# *Review of Literature*

# Review of Literature

In 1986 Human genome Project was launched by Charles DeLisi . Goal of this initiative was to understand the human genome. He said that "Knowledge of the human genome is as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine."

Human genome project was completed on 14 April 2003, In October 2004, researchers from the International Human Genome Sequencing Consortium (IHGSH) of the HGP announced a new estimate of 20,000 to 25,000 genes in the human genome **(Fiers 2004).** Previously 30,000 to 40,000 genes had been predicted, while at the start of project the estimated Number of gene was approx 2,000,000. The goals of the original HGP were not only to determine all 3 billion base pairs in the human genome, but also to identify all the genes in this vast amount of data.

It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology. Like genetic tests that can show predisposition to a variety of illnesses, including breast cancer, disorders of hemostasis, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management **(Farrer *et al.*, 1997; Barnhart & Benjamin 1989).** The Human Genome Diversity Project, also aims at mapping the DNA that varies between human ethnic groups. In the future, HGDP could possibly expose new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the susceptibility of ethnic groups to certain diseases. It could also show how human populations have adapted to these Susceptibilities **(Barnhart & Benjamin 1989).** The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of the theory of evolution. Similarities and

differences between humans and our closest relatives (the primates, and the other mammals) are expected to be illuminated by the genetic variation studies DNA from all organisms is made up of the same chemical and physical components. The DNA sequence is the particular side-by-side arrangement of bases along the DNA strand, like ATTCCGGA. The genome is an organism's complete set of DNA. Genomes vary widely in size: the smallest known genome for a free-living organism (a bacterium) contains about 600,000 DNA base pairs, while human and mouse genomes have some 3 billion. Except for mature red blood cells, all human cells contain a complete genome (**Figure 2.1**).

DNA in the human genome is arranged into 24 distinct chromosomes--physically separate molecules that range in length from about 50 million to 250 million base pairs.

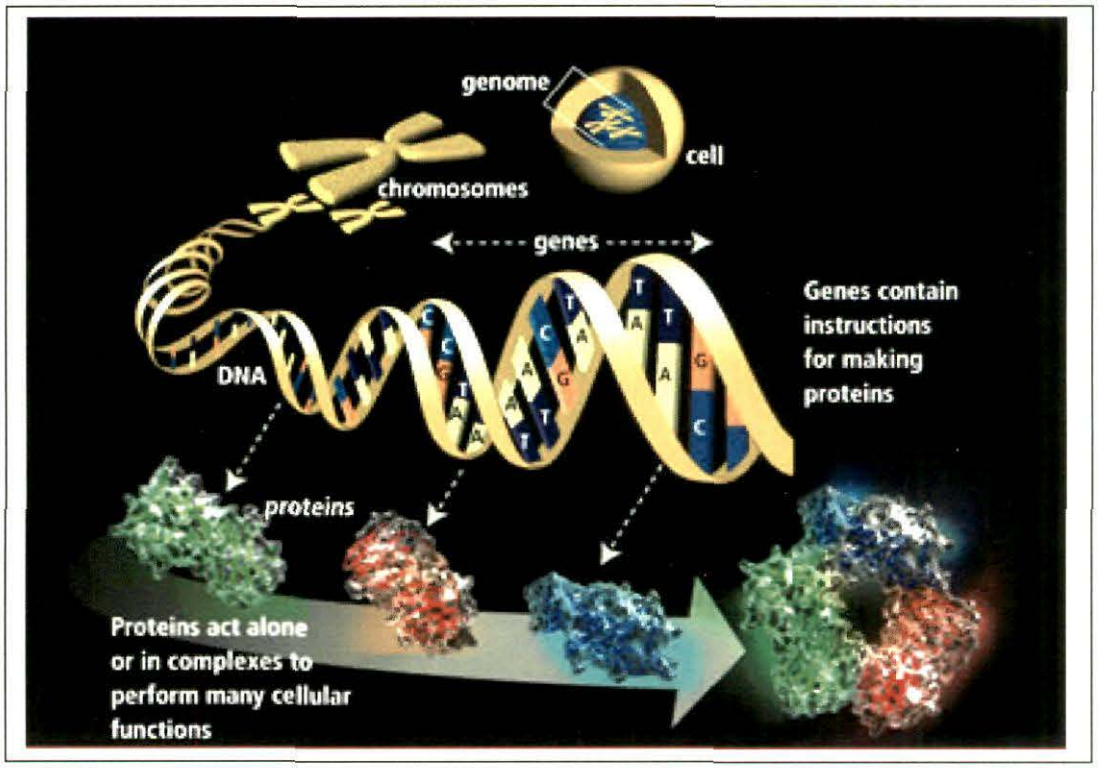



**Figure 2.1 Genome organization**

Each chromosome contains many genes, the basic physical and functional units of heredity. These genes are encoded in segments of DNA and the information is contained in 23 pairs of chromosomes. DNA transmits its information through four

different nucleotide bases, which are the "letters" of the genetic code A, C, G, and T. The human genome contains approx 3164.7 million chemical nucleotide bases. The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases. Almost all (99.9%) nucleotide bases are exactly the same in all people. These bases, when read as strings of three letter word, form the blue prints of molecules responsible for all the functions of the body. But the coding region of these genes takes up only 2-3% of the human genome .The remaining consists of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made. Although genes get a lot of attention, it's the proteins that perform most life functions and even make up the majority of cellular structures. Proteins are large, complex molecules made up of smaller subunits called amino acids. Unlike the relatively unchanging genome, the dynamic proteome changes from minute to minute in response to tens of thousands of intra- and extra cellular environmental signals. A protein's chemistry and behaviour are specified by the gene sequence and by the number and identities of other proteins made in the same cell at the same time and with which it associated.

