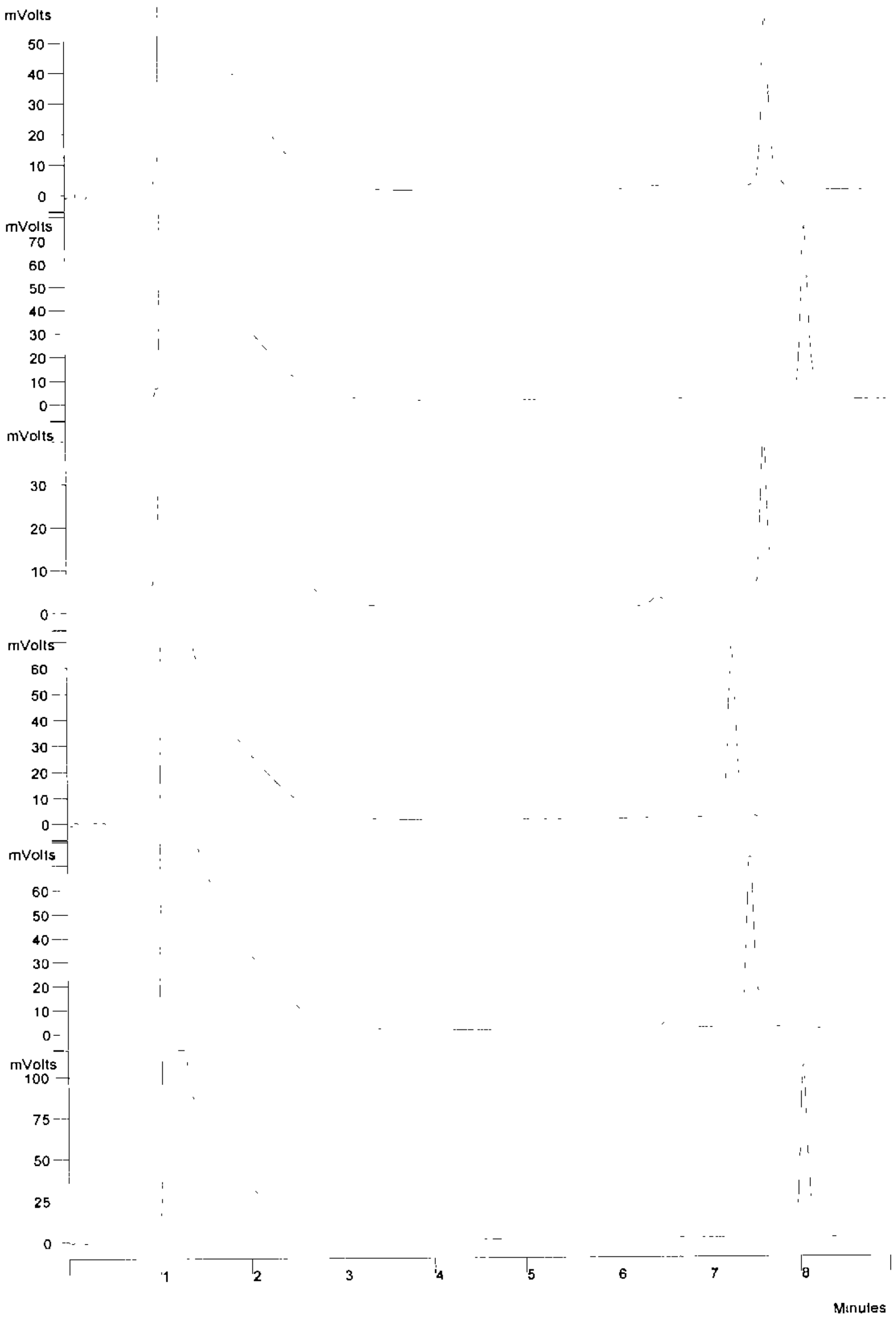

Figure 26: Non-denaturing analysis of samples on the HPLC. To determine size, yield and purity of PCR products, all 23 amplicons of the reference sample (8 1 AH) were analysed under non-denaturing conditions (50°C)
From top to bottom E14 (377 bp), E15+16 (536 bp), E29 (404 bp), E30 (291 bp), E31 (302 bp) and E32+33 (525 bp).

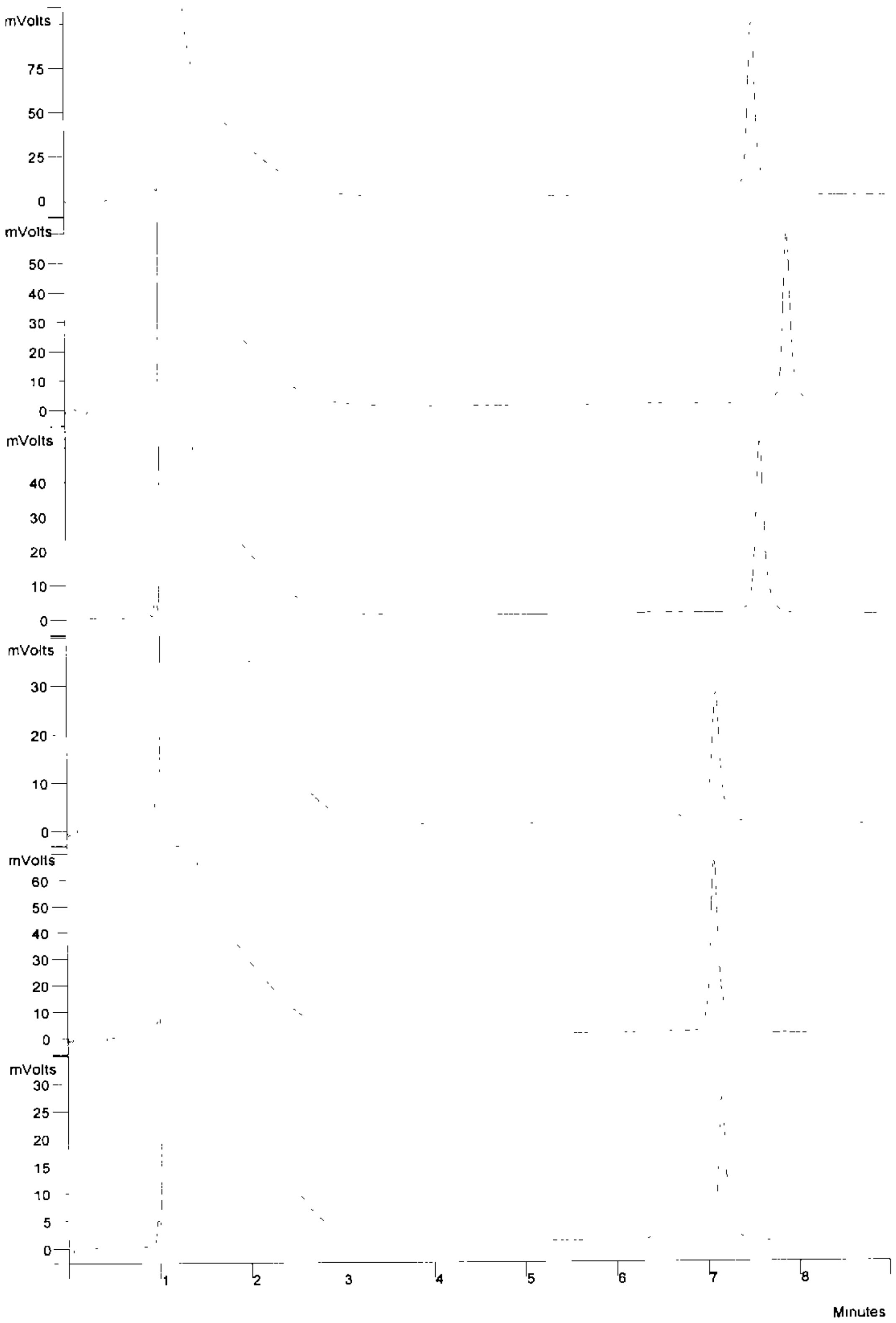

Figure 27: Non-denaturing analysis of samples on the HPLC. To determine size, yield and purity of PCR products, all 23 amplicons of the reference sample (8.1 AH) were analysed under non-denaturing conditions (50°C).
From top to bottom: E34+35 (337 bp), E36+37 (469 bp), E38 (366 bp), E39 (249 bp), E40 (268 bp) and E41 (276 bp)