Human genome varies from individual to individual and therefore no two individuals look alike (**Fischer *et al.*, 2004**). Historically, individual variation was studied on the basis of conventional markers. However, with the advancement of technology various genetic markers were discovered and the gene frequency data for studying the evolution of human races was analyzed using these markers. Initially, the classical serological and biochemical markers have played important roles in various types of human population genetic studies. However, it is important to record population variation because it is helpful to know the various mechanisms involved in causing variation and it further enhances our knowledge about the molecular basis of disease susceptibility. Basic information about the types, frequencies and distribution of common variants are essential not only for the understanding of pathological entities, but also to unravel our evolutionary past and provide guidance about our biological future (**Jorde & Wooding, 2004**).

To study the population variation various genetic markers are being used. Some of the important markers involved for studying ethnic variability are conventional markers like blood groups, protein polymorphisms, restriction fragment length polymorphisms, short tandem repeat polymorphisms, variable tandem repeat polymorphisms, single nucleotide polymorphisms and human leukocyte (HLA) polymorphisms. In **Table 2.1** the characteristic features of various markers is shown

**Table 2.1 Genetic markers used to study population variations**

| Type of marker | Year | No. of loci | Features |
|---|---|---|---|
| Blood groups | 1910-1960 | ~20 | May need fresh blood, rare antisera. Genotype cannot always be inferred from phenotype because of dominance. No easy physical localization. |
| Electrophoretic mobility variants of serum proteins | 1960-1975 | ~30 | May need fresh serum, specialized assays, no easy physical localization often limited polymorphisms |
| Human Leukocyte Antigens (HLA) | 1970 | 1 (multi locus haplotype) | One linked set highly informative. Can only test for linkage to 6p21.3 |
| DNA RFLPs | 1975 | $>10^5$ (potentially) | Two allele markers, maximum heterozygosity 0.5, initially required Southern blotting, now PCR. Easy physical localization |
| DNA VNTRs (minisatellites) | 1985- | >104 (potentially) | Many alleles, highly informative can be typed by southern blotting easy physical localization. Tend to cluster near ends of chromosomes. |
| DNA VNTRs (microsatellites) (di-,tri-, and tetranucleotide repeats) | 1989- | 105 (potentially) | Many alleles, highly informative Can be typed by automated multiplex PCR, easy physical localization. Distributed throughout genome |
| DNA SNPs | 1998- | 106 (potentially) | Less informative than microsatellites. Can be typed on a very large scale by automated equipment, without gel electrophoresis, etc. |

In present study we have taken into consideration the HLA class II antigens to decipher the social structure of 6 North Indian populations and 3 North Eastern caste and tribal populations. In the following section we have reviewed the human leukocyte antigens in detail.

**Human Leukocyte Antigens (HLA)**

The major histocompatibility complex (MHC) is a dense complex of genes with immunological and non-immunological functions and is present in all vertebrates. In humans it is known as human leukocyte antigens (HLA) **(Marsh, 2000; Trowsdale, 1995; Gruen & Weissman, 1997)**. Peter Gorer discovered it during transplantation studies in mice (the H-2 complex) in 1937. Jean Dausset described the first human MHC antigen *Mac* (HLA-A2) followed by the discovery of 4a and 4b. MHC is best known with its role in histocompatibility **(Snell, 1981)** and in immune regulation **(Jorde *et al.*, 1999; Benacerraf & Devitt, 1972)**. The main function of the MHC molecules is peptide binding and presentation of these peptides to T lymphocytes. Among the non-immune functions, the noteworthy ones are interactions with receptors on the cell surface **(Jorde *et al.*, 1999; Edidin, 1988; Svejgaard & Ryder, 1976)**, in particular with transferring receptor (TfR), epidermal growth factor **(Schreiber *et al.*, 1984)** and various other hormone receptors and signal transduction **(Abbas *et al* 1991; Schafer *et al.* 1995)**.

In humans, the genes for the HLA antigens are located on the short arm of chromosome 6 in the band 6p 21.3 **(Trachtenberg 1996; Apanius 1997)**. It contains approx 4 million nucleotide pairs and contains approx 200 genes The MHC complex is divided into three subgroups called MHC class I, MHC class II, and MHC class III. Among the genes within the MHC are more than 20 loci encoding proteins involved in binding and presentation of the peptide degradation products of proteins to the T cell antigen receptor **(Figure 2.2a & b)**.

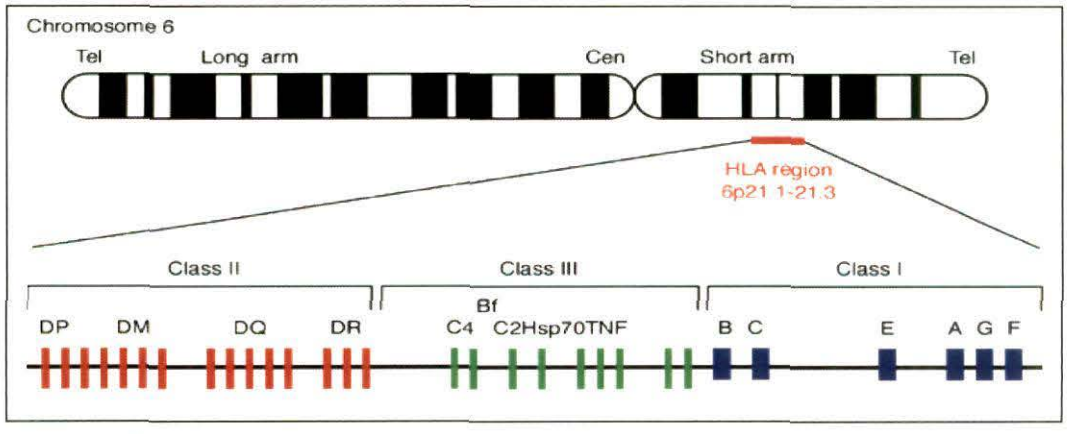
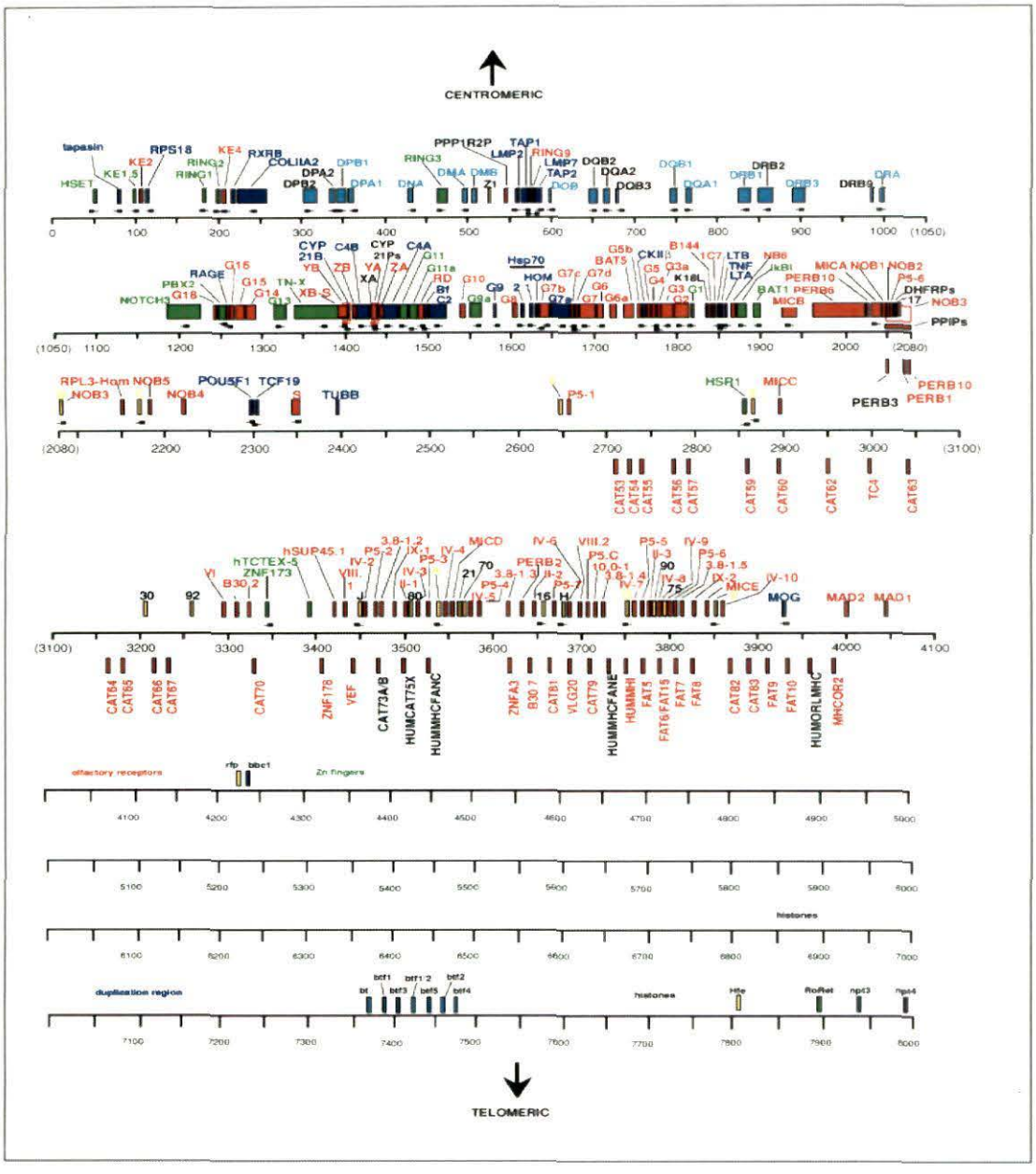



**Figure 2.2a Gene map of the human leukocyte antigen (HLA) region**




**LEGEND**

- Class II sequences
- Class I sequences
- Genes of known function
- Genes of putative function
- Genes of unknown function