C.2 Temperature titration

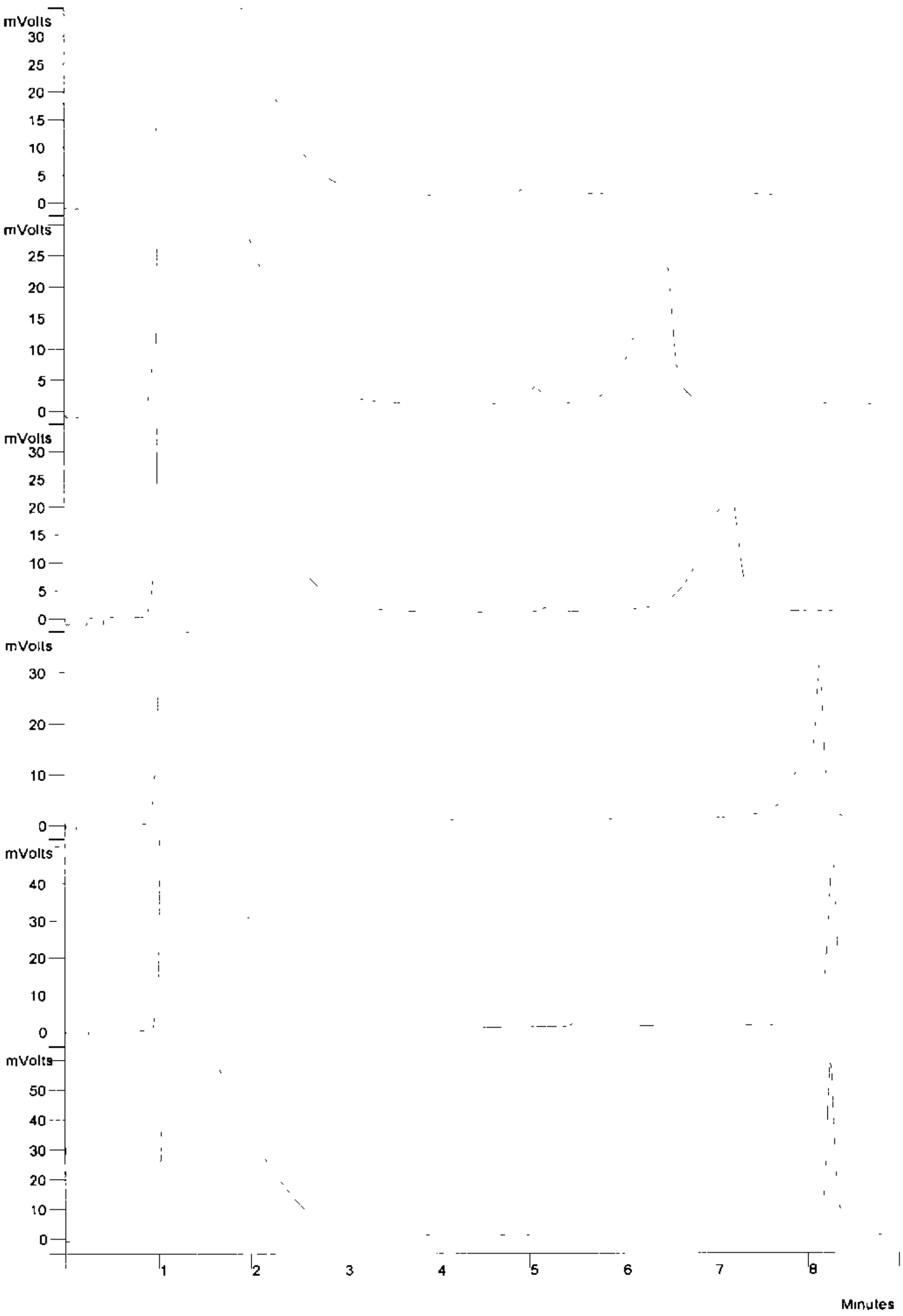

Figure 28: Temperature titration of homozygous DNA. A 586 bp homozygous DNA fragment was analysed at different temperatures ranging from 50°C to 64°C. With increasing temperature, peaks were shifted to shorter retention times and resolution of homoduplex peaks reduced until DNA was completely melted at 64°C.
From top to bottom: 64°C, 63°C, 62°C, 61°C, 59°C and 50°C.

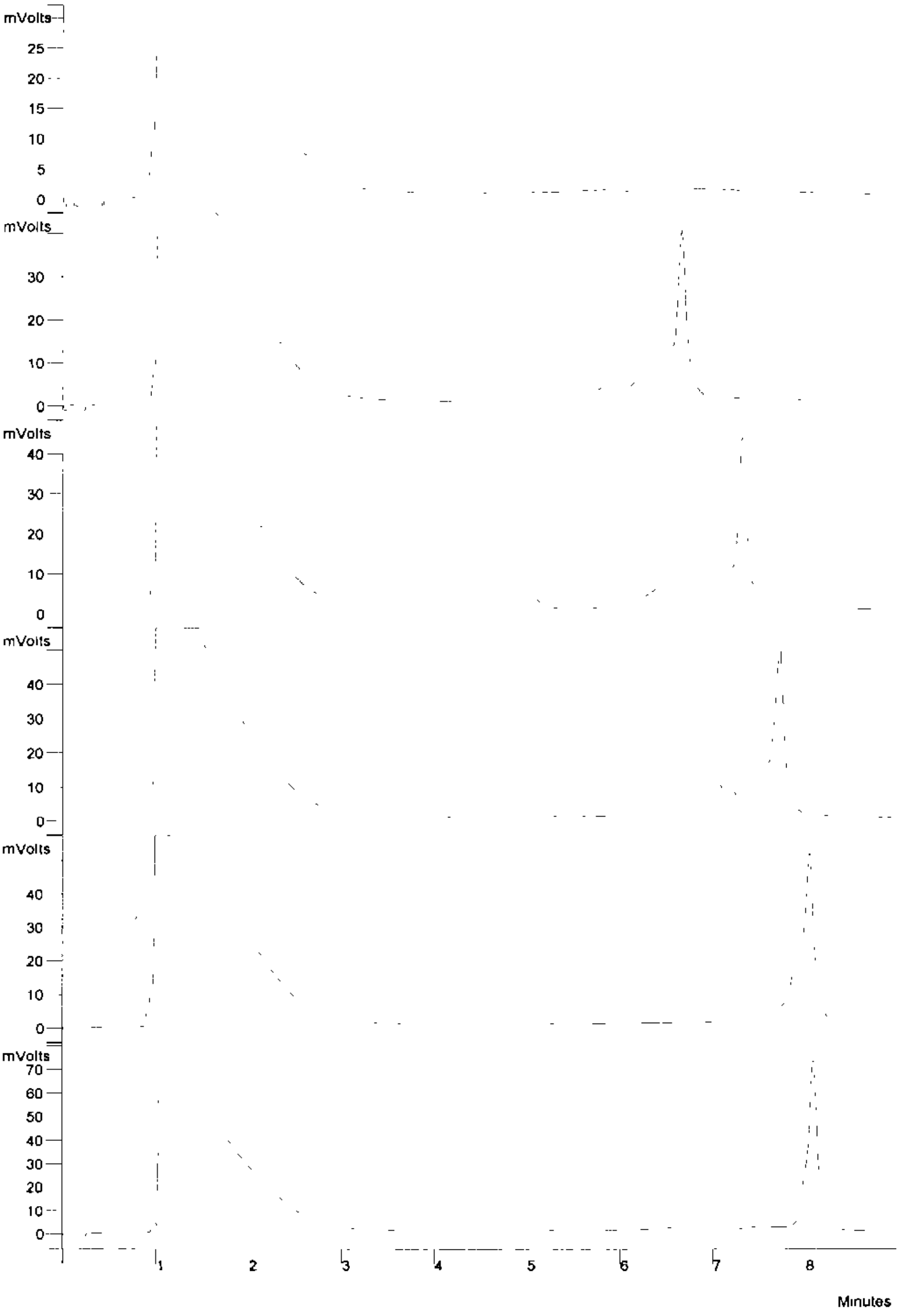

Figure 29: Temperature titration of heterozygous DNA. A 546 bp heterozygous DNA fragment was analysed at different temperatures ranging from 50°C to 65°C to determine the temperature dependent resolution of heteroduplex peaks from homoduplex peaks. From top to bottom: 65°C, 64°C, 63°C, 62°C, 60°C and 50°C

C.3 Reproducibility of results

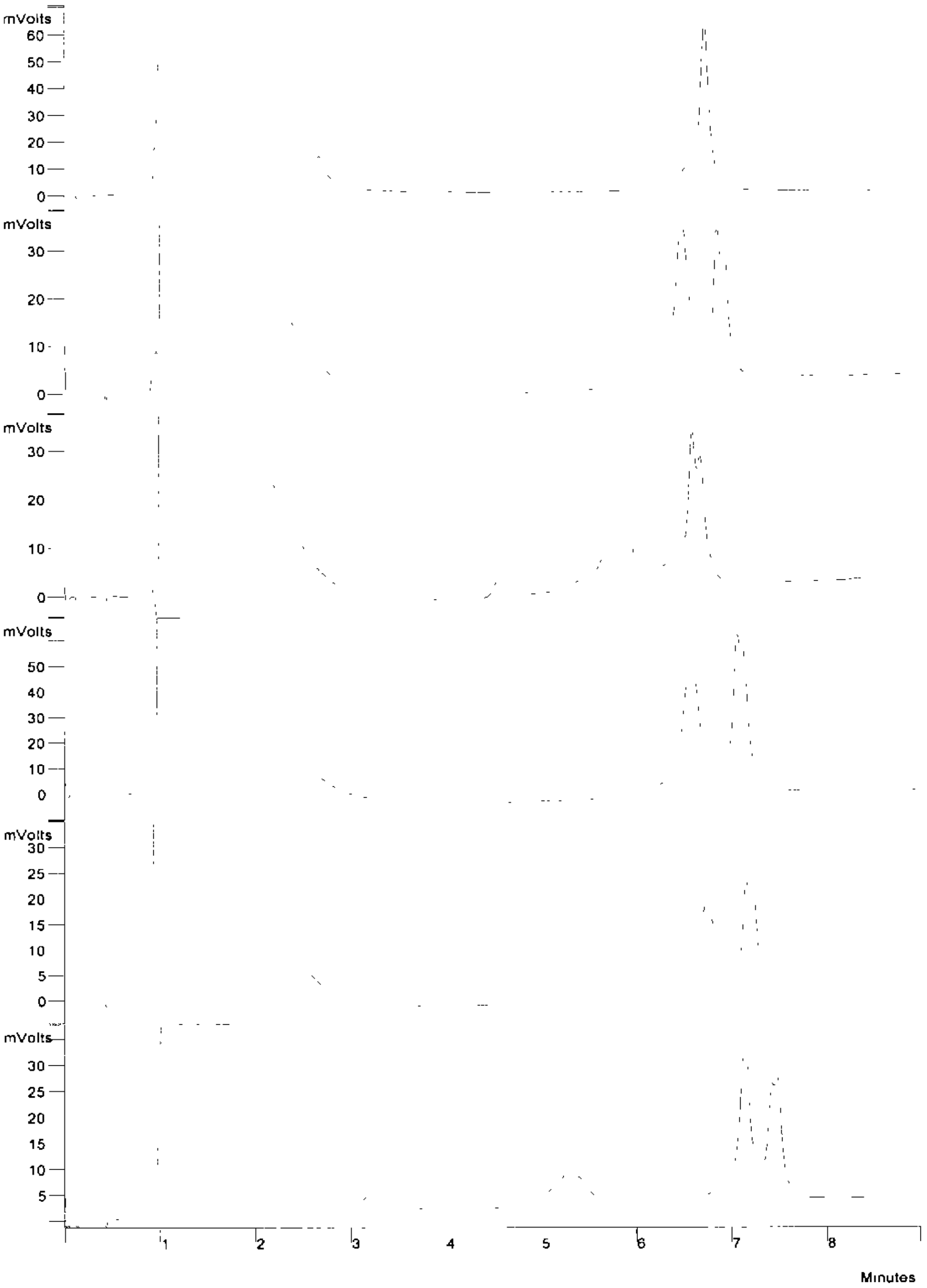

Figure 30: Reproducibility of heteroduplex peaks on the HPLC. To confirm that results were reproducible, a 302 bp heterozygous fragment was repeatedly injected on several runs of the HPLC. Slight shifts in the retention times of the fragment were observed. However, all heteroduplex peaks were significantly different from the homoduplex peak shown at the top

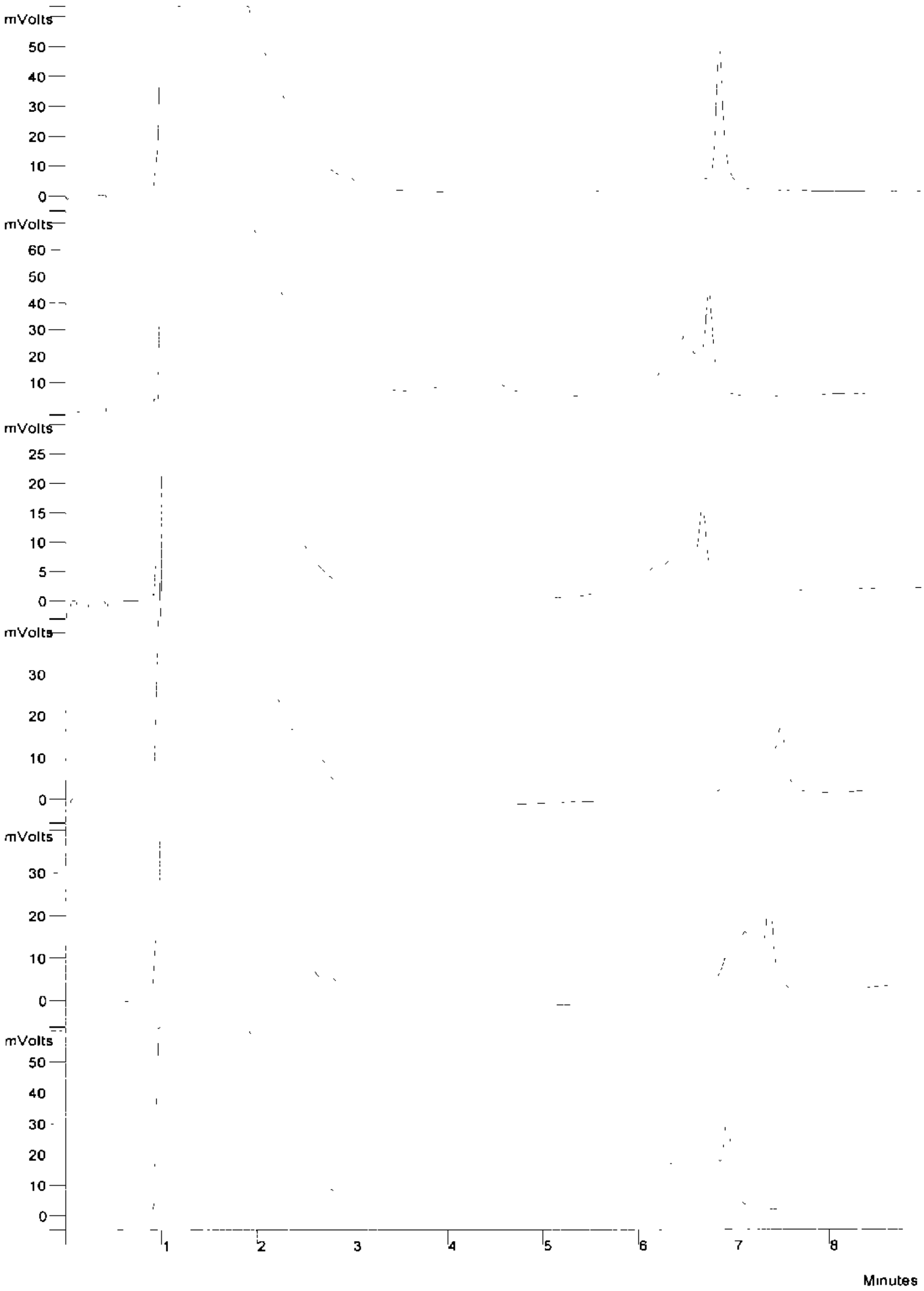

Figure 31: Reproducibility of heteroduplex peaks on the HPLC. Reproducibility was tested for a second heterozygous sample. The fragment size of the sample shown here is 404 bp. As observed for the first sample, retention times varied slightly on different runs, however, resolution of heteroduplex peaks from the homoduplex peak (shown at the top) did not change.

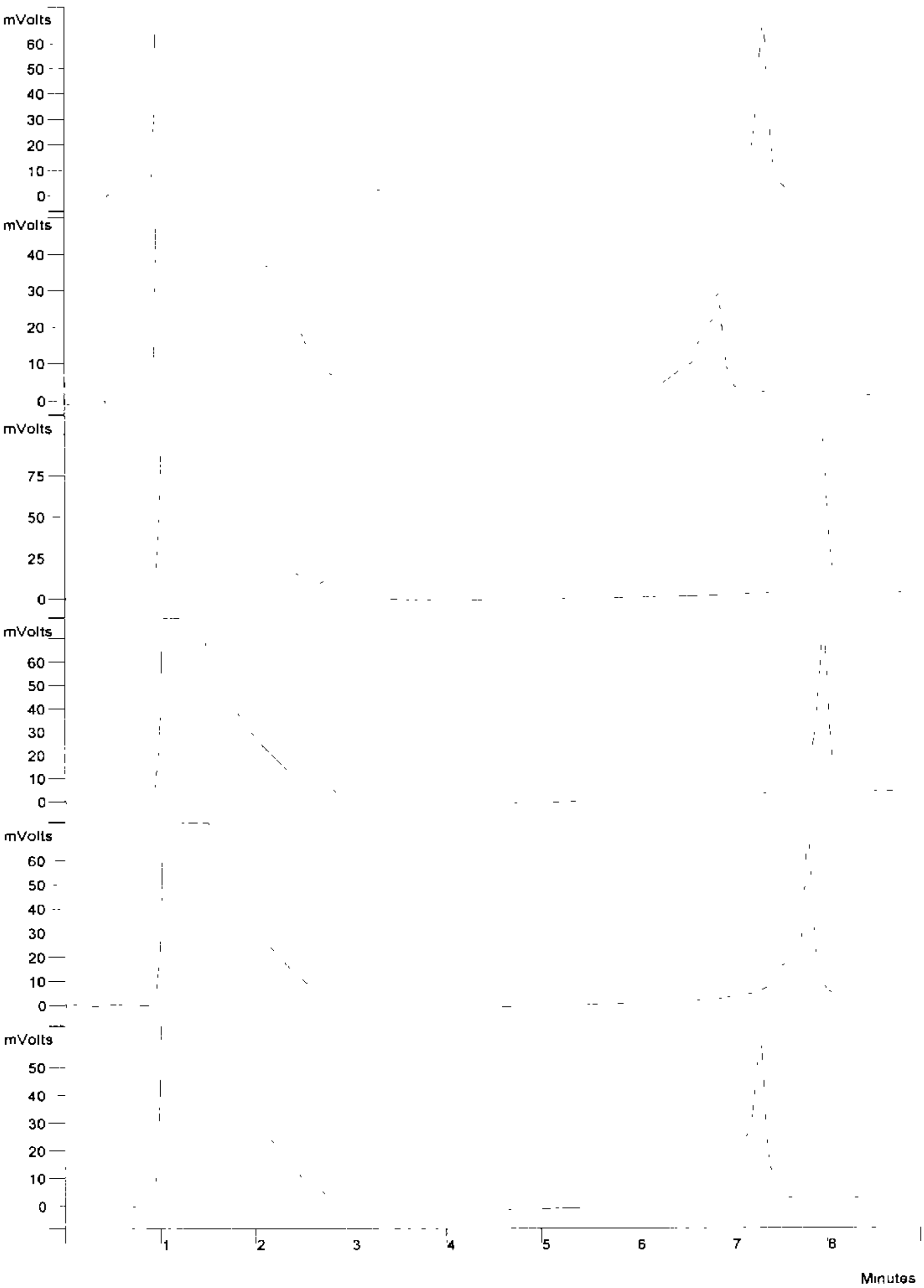

Figure 32: Temperature dependent reproducibility of HPLC results. Analysis of a 525 bp fragment indicated that instabilities of the oven temperature might lead to partial melting of homoduplex DNAs resulting in heteroduplex peak patterns.

From top to bottom: (1) homozygous reference at 63°C, (2) unknown fragment at 63°C showing heteroduplex peak pattern at initial analysis, (3) homozygous reference at 63°C (second run), (4) unknown fragment at 63°C showing homoduplex peak pattern at second run, (5) unknown fragment at 64°C, (6) unknown fragment at 65°C.

C.4 Sensitivity of the dHPLC method

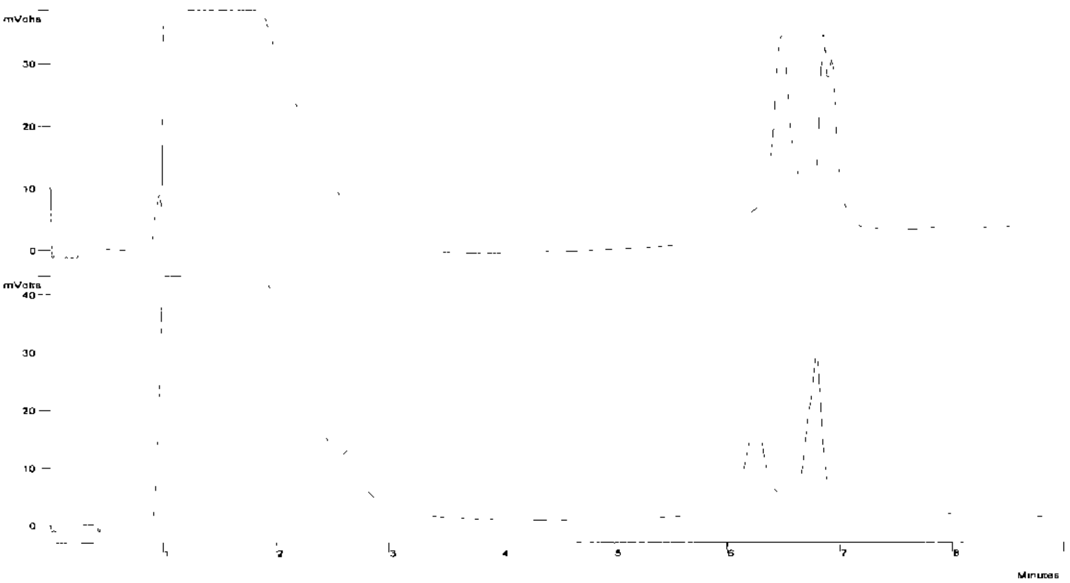

Figure 33: Evaluation of system performance. After a certain number of injections the column showed degradation of performance resulting in low resolution of heteroduplex peaks from homoduplex peaks. Three peaks were resolved in the chromatograms shown at the top, whereas the two peaks at about 7 minutes were not resolved in the second chromatogram, indicating low column quality

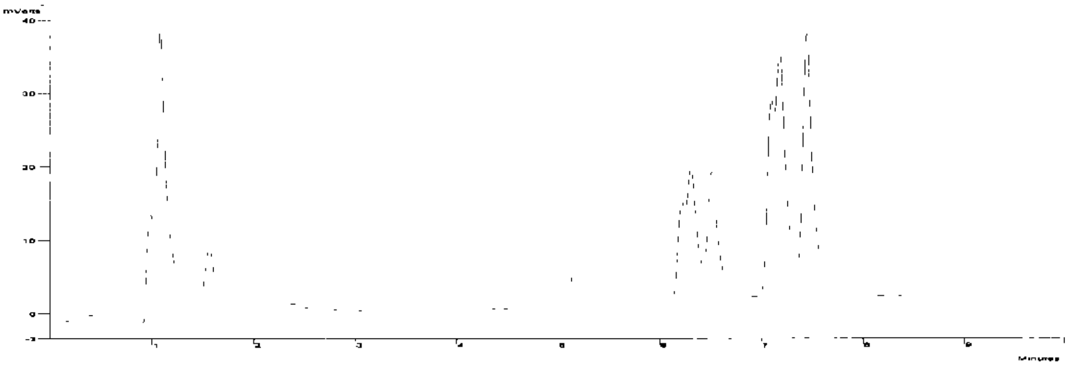

Figure 34: Chromatogram of the HaeIII digested pUC18 size marker. The pUC18/HaeIII size marker was used to evaluate performance of the dHPLC under non-denaturing conditions (50°C) Separation of the 257 bp and 267 bp peaks indicated adequate column resolution The pUC18/HaeIII size marker contains 11 fragments of sizes 11, 18, 80, 102, 174, 257, 267, 298, 434, 458 and 587 bp, 9 of which are detectable by HPLC. The 11 bp and 18 bp are too small to be resolved by the method. The first peak at about 1 minute is the injection peak.