**Figure 2.2b Gene map of the human leukocyte antigen (HLA) region**

There is not a definite candidate for the primordial MHC gene. According to one hypothesis the class II MHC evolved first **(Hughes 1993)** whereas another hypothesis holds that the class I MHC originated first as a result of a recombination between an immunoglobulin-like C-domain and the peptide-binding domain of an HSP70 heat-shock protein **(Flajnik,1991)**

A phylogenetic analysis supports a relationship between the class II MHC alpha chain and beta 2-microglobulin and between the class II MHC beta-chain and the class I alpha chain **(Hughes, 1992)**. Most evidence supports the hypothesis that the ancestral MHC molecule had a class II-like structure and it gave rise to the class I molecule **(Hughes,1992; Lawlor 1990; Klein,1990)**

The MHC class I encodes heterodimeric peptide binding proteins, as well as antigen processing molecules such as TAP and Tapasin. The MHC class II encodes heterodimeric peptide binding proteins with help of specialized chaperone invariant chain (Ii chain) The MHC class III region encodes for other immune components, such as complement components (e.g., C2, C4, factor B) and some that encode cytokines (e.g., TNF-alpha). Class I antigens are expressed on all somatic nucleated cells, where as class II is expressed on restricted population of cells, which are usually capable of antigen presentation. The complement system and other soluble molecules constitute the class III component of the MHC molecule (**Figure 2.3**).

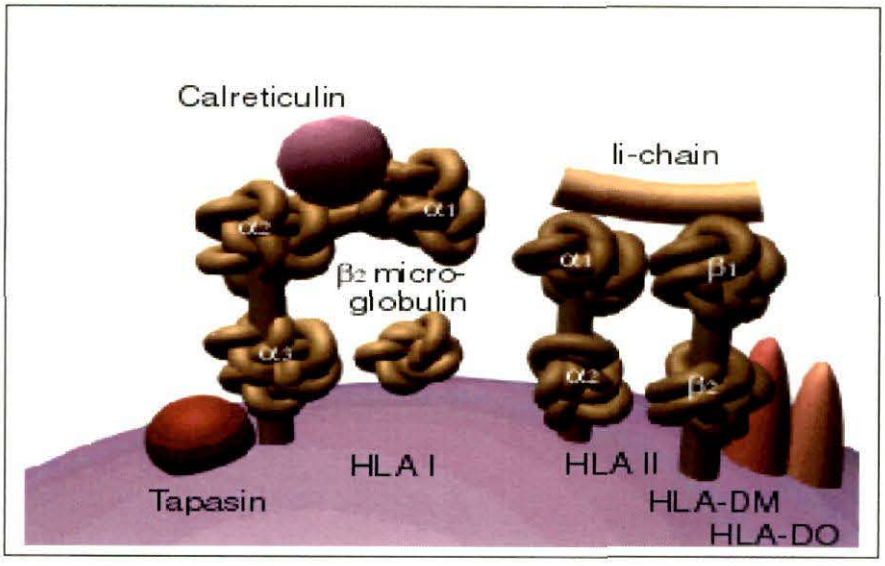



**Figure 2.3 Molecular structure of HLA class I and class II antigens**

The MHC proteins act as signposts that display fragmented pieces of an antigen on the host cell's surface (**Snell,1981**). They may be self or nonself. If they are nonself, there are two ways by which the host cell may acquire this antigen. If the host is a macrophage or microphage, such as a monocyte or neutrophil, it may engulf the particle (bacterial, viral, or particulate matter), break it apart using lysozymes, and display the fragments on class II MHC molecules. On other hand, if a host cell is infected by a bacteria or virus, or is cancerous, it may display the antigens on its surface with a class I MHC molecule (**Figure 2.4**).