References

[1] Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature*, 401(6756):921–3, 1999.

[2] C. Y. Yu, Z. Yang, C. A. Blanchong, and W. Miller. The human and mouse MHC class III region: a parade of 21 genes at the centromeric segment. *Immunol Today*, 21(7):320–8, 2000.

[3] C. M. Milner and R. D. Campbell. Genetic organization of the human MHC class III region. *Front Biosci*, 6:D914–26, 2001.

[4] J. Klein. The unity of genes in the major histocompatibility complex. *Arthritis Rheum*, 21(5 Suppl):S90–6, 1978.

[5] G. M. Schreuder, C. K. Hurley, S. G. Marsh, M. Lau, M. Maiers, C. Kollman, and H. J. Noreen. The HLA Dictionary 2001: a summary of HLA-A, -B, -C, -DRB1/3/4/5, -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR and -DQ antigens. *Tissue Antigens*, 58(2):109–40, 2001.

[6] A. Maffei and P. E. Harris. Peptides bound to major histocompatibility complex molecules. *Peptides*, 19(1):179–98, 1998.

[7] D. Allan, E. Lepin, V. Braud, C. O'Callaghan, and A. McMichael. Tetrameric complexes of HLA-E, HLA-F, and HLA-G. *J Immunol Methods*, 268(1):43, 2002.

[8] I. Khalil-Daher, B. Riteau, C. Menier, C. Sedlik, P. Paul, J. Dausset, E. D. Carosella, and N. Rouas-Freiss. Role of HLA-G versus HLA-E on NK function: HLA-G is able to inhibit NK cytolysis by itself. *J Reprod Immunol*, 43(2):175–82, 1999.

[9] A. Maffei, K. Papadopoulos, and P. E. Harris. MHC class I antigen processing pathways. *Hum Immunol*, 54(2):91–103, 1997.

[10] J. O. Koopmann, G. J. Hammerling, and F. Momburg. Generation, intracellular transport and loading of peptides associated with MHC class I molecules. *Curr Opin Immunol*, 9(1):80–8, 1997.

[11] J. C. Howard. Supply and transport of peptides presented by class I MHC molecules. *Curr Opin Immunol*, 7(1):69–76, 1995.

[12] 3rd Grandea, A. G. and L. Van Kaer. Tapasin: an ER chaperone that controls MHC class I assembly with peptide. *Trends Immunol*, 22(4):194–9, 2001.

[13] P. Brocke, N. Garbi, F. Momburg, and G. J. Hammerling. HLA-DM, HLA-DO and tapasin: functional similarities and differences. *Curr Opin Immunol*, 14(1):22–9, 2002.

[14] J. R. Gruen and S. M. Weissman. Human MHC class III and IV genes and disease associations *Front Biosci*, 6:D960–72, 2001.

[15] R. Dawkins, C. Leelayuwat, S. Gaudieri, G. Tay, J. Hui, S. Cattley, P. Martinez, and J. Kulski. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev*, 167:275–304, 1999.

[16] Z. Yang, L. Shen, A. W. Dangel, L. C. Wu, and C. Y. Yu. Four ubiquitously expressed genes, RD (D6S45)-SKI2W (SKIV2L)-DOM3Z-RP1 (D6S60E), are present between complement component genes factor B and C4 in the class III region of the HLA. *Genomics*, 53(3):338–47, 1998.

References

[1] Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature*, 401(6756):921–3, 1999.

[2] C. Y. Yu, Z. Yang, C. A. Blanchong, and W. Miller. The human and mouse MHC class III region: a parade of 21 genes at the centromeric segment. *Immunol Today*, 21(7):320–8, 2000.

[3] C. M. Milner and R. D. Campbell. Genetic organization of the human MHC class III region. *Front Biosci*, 6:D914–26, 2001.

[4] J. Klein. The unity of genes in the major histocompatibility complex. *Arthritis Rheum*, 21(5 Suppl):S90–6, 1978.

[5] G. M. Schreuder, C. K. Hurley, S. G. Marsh, M. Lau, M. Maiers, C. Kollman, and H. J. Noreen. The HLA Dictionary 2001: a summary of HLA-A, -B, -C, -DRB1/3/4/5, -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR and -DQ antigens. *Tissue Antigens*, 58(2):109–40, 2001.

[6] A. Maffei and P. E. Harris. Peptides bound to major histocompatibility complex molecules. *Peptides*, 19(1):179–98, 1998.

[7] D. Allan, E. Lepin, V. Braud, C. O'Callaghan, and A. McMichael. Tetrameric complexes of HLA-E, HLA-F, and HLA-G. *J Immunol Methods*, 268(1):43, 2002.

[8] I. Khalil-Daher, B. Riteau, C. Menier, C. Sedlik, P. Paul, J. Dausset, E. D. Carosella, and N. Rouas-Freiss. Role of HLA-G versus HLA-E on NK function: HLA-G is able to inhibit NK cytolysis by itself. *J Reprod Immunol*, 43(2):175–82, 1999.

[9] A. Maffei, K. Papadopoulos, and P. E. Harris. MHC class I antigen processing pathways. *Hum Immunol*, 54(2):91–103, 1997.

[10] J. O. Koopmann, G. J. Hammerling, and F. Momburg. Generation, intracellular transport and loading of peptides associated with MHC class I molecules. *Curr Opin Immunol*, 9(1):80–8, 1997.

[11] J. C. Howard. Supply and transport of peptides presented by class I MHC molecules. *Curr Opin Immunol*, 7(1):69–76, 1995.

[12] 3rd Grandea, A. G. and L. Van Kaer. Tapasin: an ER chaperone that controls MHC class I assembly with peptide. *Trends Immunol*, 22(4):194–9, 2001.

[13] P. Brocke, N. Garbi, F. Momburg, and G. J. Hammerling. HLA-DM, HLA-DO and tapasin: functional similarities and differences. *Curr Opin Immunol*, 14(1):22–9, 2002.

[14] J. R. Gruen and S. M. Weissman. Human MHC class III and IV genes and disease associations *Front Biosci*, 6:D960–72, 2001.

[15] R. Dawkins, C. Leelayuwat, S. Gaudieri, G. Tay, J. Hui, S. Cattley, P. Martinez, and J. Kulski. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev*, 167:275–304, 1999.

[16] Z. Yang, L. Shen, A. W. Dangel, L. C. Wu, and C. Y. Yu. Four ubiquitously expressed genes, RD (D6S45)-SKI2W (SKIV2L)-DOM3Z-RP1 (D6S60E), are present between complement component genes factor B and C4 in the class III region of the HLA. *Genomics*, 53(3):338–47, 1998.

[17] Z. Yang, X. Qu, and C. Y. Yu. Features of the two gene pairs RD-SKI2W and DOM3Z-RP1 located between complement component genes factor B and C4 at the MHC class III region. *Front Biosci*, 6:D927–35, 2001.

[18] S. D. Wijesuriya, J. Bristow, and W. L. Miller. Localization and analysis of the principal promoter for human tenascin-X. *Genomics*, 80(4):443–52, 2002.

[19] B. Aguado and R. D. Campbell. The novel gene G17, located in the human major histocompatibility complex, encodes PBX2, a homeodomain-containing protein. *Genomics*, 25(3):650–9, 1995.

[20] R. L. Brake, U. R. Kees, and P. M. Watt. A complex containing PBX2 contributes to activation of the proto-oncogene HOX11. *Biochem Biophys Res Commun*, 294(1):23–34, 2002.

[21] K. Sugaya, S. Sasanuma, J. Nohata, T. Kimura, T. Fukagawa, Y. Nakamura, A. Ando, H. Inoko, T. Ikemura, and K. Mita. Gene organization of human NOTCH4 and (CTG)n polymorphism in this human counterpart gene of mouse proto-oncogene Int3. *Gene*, 189(2):235–44, 1997.

[22] L. Li, G. M. Huang, A. B. Banta, Y. Deng, T. Smith, P. Dong, C. Friedman, L. Chen, B. J. Trask, T. Spies, L. Rowen, and L. Hood. Cloning, characterization, and the complete 56.8-kilobase DNA sequence of the human NOTCH4 gene. *Genomics*, 51(1):45–58, 1998.

[23] A. M. Schmidt and D. M. Stern. Receptor for age (RAGE) is a gene within the major histocompatibility class III region: implications for host response mechanisms in homeostasis and chronic disease. *Front Biosci*, 6:D1151–60, 2001.

[24] K. Sugaya, T. Fukagawa, K. Matsumoto, K. Mita, E. Takahashi, A. Ando, H. Inoko, and T. Ikemura. Three genes in the human MHC class III region near the junction with the class II: gene for receptor of advanced glycosylation end products, PBX2 homeobox gene and a notch homolog, human counterpart of mouse mammary tumor gene int-3. *Genomics*, 23(2):408–19, 1994.

[25] M. I. New. Steroid 21-hydroxylase deficiency (congenital adrenal hyperplasia). *Am J Med*, 98(1A):2S–8S, 1995.

[26] A. M. Fourie, P. A. Peterson, and Y. Yang. Characterization and regulation of the major histocompatibility complex-encoded proteins Hsp70-Hom and Hsp70-1/2. *Cell Stress Chaperones*, 6(3):282–95, 2001.