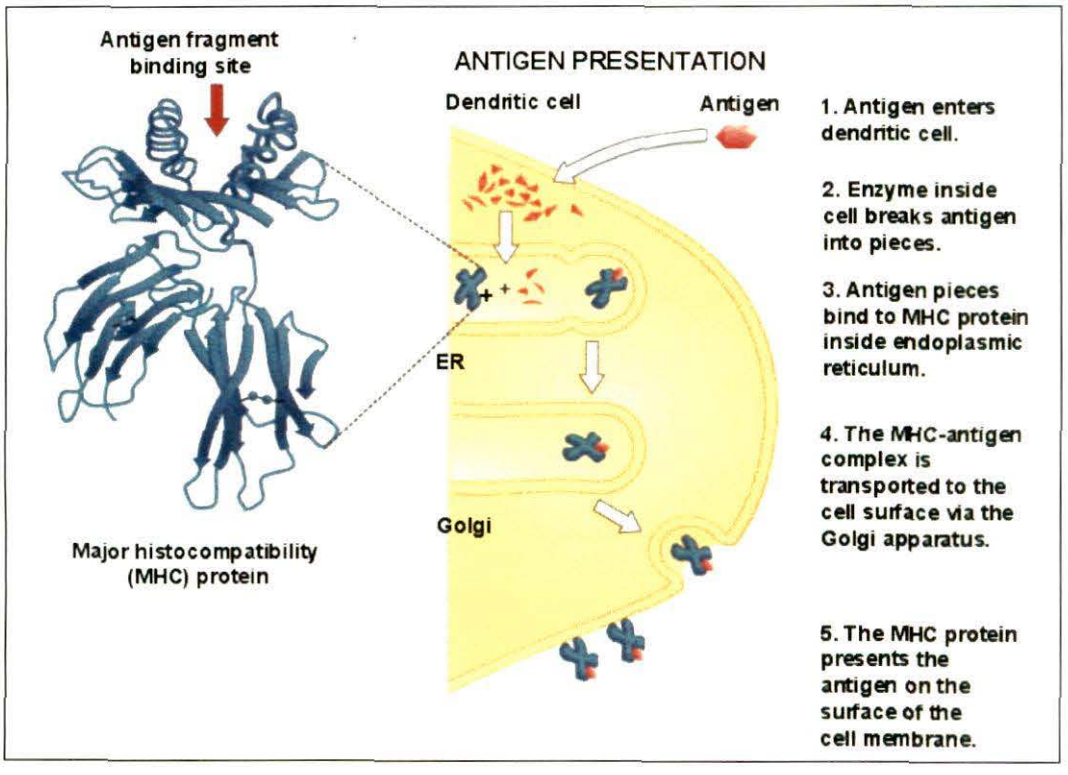



**Figure 2.4 Mechanism of MHC processing**

Viruses and cancerous cells have a tendency to display unusual, nonself antigens on their surface. These nonself antigens, regardless of which type of MHC molecule they are displayed on, will initiate the specific immunity of the host's body.

The best-known genes in the MHC region are the subset that encodes cell-surface antigen-presenting proteins. In humans, these genes are referred to as human leukocyte antigen (HLA) genes, although people often use the term MHC to refer to

HLA gene products) **(Trowsdale 1995; Gruen & Weissman,1997).** In mouse it is H-2 (Histocompatibility System – 2), in rabbits it is RLA (Rabbit Leukocyte antigens), Guinea Pig it is GLA (Guinea Pig leukocyte antigens) in Chimpanzee ChLA (Chimpanzee Leukocyte Antigen) in Cattle BoLA (Bovine Leukocyte Antigens).

In humans the most intensely studied HLA genes are the nine so-called classical MHC genes: HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1. The A, B, and C genes belong to MHC class I, whereas the six D genes belong to class II.

The MHC molecules have a vital role in the complex immunological dialog that must occur between T cells and other cells of the body **(Benacerraf & Devitt, 1972).** At maturity, MHC molecules are anchored in the cell membrane, where they display short polypeptides to T cells, via the T cell receptors (TCRs) **(Janeway, 1997).** The polypeptides may be self, that is, originating from a protein created by the organism itself, or they may be foreign, originating from bacteria, viruses, pollen, etc.

The immune system has another and equally important method to identify antigen: B cells with their membrane-bound antibodies, also known as B cell receptors (BCRs). However, whereas the BCRs of B cells can bind to antigens without much outside help, the TCRs of T cells require "presentation" of the antigen. It is important to realize that, during the vast majority of the time, MHC is kept busy presenting self-peptides, which the T cells should appropriately ignore. A full-force immune response usually requires the activation of B cells via BCRs *and* T cells via the MHC-TCR interaction. All MHC molecules receive polypeptides from inside the cells they are part of and display them on the cell's exterior surface for recognition by T cells. However, there are major differences between MHC class I and II in the method and outcome of peptide presentation.

Besides being scrutinized by immunologists for its pivotal role in the immune system, the MHC has also attracted the attention of many evolutionary biologists, due to the high levels of allelic diversity found within many of its genes

**MHC Class I**

MHC class I molecules are found on almost every nucleated cell of the body except central nervous system, skeletal and smooth muscle cells, parathyroid cells, pancreatic cells and corneal epithelium. Both male and female germinal cells are also devoid of classical class I antigens. The placenta and extra villous membranes express non-classical class I antigens **(Parham 1996)**. MHC class I molecules are heterodimers, consisting of a single transmembrane polypeptide chain (the α-chain) which is about 44 kiloDaltons (kD) and a $\beta_2$ microglobulin is 12 KD protein alpha chain is non-covalently linked with Beta chain **(Figure 2.5)**. The heavy α chain comprises of domains, α1, α2 and α3. β2 micro-globulin is also found as a soluble protein in serum, transmembrane region and a cytoplasm. Close to plasma membrane is located α3 domains.

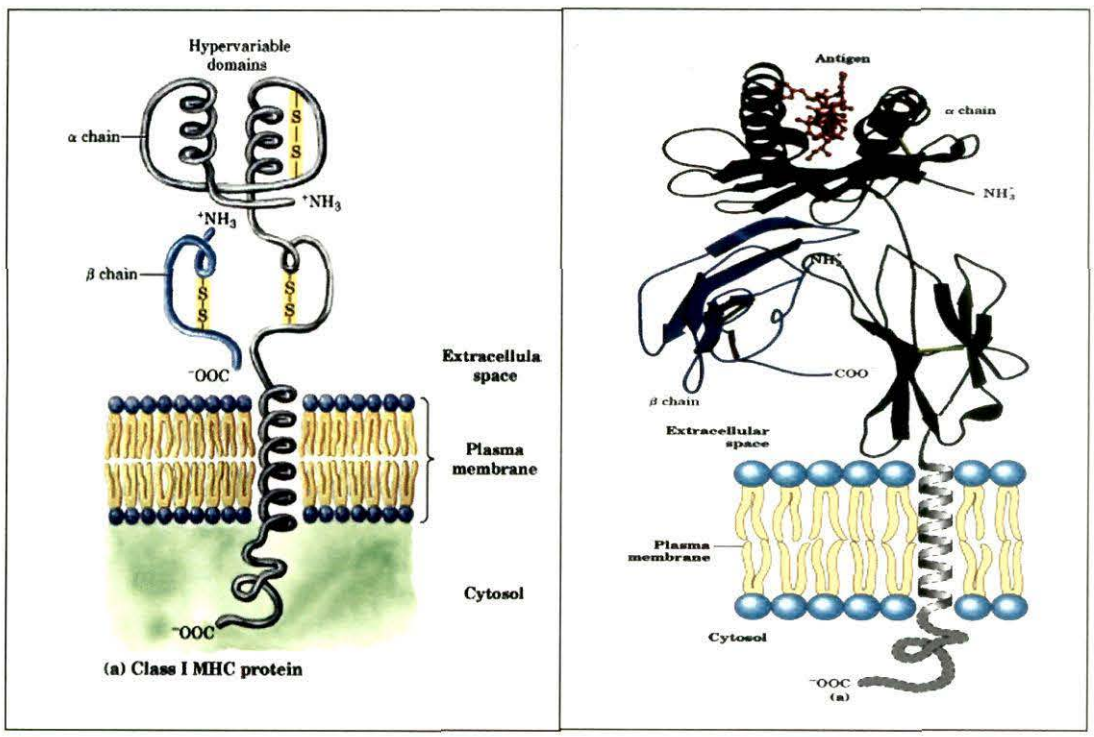



**Figure 2.5 Structure of MHC class I antigens**

In all the three domains 90 amino acids are present seperately, α2 and α3 domains are linked with inter chain disulphide bonds. In α3 domains at position 86 A glycosylated asparagine residue is situated. In transmembrane region 23 hydrophobic amino acid residues are present due to which α-helical conformation is

achieved. To the C terminal of the membrane arginine and lysine forms a cluster. These then get linked with the polypeptide chain in the membrane by interacting with negatively charged phospholipid groups of the inner membrane. The hydrophobic cytoplasmic domain consists of 30 amino acids; of which 50 are polar amino acids particularly serine and some are phosphorylated by a cAMP dependent protein kinase.

**Mechanism of MHC Class 1 Processing**

MHC proteins must bind peptide, and Class I must be complexed with $\beta_2$-microglobulin in intracellular compartments before MHC can be expressed on the cell surface. Although many MHC alleles have been identified in the human population, each individual has a limited number of MHC proteins with which to present a great many pathogen epitopes to T cells. Peptide binding to MHC is less specific than epitope binding to Ig or TCR; each MHC presents many different epitopes. Peptide must bind MHC with enough affinity to be retained on the plasma membrane and not exchange with soluble peptide. MHC molecules are unstable in the absence of bound peptide and are folded around peptide before transport to the plasma membrane (**Petrovsky, 2004; Terhorst ,1976**).

A virus-infected cell synthesizes virus proteins on ribosomes in its cytoplasm. In order to be presented, these proteins must be broken down into short peptides and transported into the endoplasmic reticulum (ER) to bind to newly synthesized Class I MHC proteins. In the cytosolic processing pathway, cytosolic proteins are degraded to peptides in proteasomes, cylindrical arrays of proteolytic enzymes with their active sites towards the center of the cylinder. Both pathogen proteins and self cell proteins can be complexed with ubiquitin to target them to the proteasome for processing. Two proteases encoded in the MHC Class II region (LMP2 and LMP7) and a third subunit not encoded in MHC is produced in response to interferon, which is synthesized in response to virus infection. These inducible proteases replace constitutive proteases in the proteasome and produce peptides with basic and hydrophobic carboxyl terminal residues preferred as anchor residues in Class I peptide binding sites and for transport from the cytosol into the ER. Two polypeptide

chains. Transporters of Antigen Peptide TAP-1 and TAP-2, are present in the ER membrane with ATP-binding domains on the cytosolic side and hydrophobic transmembrane domains spanning the ER membrane. The TAP-1/TAP-2 complex transports cytosolic peptides into the lumen of the ER with the expenditure of ATP. Both TAP molecules are required for membrane expression of Class I. Newly synthesized, partly folded Class I MHC $\alpha$ chain binds the chaperone calnexin in the lumen of the ER. Chaperone binding prevents mis-folding that otherwise would occur in the absence of antigen peptide. When $\beta_2$-microglobulin binds to Class I $\alpha$ chain, calnexin dissociates and Class I $\alpha$ plus $\beta_2$-microglobulin form a complex with calreticulin, tapasin, and TAP transporter. When Class I MHC binds peptide, it is released from TAP transporter and the Class I MHC-peptide complex is transported through the Golgi to the plasma membrane. Unbound peptides are thought to be transported back into the cytoplasm for reprocessing and retransport. In uninfected cells, membrane Class I MHC presents self peptides. Viruses which can interfere with TAP function or with transport of Class I MHC to the plasma membrane can evade destruction by cytotoxic T cells (**Petrovsky, 2004; Neumann, 1997**) **(Figure 2.6)**.