[27] A. H. Hajeer and I. V. Hutchinson. Influence of TNFalpha gene polymorphisms on TNFalpha production and disease. *Hum Immunol*, 62(11):1191–9, 2001.

[28] N. J. Makhatadze. Tumor necrosis factor locus: genetic organisation and biological implications. *Hum Immunol*, 59(9):571–9, 1998.

[29] G. Ribas, M. Neville, J. L. Wixon, J. Cheng, and R. D. Campbell. Genes encoding three new members of the leukocyte antigen 6 superfamily and a novel member of Ig superfamily, together with genes encoding the regulatory nuclear chloride ion channel protein (hRNCC) and an N omega-N omega-dimethylarginine dimethylaminohydrolase homologue, are found in a 30-kb segment of the MHC class III region. *J Immunol*, 163(1):278–87, 1999.

[30] M. Mallya, R. D. Campbell, and B. Aguado. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. *Genomics*, 80(1):113–23, 2002.

[31] E. C. de Vet, B. Aguado, and R. D. Campbell. G6b, a novel immunoglobulin superfamily member encoded in the human major histocompatibility complex, interacts with SHP-1 and SHP-2. *J Biol Chem*, 276(45):42070–6, 2001.

[32] M. Kimoto, S. Miyatake, T. Sasagawa, H. Yamashita, M. Okita, T. Oka, T. Ogawa, and H. Tsuji. Purification, cDNA cloning and expression of human NG,NG-dimethylarginine dimethylaminohydrolase. *Eur J Biochem*, 258(2):863–8, 1998.

[33] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29(2):217–22, 2001.

[34] J. Klein, C. O'h Uigin, M. Kasahara, A. Vincek, D. Klein, and F. Figueroa. Frozen haplotypes in MHC evolution. In J. Klein and D. Klein, editors, *Molecular evolution of the major histocompatibility complex*. Springer Verlag, Berlin Heidelberg, 1991.

[35] S. Gaudieri, J. K. Kulski, R. L. Dawkins, and T. Gojobori. Extensive nucleotide variability within a 370 kb sequence from the central region of the major histocompatibility complex. *Gene*, 238(1):157–61, 1999.

[36] N. Keicho, J. Ohashi, G. Tamiya, K. Nakata, Y. Taguchi, A. Azuma, N. Ohishi, M. Emi, M. H. Park, H. Inoko, K. Tokunaga, and S. Kudoh. Fine localization of a major disease-susceptibility locus for diffuse panbronchiolitis. *Am J Hum Genet*, 66(2):501–7, 2000.

[37] S. Gaudieri, R. L. Dawkins, K. Habara, J. K. Kulski, and T. Gojobori. SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Res*, 10(10):1579–86, 2000.

[38] A. Levo, P. Westman, and J. Partanen. An approach to mapping haplotype-specific recombination sites in human MHC class III. *Immunogenetics*, 43(3):136–40, 1996.

[39] S. Mallal, D. Nolan, C. Witt, G. Masel, A. M. Martin, C. Moore, D. Sayer, A. Castley, C. Mamotte, D. Maxwell, I. James, and F. T. Christiansen. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet*, 359(9308):727–32, 2002.

[40] M. A. Degli-Esposti, A. L. Leaver, F. T. Christiansen, C. S. Witt, L. J. Abraham, and R. L. Dawkins. Ancestral haplotypes: conserved population MHC haplotypes. *Hum Immunol*, 34(4):242–52, 1992.

[41] M. A. Degli-Esposti, D. C. Townsend, L. K. Smith, M. Finlay, and R. L. Dawkins. Complotypes and ancestral haplotypes: association with autoimmune disease. In K. Tsuji, M. Aizawa, and T. Sasazuki, editors, *HLA 1991*, volume 1, pages 984–985. Oxford University Press, New York, 1992.

[42] S. Gaudieri, C. Leelayuwat, G. K. Tay, D. C. Townend, and R. L. Dawkins. The major histocompatability complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. *J Mol Evol*, 45(1):17–23, 1997.

[43] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–9, 2002.

[44] W. J. Zhang, M. A. Degli-Esposti, T. J. Cobain, P. U. Cameron, F. T. Christiansen, and R. L. Dawkins. Differences in gene copy number carried by different MHC ancestral haplotypes. Quantitation after physical separation of haplotypes by pulsed field gel electrophoresis. *J Exp Med*, 171(6):2101–14, 1990.

[45] M. A. Degli-Esposti, L. J. Abraham, C. S. Witt, U. P. Cameron, W. J. Zhang, F. T. Christiansen, and R. J. Dawkins. *Ancestral haplotypes are conserved population haplotypes with a unique MHC structure*. HLA. Oxford University Press, Oxford, 1991.

[46] G. Candore, D. Lio, G. C. Romano, and C. Caruso. Pathogenisis of autoimmune disease associated with 8.1 ancestral haplotype: effect of multiple gene interactions. *Autoimmunity Reviews*, 1:29–35, 2002.

[47] P. Price, C. Witt, R. Allcock, D. Sayer, M. Garlepp, C. C. Kok, M. French, S. Mallal, and F. Christiansen. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol Rev*, 167:257–74, 1999.

[48] S. Jenisch, J. T. Elder, E. Westphal, R P. Nair, E. Christophers, M. Kronke, and T. Henseler. Localization of a psoriasis vulagaris susceptibility allele on specific ancestral HLA-haplotypes. *J Dermatological Science*, 16(1):147, 1998.

[49] H. Kelly, V. J. McCann, P. H. Kay, and R. L. Dawkins. Susceptibility to IDDM is marked by MHC supratypes rather than individual alleles. *Immunogenetics*, 22(6):643–51, 1985.

[50] D. Raum, Z. Awdeh, E. J. Yunis, C. A. Alper, and K. H. Gabbay. Extended major histocompatibility complex haplotypes in type I diabetes mellitus. *J Clin Invest*, 74(2):449–54, 1984.

[51] R. L. Dawkins, W. J. Zhang, M. A. Degli-Esposti, L. Abraham, V. McCann, and F. T. Christiansen. Genetics of diabetes. Studies of MHC haplotypes by pulsed field gel electrophoresis. *Baillieres Clin Endocrinol Metab*, 5(2):285–97, 1991.

[52] J. J. Just. Genetic predisposition to HIV-1 infection and acquired immune deficiency virus syndrome: a review of the literature examining associations with HLA [corrected]. *Hum Immunol*, 44(3):156–69, 1995.

[53] F. T. Christiansen, G. C. Saueracker, A. L. Leaver, K. Tokunaga, P. U. Cameron, and R. L. Dawkins. Characterization of MHC ancestral haplotypes associated with insulin-dependent diabetes mellitus: evidence for involvement of non-HLA genes. *J Immunogenet*, 17(6):379–86, 1990.

[54] J. M. Aparicio, A. Wakisaka, A. Takada, N. Matsuura, and M. Aizawa. HLA-DQ system and insulin-dependent diabetes mellitus in Japanese: does it contribute to the development of IDDM as it does in Caucasians? *Immunogenetics*, 28(4):240–6, 1988.

[55] M. A. Degli-Esposti, C. Leelayuwat, L. N. Daly, C. Carcassi, L. Contu, L. F. Versluis, M. G. Tilanus, and R. L. Dawkins. Updated characterization of ancestral haplotypes using the Fourth Asia-Oceania Histocompatibility Workshop panel. *Hum Immunol*, 44(1):12–8, 1995.

[56] O. P. Martinez, N. Longman-Jacobsen, R. Davies, E. K. Chung, Y. Yang, S. Gaudieri, R. L. Dawkins, and C. Y. Yu. Genetics of human complement component C4 and evolution the central MHC. *Front Biosci*, 6:D904–13, 2001.

[57] M. Sakamoto, Y. Fujisawa, and K. Nishioka. Physiologic role of the complement system in host defense, disease, and malnutrition. *Nutrition*, 14(4):391–8, 1998.

[58] H. Buchner. Über die Natur der bakterientötenden Substanz im Blutserum. *Zentralbl Bakteriol*, 6:561572, 1889.

[59] B. Z. Schmidt and H. R. Colten. Complement: a critical test of its biological importance. *Immunol Rev*, 178:166–76, 2000.

[60] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik. The complement system and innate immunity. In *Immunobiology: the immune system in health and disease*, pages 43–64. Garland Publishing, New York, 5th edition, 2001.

[61] M. Tolnay and G. C. Tsokos. Complement receptor 2 in the regulation of the immune response. *Clin Immunol Immunopathol*, 88(2):123–32, 1998.

[62] M. C. Carroll and A. P. Prodeus. Linkages of innate and adaptive immunity. *Curr Opin Immunol*, 10(1):36–40, 1998.

[63] Z. Fishelson, G. Attali, and D. Mevorach. Complement and apoptosis. *Mol Immunol*, 38(2-3):207–19, 2001.

[64] V. D. Miletic and M. M. Frank. Complement-immunoglobulin interactions. *Curr Opin Immunol*, 7(1):41–7, 1995.

[65] D. C. Kilpatrick. Mannan-binding lectin: clinical significance and applications. *Biochim Biophys Acta*, 1572(2-3):401–13, 2002.

[66] B. B Vuagnat, J. Mach, and J. M. Le Doussal. Activation of the alternative pathway of human complement by autologous cells expressing transmembrane recombinant properdin. *Mol Immunol*, 37(8):467–78, 2000.

[67] K. Crawford and C. A. Alper. Genetics of the complement system. *Rev Immunogenet*, 2(3):323–38, 2000.

[68] C. Mold, H. Gewurz, and T. W. Du Clos. Regulation of complement activation by C-reactive protein. *Immunopharmacology*, 42(1-3):23–30, 1999.

[69] G. J. Arlaud, C. Gaboriaud, N. M. Thielens, V. Rossi, B. Bersch, J. F. Hernandez, and J. C. Fontecilla-Camps. Structural biology of C1: dissection of a complex molecular machinery. *Immunol Rev*, 180:136–45, 2001.

[70] Q. Pan, R. O. Ebanks, and D. E. Isenman. Two clusters of acidic amino acids near the NH2 terminus of complement component C4 alpha'-chain are important for C2 binding. *J Immunol*, 165(5):2518–27, 2000.

[71] J. M. Inal and J. A. Schifferli. Complement C2 receptor inhibitor trispanning and the beta-chain of C4 share a binding site for complement C2. *J Immunol*, 168(10):5213–21, 2002.

[72] M. Kirschfink. Controlling the complement system in inflammation. *Immunopharmacology*, 38(1-2):51–62, 1997.