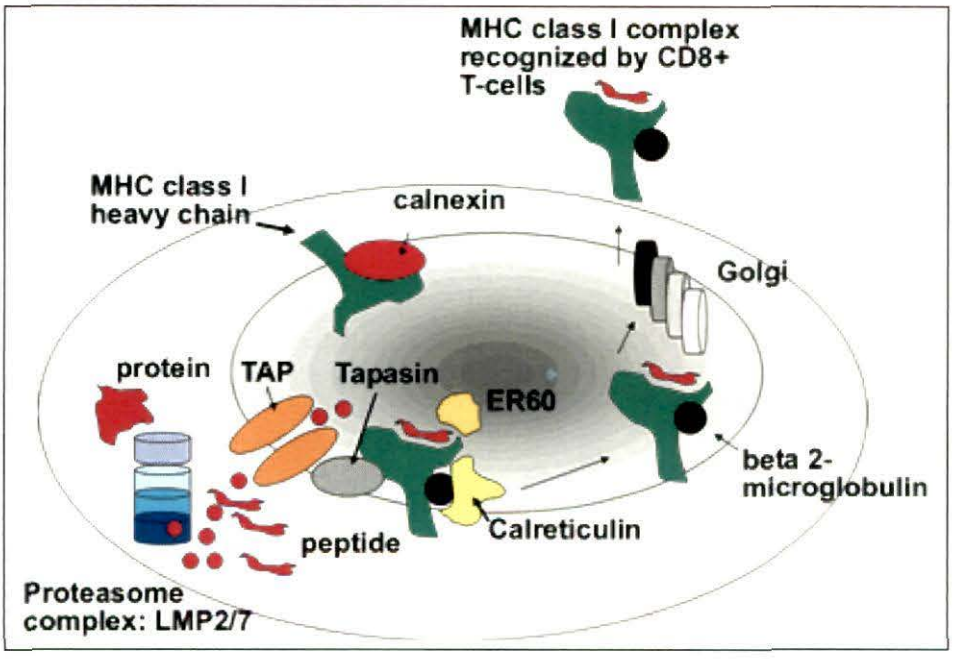



**Figure 2.6 Mechanism of MHC Class 1 Processing**

Exogenous antigen is processed in the endosomal processing pathway. Bacteria, soluble protein antigens, and antibody-coated viruses which have been taken up by macrophages and B cells, envelope proteins from the plasma membrane of DC, and some bacteria and parasites that live in endosomal vesicles, enter the endosomal processing pathway. The endosomes become increasingly acidic as they move from the plasma membrane farther into the cytoplasm. Increased acidity activates proteases that cut the antigen into peptides.

**MHC Class I Gene Expression**

Various biological or chemical modulators can modulate MHC gene expression. The promoter region of class I genes can be activated through several pathways. Hence, these genes are not typical domestic genes. Extensive studies on regulation of MHC class I expression, using transfection techniques and transgenic animal models, have resulted in identification of various cis-acting sequences involved in positive and negative regulation of class I genes. Work is in progress to identify the transacting proteins that bind to these sites and to delineate the mechanisms that regulate constitutive and inducible expression of class I genes in normal and diseased cells. Various biological molecules (IFN, GM-CSF, IL-2) and other chemicals up-regulate the MHC expression. If the exact mechanisms are known by which the expression of class I genes is up regulated, the efforts can be made to balance the beneficial and toxic effects of biological molecules with one another, which may facilitate the use of combination of these molecules in subpharmacological doses (to eliminate toxicity) for early and better management of neoplastic diseases, as it is well-known that during malignancy MHC gene expression is down-regulated **(Agrawal & Kishore,2000; Neumann,1997).**

**MHC Class II**

MHC Class II molecules are found only on a few specialized cell types, including macrophages, dendritic cells, activated T cells, and B cells, all of which are professional antigen-presenting cells (APCs). Like MHC class I molecules, class II molecules are also heterodimers, but in this case consist of two homologous

peptides, an α and β chain. α chain is of 34,000 Daltons and β- chain is of 28000 dalton. These are independently synthesized in the endoplasmic reticulum and become associated with a third type of polypeptide chain, the invariant chain (Ii chain) **(Ceman 1995; Schafer 1995)**. The α-chains are encoded by DRα, DQα, DPα etc. genes whereas β-chains are encoded by DRβ, DQβ, DPβ etc. genes. Each DR, DQ or DP sub-region consists of at least one expressed α and one expressed β gene. Both α and β genes can be polymorphic but most of the polymorphism resides in β genes. All the class II genes lie within the MHC complex **(Parham and Ohta, 1996; Terhorst,1976; Zhong, 1997)**.

The α and β chains consist of 2 extra cellular domains of approximately 90 amino acids each. In the human DR region there is more than one functional β chain gene; the number of genes varies with different class II haplotypes. All these products can associate with the α chain to form class II cell surface molecules, this providing a mechanism for increasing the polymorphism of the DR molecule **(Figure 2.7)**.