[73] S. Thiel, T. Vorup-Jensen, C. M. Stover, W. Schwaeble, S. B. Laursen, K. Poulsen, A. C. Willis, P. Eggleton, S. Hansen, U. Holmskov, K. B. Reid, and J. C. Jensenius. A second serine protease associated with mannan-binding lectin that activates complement. *Nature*, 386(6624):506–10, 1997.

[74] C. Suankratay, X. H. Zhang, Y. Zhang, T. F. Lint, and H. Gewurz. Requirement for the alternative pathway as well as C4 and C2 in complement-dependent hemolysis via the lectin pathway. *J Immunol*, 160(6):3006–13, 1998.

[75] S. V. Petersen, S. Thiel, and J. C. Jensenius. The mannan-binding lectin pathway of complement activation: biology and disease association. *Mol Immunol*, 38(2-3):133–49, 2001.

[76] T. C. Farries and J. P. Atkinson. Evolution of the complement system. *Immunol Today*, 12(9):295–300, 1991.

[77] S. K. Law and A. W. Dodds. Catalysed hydrolysis–the complement quickstep. *Immunol Today*, 17(3):105, 1996.

[78] N. Rawal and M. K. Pangburn. Structure/function of C5 convertases of complement. *Int Immunopharmacol*, 1(3):415–22, 2001.

[79] D. Hourcade, M. K. Liszewski, M. Krych-Goldberg, and J. P. Atkinson. Functional domains, structural variations and pathogen interactions of MCP, DAF and CR1. *Immunopharmacology*, 49(1-2):103–16, 2000.

[80] Y. Sugita and Y. Masuho. CD59: its role in complement regulation and potential for therapeutic use. *Immunotechnology*, 1(3-4):157–68, 1995.

[81] L. A. Trouw, A. Roos, and M. R. Daha. Autoantibodies to complement components. *Mol Immunol*, 38(2-3):199–206, 2001.

[82] P. M. Schneider and R. Wurzner. Complement genetics: biological implications of polymorphisms and deficiencies. *Immunol Today*, 20(1):2–5, 1999.

[83] C. A. Blanchong, E. K. Chung, K. L. Rupert, Y. Yang, Z. Yang, B. Zhou, J. M. Moulds, and C. Y. Yu. Genetic, structural and functional diversities of human complement components C4A and C4B and their mouse homologues, Slp and C4. *Int Immunopharmacol*, 1(3):365–92, 2001.

[84] T. J. Mitchell, M. Naughton, P. Norsworthy, K. A. Davies, M. J. Walport, and B. J. Morley. IFN-gamma up-regulates expression of the complement components C3 and C4 by stabilization of mRNA. *J Immunol*, 156(11):4429–34, 1996.

[85] J. J. Timmerman, C. L. Verweij, D. J. van Gijlswijk-Janssen, F. J. van der Woude, L. A. van Es, and M. R. Daha. Cytokine-regulated production of the major histocompatibility complex class-III-encoded complement proteins factor B and C4 by human glomerular mesangial cells. *Hum Immunol*, 43(1):19–28, 1995.

[86] T. Seya, S. Nagasawa, and J. P. Atkinson. Location of the interchain disulfide bonds of the fourth component of human complement (C4): evidence based on the liberation of fragments secondary to thiol-disulfide interchange reactions. *J Immunol*, 136(11):4152–6, 1986.

[87] C. Y. Yu. The complete exon-intron structure of a human complement component C4A gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. *J Immunol*, 146(3):1057–66, 1991.

[88] I. von Zabern, E. L. Bloom, V. Chu, and I. Gigli. The fourth component of human complement treated with amines or chaotropes or frozen-thawed (C4b-like C4): interaction with C4 binding protein and cleavage by C3b/C4b inactivator. *J Immunol*, 128(3):1433–8, 1982.

[89] G. Uko, F. T. Christiansen, R. L. Dawkins, and V. J. McCann. Reference ranges for serum C4 concentrations in subjects with and without C4 null alleles. *J Clin Pathol*, 39(5):573–6, 1986.

[90] D. E. Isenman and J. R. Young. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. *J Immunol*, 132(6):3019–27, 1984.

[91] B. D. Reilly, R. P. Levine, and V. M. Skanes. Amino acid residues 1101-1105 of the isotypic region of human C4B is important to the covalent binding activity of complement component C4. *J Immunol*, 147(9):3018–23, 1991.

[92] A. W. Dodds, X. D. Ren, A. C. Willis, and S. K. Law. The reaction mechanism of the internal thioester in the human complement component C4. *Nature*, 379(6561):177–9, 1996.

[93] M. C. Carroll, D. M. Fathallah, L. Bergamaschini, E. M. Alicot, and D. E. Isenman. Substitution of a single amino acid (aspartic acid for histidine) converts the functional activity of human complement C4B to C4A. *Proc Natl Acad Sci U S A*, 87(17):6868–72, 1990.

[94] C. A. Blanchong, B. Zhou, K. L. Rupert, E. K. Chung, K. N. Jones, J. F. Sotos, W. B. Zipf, R. M. Rennebohm, and C. Yung Yu. Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J Exp Med*, 191(12):2183–96, 2000.

[95] M. H. Roos, C. M. Giles, P. Demant, E. Mollenhauer, and C. Rittner. Rodgers (Rg) and Chido (Ch) determinants on human C4: characterization of two C4 B5 subtypes, one of which contains Rg and Ch determinants. *J Immunol*, 133(5):2634–40, 1984.

[96] O. G. Segurado, C. M. Giles, P. Iglesias-Casarrubios, A. Corell, J. Martinez-Laso, J. L. Vicario, and A. Arnaiz-Villena. C4 Chido 3 and 6 distinguish two diabetogenic haplotypes: HLA-B49, SC01,DR4,DQw8 and B8,SC01,DR3,DQw2. *Immunobiology*, 183(1-2):12–22, 1991.

[97] V. F. Chu, W. L. Marsh, and I. Gigli. Chido and Rogers antigenic determinant on the fourth component of human complement. *J Immunol*, 128(1):181–5, 1982.

[98] J. M. Moulds, S. L. Roberts, and T. D. Wells. DNA sequence analysis of the C4 antigen WH: evidence for two mechanisms of expression. *Immunogenetics*, 44(2):104–7, 1996.

[99] E. Paz-Artal, A. Corell, M. Alvarez, P. Varela, L. Allende, A. Madrono, M. Rosal, and A. Arnaiz-Villena. C4 gene polymorphism in primates: evolution, generation, and Chido and Rodgers antigenicity. *Immunogenetics*, 40(6):381–96, 1994.

[100] C. M. Giles, K. Tokunaga, W. J. Zhang, H. Tanaka, N. Endoh, and T. Juji. The antigenic determinants, Rg/Ch/WH, expressed by Japanese C4 allotypes. *J Immunogenet*, 15(5-6):267–75, 1988.

[101] G. M. Barba, L. Braun-Heimer, C. Rittner, and P. M. Schneider. A new PCR-based typing of the Rodgers and Chido antigenic determinants of the fourth component of human complement. *Eur J Immunogenet*, 21(5):325–39, 1994.

[102] N. Martinez-Quiles, E. Paz-Artal, M. A. Moreno-Pelayo, J. Longas, S. Ferre-Lopez, M. Rosal, and A. Arnaiz-Villena. C4d DNA sequences of two infrequent human allotypes (C4A13 and C4B12) and the presence of signal sequences enhancing recombination. *J Immunol*, 161(7):3438–43, 1998.

[103] D. Ulgiati and L. J. Abraham. Extensive conservation of upstream C4 promoter sequences: a comparison between C4A and C4B. *Tissue Antigens*, 48(5):600–3, 1996.

[104] D. Ulgiati and L. J. Abraham. Comparative analysis of the disease-associated complement C4 gene from the HLA-A1, B8, DR3 haplotype. *Exp Clin Immunogenet*, 13(1):43–54, 1996.

[105] A. K. Vaishnaw, T. J. Mitchell, S. J. Rose, M. J. Walport, and B. J. Morley. Regulation of transcription of the TATA-less human complement component C4 gene. *J Immunol*, 160(9):4353–60, 1998.

[106] D. Ulgiati, L. S. Subrata, and L. J. Abraham. The role of Sp family members, basic Kruppel-like factor, and E box factors in the basal and IFN-gamma regulated expression of the human complement C4 promoter. *J Immunol*, 164(1):300–7, 2000.

[107] S. F. Grant, H. Kristjansdottir, K. Steinsson, T. Blondal, A. Yuryev, K. Stefansson, and J. R. Gulcher. Long PCR detection of the C4A null allele in B8-C4AQ0-C4B1-DR3. *J Immunol Methods*, 244(1-2):41–7, 2000.

[108] A. W. Dangel, A. R. Mendoza, B. J. Baker, C. M. Daniel, M. C. Carroll, L. C. Wu, and C. Y. Yu. The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. *Immunogenetics*, 40(6):425–36, 1994.

[109] P. M. Schneider, K. Witzel-Schlomp, C. Rittner, and L. Zhang. The endogenous retroviral insertion in the human complement C4 gene modulates the expression of homologous genes by antisense inhibition. *Immunogenetics*, 53(1):1–9, 2001.

[110] E. D. Sverdlov. Perpetually mobile footprints of ancient infections in human genome. *FEBS Lett*, 428(1-2):1–6, 1998.

[111] G. Andersson, A. C. Svensson, N. Setterblad, and L. Rask. Retroelements in the human MHC class II region. *Trends Genet*, 14(3):109–14, 1998.

[112] R. Lower, J. Lower, and R. Kurth. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A*, 93(11):5177–84, 1996.

[113] M. A. Pani, J. P. Wood, K. Bieda, R. R. Toenjes, K. H. Usadel, and K. Badenhoop. The variable endogenous retroviral insertion in the human complement C4 gene: a transmission study in type I diabetes mellitus. *Hum Immunol*, 63(6):481–4, 2002.

[114] Z. Yang, A. R. Mendoza, T. R. Welch, W. B. Zipf, and C. Y. Yu. Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J Biol Chem*, 274(17):12147–56, 1999.

[115] T. Jaatinen, E. K. Chung, O. Ruuskanen, and M. L. Lokki. An unequal crossover event in RCCX modules of the human MHC resulting in the formation of a TNXB/TNXA hybrid and deletion of the CYP21A. *Hum Immunol*, 63(8):683–9, 2002.

[116] E. K. Chung, Y. Yang, R. M. Rennebohm, M. L. Lokki, G. C. Higgins, K. N. Jones, B. Zhou, C. A. Blanchong, and C. Y. Yu. Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. *Am J Hum Genet*, 71(4):823–37, 2002.

[117] E. K. Chung, Y. Yang, K. L. Rupert, K. N. Jones, R. M. Rennebohm, C. A. Blanchong, and C. Y. Yu. Determining the one, two, three, or four long and short loci of human complement C4 in a major histocompatibility complex haplotype encoding C4A or C4B proteins. *Am J Hum Genet*, 71(4):810–22, 2002.

[118] L. Braun, P. M. Schneider, C. M. Giles, J. Bertrams, and C. Rittner. Null alleles of human complement C4. Evidence for pseudogenes at the C4A locus and for gene conversion at the C4B locus. *J Exp Med*, 171(1):129–40, 1990.

[119] K. L. Rupert, J. M. Moulds, Y. Yang, F. C Arnett, R. W. Warren, J. D. Reveille, B. L. Myones, C. A. Blanchong, and C. Y. Yu. The molecular basis of complete complement C4A and C4B deficiencies in a systemic lupus erythematosus patient with homozygous C4A and C4B mutant genes. *J Immunol*, 169(3):1570–8, 2002.

[120] M. L. Lokki, A. Circolo, P. Ahokas, K. L. Rupert, C. Y. Yu, and H. R. Colten. Deficiency of human complement protein C4 due to identical frameshift mutations in the C4A and C4B genes. *J Immunol*, 162(6):3687–93, 1999.

[121] G. N. Fredrikson, B. Gullstrand, P. M. Schneider, K. Witzel-Schlomp, A. G. Sjoholm, C. A. Alper, Z. Awdeh, and L. Truedsson. Characterization of non-expressed C4 genes in a case of complete C4 deficiency: identification of a novel point mutation leading to a premature stop codon. *Hum Immunol*, 59(11):713–9, 1998.

[122] M. Brai, P. Accardo, and D. Bellavia. Polymorphism of the complement components in human pathology. *Ann Ital Med Int*, 9(3):167–72, 1994.

[123] M. J. Anderson, C. M. Milner, R. G. Cotton, and R. D. Campbell. The coding sequence of the hemolytically inactive C4A6 allotype of human complement component C4 reveals that a single arginine to tryptophan substitution at beta-chain residue 458 is the likely cause of the defect. *J Immunol*, 148(9):2795–802, 1992.

[124] T. Jaatinen. *Genetic studies of the human complement C4 region in MHC class III.* Academic dessertation, University of Helsinki, 2002.

[125] P. F. Howard, M. C. Hochberg, W. B. Bias, Jr. Arnett, F. C., and R. H. McLean. Relationship between C4 null genes, HLA-D region antigens, and genetic susceptibility to systemic lupus erythematosus in Caucasian and black Americans. *Am J Med*, 81(2):187–93, 1986.

[126] M. L. Olsen, R. Goldstein, F. C. Arnett, M. Duvic, M. Pollack, and J. D. Reveille. C4A gene deletion and HLA associations in black Americans with systemic lupus erythematosus. *Immunogenetics*, 30(1):27–33, 1989.

[127] F. T. Christiansen, W. J. Zhang, M. Griffiths, S. A. Mallal, and R. L. Dawkins. Major histocompatibility complex (MHC) complement deficiency, ancestral haplotypes and systemic lupus erythematosus (SLE): C4 deficiency explains some but not all of the influence of the MHC. *J Rheumatol*, 18(9):1350–8, 1991.

[128] F. T. Christiansen, R. L. Dawkins, G. Uko, J. McCluskey, P. H. Kay, and P. J. Zilko. Complement allotyping in SLE: association with C4A null. *Aust N Z J Med*, 13(5):483–8, 1983.

[129] X. Z. Zhao, W. J. Zhang, Y. W. Tian, F. Wu, L. Zhang, X. D. Jiang, Z. Z. Sun, C. F. Hu, W. Z. Wan, and L. Gan. Allotypic differences and frequencies of C4 null alleles (C4Q0) detected in patients with systemic lupus erythematosus (SLE). *Chinese Science Bulletin*, 34(3), 1989.

[130] T. Jaatinen, O. Ruuskanen, L. Truedsson, and M. L. Lokki. Homozygous deletion of the CYP21A-TNXA-RP2-C4B gene region conferring C4B deficiency associated with recurrent respiratory infections. *Hum Immunol*, 60(8):707–14, 1999.

[131] O. Finco, S. Li, M. Cuccia, F. S. Rosen, and M. C. Carroll. Structural differences between the two human complement C4 isotypes affect the humoral immune response. *J Exp Med*, 175(2):537–43, 1992.

[132] H. S. Howe, A. K. So, J. Farrant, and A. D. Webster. Common variable immunodeficiency is associated with polymorphic markers in the human major histocompatibility complex. *Clin Exp Immunol*, 83(3):387–90, 1991.

[133] T. R. Welch, L. S. Beischel, and E. M. Choi. Molecular genetics of C4B deficiency in IgA nephropathy. *Hum Immunol*, 26(4):353–63, 1989.

[134] D. K. Jin, T. Kohsaka, J. W. Koo, I. S. Ha, H. I. Cheong, and Y. Choi. Complement 4 locus II gene deletion and DQA1*0301 gene: genetic risk factors for IgA nephropathy and Henoch-Schonlein nephritis. *Nephron*, 73(3):390–5, 1996.

[135] L. J. Scully, C. Toze, D. P. Sengar, and R. Goldstein. Early-onset autoimmune hepatitis is associated with a C4A gene deletion. *Gastroenterology*, 104(5):1478–84, 1993.

[136] P. U. Cameron, T. J. Cobain, W. J. Zhang, P. H. Kay, and R. L. Dawkins. Influence of C4 null genes on infection with human immunodeficiency virus. *Br Med J (Clin Res Ed)*, 296(6637):1627–8, 1988.

[137] P. U. Cameron, S. A. Mallal, M. A. French, and R. L. Dawkins. Major histocompatibility complex genes influence the outcome of HIV infection. Ancestral haplotypes with C4 null alleles explain diverse HLA associations. *Hum Immunol*, 29(4):282–95, 1990.

[138] P. M. Schneider, C. Wendler, T. Riepert, L. Braun, U. Schacker, M. Horn, H. Althoff, R. Mattern, and C. Rittner. Possible association of sudden infant death with partial complement C4 deficiency revealed by post-mortem DNA typing of HLA class II and III genes. *Eur J Pediatr*, 149(3):170–4, 1989.

[139] S. H. Opdal, A. Vege, A. K. Stave, and T. O. Rognum. The complement component C4 in sudden infant death. *Eur J Pediatr*, 158(3):210–2, 1999.

[140] D. Franciotta, E. Dondi, R. Bergamaschi, G. Piccolo, G. V. d'Eril, V. Cosi, and M. Cuccia. HLA complement gene polymorphisms in multiple sclerosis. A study on 80 Italian patients. *J Neurol*, 242(2):64–8, 1995.

[141] G. T. Venneker, W. Westerhof, I. J. de Vries, N. M. Drayer, B. G. Wolthers, L. P. de Waal, J. D. Bos, and S. S. Asghar. Molecular heterogeneity of the fourth component of complement (C4) and its genes in vitiligo. *J Invest Dermatol*, 99(6):853–8, 1992.

[142] L. J. Abraham, M.A. Degli-Esposti, W. Zhang, F. T. Christiansen, and R. L. Dawkins. Mapping autoimmune disease susceptibility genes in the major histocompatibility complex. *Proceedings of the Australian Physiological and Pharmacological Society*, 22(2):201–203, 1991.

[143] H. Ikegami, S. Makino, E. Yamato, Y. Kawaguchi, H. Ueda, T. Sakamoto, K. Takekawa, and T. Ogihara. Identification of a new susceptibility locus for insulin-dependent diabetes mellitus by ancestral haplotype congenic mapping. *J Clin Invest*, 96(4):1936–42, 1995.

[144] M. P. Martin, A. Harding, R. Chadwick, M. Kronick, M. Cullen, L. Lin, E. Mignot, and M. Carrington. Characterization of 12 microsatellite loci of the human MHC in a panel of reference cell lines. *Immunogenetics*, 47(2):131–8, 1998.

[145] A. Foissac and A. Cambon-Thomsen. Microsatellites in the HLA region: 1998 update. *Tissue Antigens*, 52(4):318–52, 1998.

[146] X. Wu, W. J. Zhang, C. S. Witt, L. J. Abraham, F. T. Christiansen, and R. L. Dawkins. Haplospecific polymorphism between HLA B and tumor necrosis factor. *Hum Immunol*, 33(2):89–97, 1992.

[147] W. J. Zhang, X. Wu, C. Witt, C. Leelayuwat, L. J. Abraham, G. Grimsley, M. A. Degli-Esposti, F. T. Christiansen, and R. L. Dawkins. New genomic markers between the HLA-B and TNF loci define MHC ancestral haplotypes. In K. Tsuji, M. Aizawa, and T. Sasazuki, editors, *HLA 1991*, volume 2, pages 176–179. Oxford University Press, New York, 1992.

[148] C. Leelayuwat, L. J. Abraham, H. A. Tabarias, A. J. Mann, G. Grimsley, M. A. Degli-Esposti, W. J. Zhang, F. T. Christiansen, and R. L. Dawkins. Haplospecific polymorphism in a duplicated region centromeric of HLA-B. In K. Tsuji, M. Aizawa, and T. Sasazuki, editors, *HLA 1991*, volume 2, pages 172–175. Oxford University Press, New York, 1992.

[149] A. M. Uglialoro, D. Turbay, P. A. Pesavento, J. C. Delgado, F. E. McKenzie, J. G. Gribben, D. Hartl, E. J. Yunis, and A. E. Goldfeld. Identification of three new single nucleotide polymorphisms in the human tumor necrosis factor-alpha gene promoter. *Tissue Antigens*, 52(4):359–67, 1998.

[150] M. A. Degli-Esposti, C. Leelayuwat, and R. L. Dawkins. Ancestral haplotypes carry haplotypic and haplospecific polymorphisms of BAT1: possible relevance to autoimmune disease. *Eur J Immunogenet*, 19(3):121–7, 1992.

[151] D. P. Singal, J. Li, and Y. Zhu. HLA class III region and susceptibility to rheumatoid arthritis. *Clin Exp Rheumatol*, 18(4):485–91, 2000.

[152] D. Franciotta, M. Cuccia, E. Dondi, G. Piccolo, and V. Cosi. Polymorphic markers in MHC class II/III region: a study on Italian patients with myasthenia gravis. *J Neurol Sci*, 190(1-2):11–6, 2001.

[153] V. B. Matthews, C. S. Witt, M. A. French, H. K. Machulla, E. G. De la Concha, K. Y. Cheong, P. Vigil, P. N. Hollingsworth, K. J. Warr, F. T. Christiansen, and P. Price. Central MHC genes affect IgA levels in the human: reciprocal effects in IgA deficiency and IgA nephropathy. *Hum Immunol*, 63(5):424–33, 2002.

[154] M. Thomsen, J. Molvig, A. Zerbib, C. de Preval, M. Abbal, J. M. Dugoujon, E. Ohayon, A. Svejgaard, A. Cambon-Thomsen, and J. Nerup. The susceptibility to insulin-dependent diabetes mellitus is associated with C4 allotypes independently of the association with HLA-DQ alleles in HLA-DR3,4 heterozygotes. *Immunogenetics*, 28(5):320–7, 1988.

[155] I. Durinovic-Bello. Autoimmune diabetes: the role of T cells, MHC molecules and autoantigens. *Autoimmunity*, 27(3):159–77, 1998.

[156] M. A. Atkinson and N. K. Maclaren. Mechanisms of disease: The pathogenesis of insulin-dependent diabetes mellitus. *New England Journal of Medicine*, 331(21):1428–1436, 1994.

[157] R. Tisch and H. McDevitt. Insulin-dependent diabetes mellitus. *Cell*, 85(3):291–7, 1996.

[158] F. Luhder, J. Katz, C. Benoist, and D. Mathis. Major histocompatibility complex class II molecules can protect from diabetes by positively selecting T cells with additional specificities. *J Exp Med*, 187(3):379–87, 1998.

[159] W. M. Ridgway and C. G. Fathman. The association of MHC with autoimmune diseases: understanding the pathogenesis of autoimmune diabetes. *Clin Immunol Immunopathol*, 86(1):3–10, 1998.

[160] P. A. Gottlieb and G. S. Eisenbarth. Human autoimmune diabetes. In A. N. Theofilopoulos and C. A. Bona, editors, *The molecular pathology of autoimmune diseases*, pages 588–610. Taylor Francis, 2nd edition, 2002.

[161] M. J. Redondo, M. Rewers, L. Yu, S. Garg, C. C. Pilcher, R. B. Elliott, and G. S. Eisenbarth. Genetic determination of islet cell autoimmunity in monozygotic twin, dizygotic twin, and non-twin siblings of patients with type 1 diabetes: prospective twin study. *Bmj*, 318(7185):698–702, 1999.

[162] M. A. Atkinson and G. S. Eisenbarth. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. *Lancet*, 358(9277):221–9, 2001.

[163] P. Concannon, K. J. Gogolin-Ewens, D. A. Hinds, B. Wapelhorst, V. A. Morrison, B. Stirling, M. Mitra, J. Farmer, S. R. Williams, N. J. Cox, G. I. Bell, N. Risch, and R. S. Spielman. A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat Genet*, 19(3):292–6, 1998.

[164] C. A. Mein, L. Esposito, M. G. Dunn, G. C. Johnson, A. E. Timms, J. V. Goy, A. N. Smith, L. Sebag-Montefiore, M. E. Merriman, A. J. Wilson, L. E. Pritchard, F. Cucca, A. H. Barnett, S. C. Bain, and J. A. Todd. A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat Genet*, 19(3):297–300, 1998.

[165] J. A. Todd. Genetic analysis of type 1 diabetes using whole genome approaches. *Proc Natl Acad Sci U S A*, 92(19):8560–5, 1995.

[166] D. Owerbach and K. H. Gabbay. The search for IDDM susceptibility genes: the next generation. *Diabetes*, 45(5):544–51, 1996.

[167] D. E. Undlien and E. Thorsby. HLA associations in type 1 diabetes: merging genetics and immunology. *Trends Immunol*, 22(9):467–9, 2001.

[168] K. Y. Cheong, R. J. Allcock, P. Eerligh, C. S. Witt, F. T. Christiansen, V. McCann, and P. Price. Localization of central MHC genes influencing type I diabetes. *Hum Immunol*, 62(12):1363–70, 2001.

[169] M. A. Degli-Esposti, L. J. Abraham, V. McCann, T. Spies, F. T. Christiansen, and R. L. Dawkins. Ancestral haplotypes reveal the role of the central MHC in the immunogenetics of IDDM. *Immunogenetics*, 36(6):345–56, 1992.

[170] H. X. Yu, A. C. Thai, and S. H. Chan. HLA microsatellite associations with insulin-dependent diabetes mellitus in Singaporean Chinese. *Hum Immunol*, 60(9):894–900, 1999.

[171] S. G. E. Marsh, R. Packer, J. M. Heyes, B. Bolton, R. Fauchet, D. Charron, and J. G. Bodmer. *The 12th International Histocompatibility Workshop cell lines panel*. Genetic diversity of HLA. Paris, 1997.

[172] A. C. Jones, J. Austin, N. Hansen, B. Hoogendoorn, P. J. Oefner, J. P. Cheadle, and M. C. O'Donovan. Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. *Clin Chem*, 45(8 Pt 1):1133–40, 1999.

[173] K. H. Hecker, P. D. Taylor, and D. T. Gjerde. Mutation detection by denaturing DNA chromatography using fluorescently labeled polymerase chain reaction products. *Anal Biochem*, 272(2):156–64, 1999.

[174] T. J. Schmitt, M. L. Robinson, and J. Doyle. Single nucleotide polymorphism (SNP), insertion and deletion detection on the WAVE Nucleid Acid Fragment Analysis System. Technical Report Application Note 112, Transgenomic Inc. Omaha.

[175] K. L. Rupert, R. M. Rennebohm, and C. Y. Yu. An unequal crossover between the RCCX modules of the human MHC leading to the presence of a CYP21B gene and a tenascin TNXB/TNXA-RP2 recombinant between C4A and C4B genes in a patient with juvenile rheumatoid arthritis. *Exp Clin Immunogenet*, 16(2):81–97, 1999.

[176] D. Ulgiati, D. C. Townend, F. T. Christiansen, R. L. Dawkins, and L. J. Abraham. Complete sequence of the complement C4 gene from the HLA-A1, B8, C4AQ0, C4B1, DR3 haplotype. *Immunogenetics*, 43(4):250–2, 1996.

[177] A. Kuklin, A. P. Davıs, K. H. Hecker, D. T. Gjerde, and P. D. Taylor. A novel technique for rapıd automated genotyping of DNA polymorphısms in the mouse. *Mol Cell Probes*, 13(3):239–42, 1999.

[178] A. K. Vaishnaw, R. Hargreaves, R. D. Campbell, B. J. Morley, and M. J. Walport. DNase I hypersensitivity mapping and promoter polymorphism analysis of human C4. *Immunogenetics*, 41(6):354–8, 1995.

[179] R. D. Campbell, I. Dunham, E. Kendall, and C. A. Sargent. Polymorphism of the human complement component C4. *Exp Clin Immunogenet*, 7(1):69–84, 1990.

[180] W. J. Zhang, P. H. Kay, T. J. Cobain, and R. L. Dawkins. C4 allotyping on plasma or serum: application to routıne laboratories. *Hum Immunol*, 21(3):165–71, 1988.

[181] Z. L. Awdeh and C. A. Alper. Inherited structural polymorphism of the fourth component of human complement. *Proc Natl Acad Sci U S A*, 77(6):3576–80, 1980.

[182] E. Sim and S. J. Cross. Phenotyping of human complement component C4, a class-III HLA antigen. *Biochem J*, 239(3):763–7, 1986.

[183] G. Hortın, H. Sims, and A. W. Strauss. Identification of the site of sulfation of the fourth component of human complement. *J Biol Chem*, 261(4):1786–93, 1986.

[184] J. W. Kehoe and C. R. Bertozzi. Tyrosine sulfation: a modulator of extracellular proteın-protein interactions. *Chem Biol*, 7(3):R57–61, 2000.

[185] R. Horton, D. Niblett, S. Milne, S. Palmer, B. Tubby, J. Trowsdale, and S. Beck. Large-scale sequence comparisons reveal unusually hıgh levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J Mol Biol*, 282(1):71–97, 1998.

[186] W. H. Li and L. A. Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–23, 1991.

[187] C. F. Aquadro, V. Bauer DuMont, and F. A. Reed. Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev*, 11(6):627–34, 2001.

[188] T. Jaatinen, M. Eholuoto, T. Laitinen, and M. L. Lokki. Characterization of a de novo conversion in human complement C4 gene producing a C4B5-like protein. *J Immunol*, 168(11):5652–8, 2002.

[189] C. Y. Yu. Molecular genetics of the human MHC complement gene cluster. *Exp Clin Immunogenet*, 15(4):213–30, 1998.

[190] H. Kawaguchı, Z. Zaleska-Rutczynska, F. Figueroa, C. O'HUigin, and J. Klein. C4 genes of the chimpanzee, gorilla, and orang-utan: evıdence for extensive homogenization. *Immunogenetics*, 35(1):16–23, 1992.

[191] G. Mauff, C. A. Alper, R. Dawkins, G. Doxiadis, C. M. Giles, G. Hauptmann, C. Rittner, and P. M. Schneıder. C4 nomenclature statement (1990). *Complement Inflamm*, 7(4-6):261–8, 1990.

[192] P. M. Schneider, B. Stradmann-Bellinghausen, and C. Rittner. Genetic polymorphism of the fourth component of human complement: population study and proposal for a revised nomenclature based on genomic PCR typing of Rodgers and Chido determınants. *Eur J Immunogenet*, 23(5):335–44, 1996.

[193] S. Hetherıngton, A. R. Hughes, M. Mosteller, D. Shortino, K. L. Baker, W. Spreen, E. Laı, K. Davies, A. Handley, D. J. Dow, M. E. Fling, M. Stocum, C. Bowman, L. M. Thurmond, and A. D. Roses. Genetic variatıons in HLA-B region and hypersensitivity reactıons to abacavir. *Lancet*, 359(9312):1121–2, 2002.

[194] E. De Clercq. New developments in anti-HIV chemotherapy. *Biochim Biophys Acta*, 1587(2-3):258–75, 2002.

[195] S. Hetherington, S. McGuirk, G. Powell, A. Cutrell, O. Naderer, B. Spreen, S. Lafon, G. Pearce, and H. Steel. Hypersensitivity reactions during therapy with the nucleoside reverse transcriptase inhibitor abacavir. *Clin Ther*, 23(10):1603–14, 2001.

[196] P. G. Clay. The abacavir hypersensitivity reaction: a review. *Clin Ther*, 24(10):1502–14, 2002.

[197] R. H. McLean, G. Niblack, B. Julian, T. Wang, R. Wyatt, 3rd Phillips, J. A., T. S. Collins, J. Winkelstein, and D. Valle. Hemolytically inactive C4B complement allotype caused by a proline to leucine mutation in the C5-binding site. *J Biol Chem*, 269(44):27727–31, 1994.

[198] N. B. Blatt and G. D. Glick. Anti-DNA autoantibodies and systemic lupus erythematosus. *Pharmacol Ther*, 83(2):125–39, 1999.