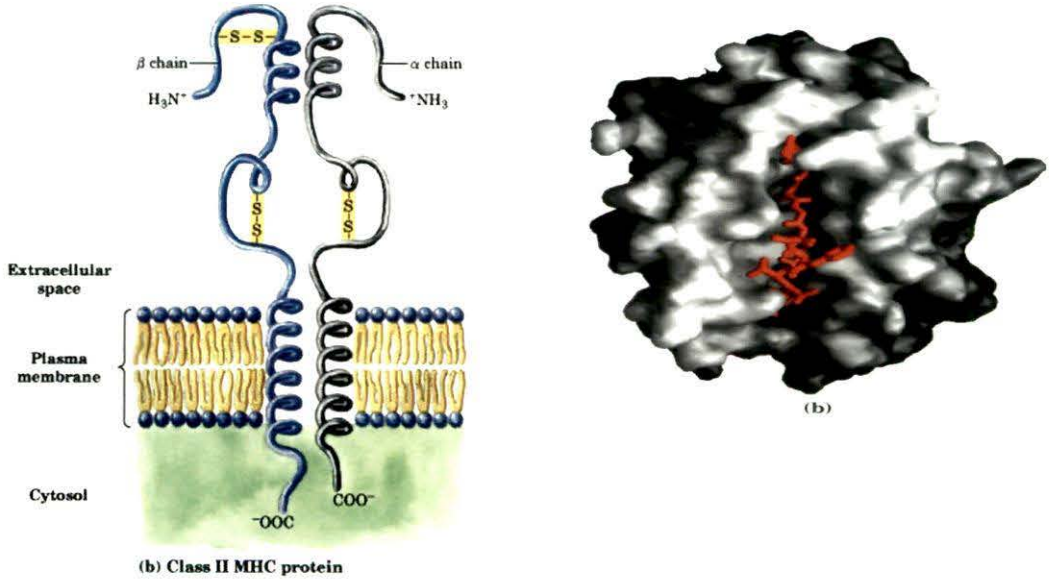



**Figure 2.7 Structure of MHC class II antigens**

**Mechanism of MHC Class II processing**

Class II MHC α and β chains are synthesized on the rough endoplasmic reticulum (ER) and transported into the ER lumen, where they assemble with another polypeptide, invariant chain (Ii). Ii and Class II MHC form nine-chain trimeric complexes, with part of each Ii molecule occupying the Class II MHC peptide-binding site. Ii allows Class II MHC to assemble in the absence of foreign peptide and blocks association with normal cell (self) and endogenous peptides present in the ER lumen. Ii also directs the transfer of Class II MHC through the Golgi to a specialized vesicular compartment **(Ceman, 1995)**. MHC (MHC Class II Compartment), where over several hours Ii is degraded by cathepsins and proteases in the low pH. Occupying the peptide –binding groove called CLIP. HLA-DM, a Class II MHC αβ heterodimer resembling other Class II MHC molecules but not expressed on the cell surface, facilitates CLIP removal and peptide binding in the MIIC compartment **(Figure 2.8)**.

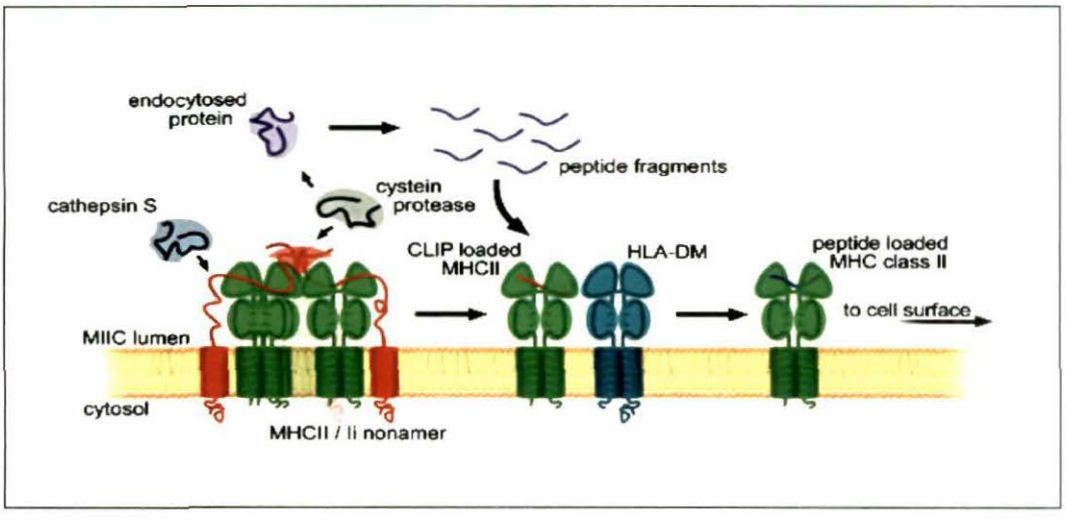



**Figure 2.8 Mechanism of MHC Class II Processing**

Class II MHC-peptide complex is then transported to the plasma membrane. Class II MHC which does not bind peptide when CLIP dissociates is unstable and is rapidly degraded. In the absence of infection, APC present Class II MHC containing self peptides, including self Class II MHC peptides. Peptide/Class II MHC

complexes are very stable, ensuring that the APC presents its own exogenous peptides and not peptides released from another cell. Like other plasma membrane proteins, Class II MHC proteins shuttle between the membrane and endosomal compartments where they can pick up new exogenous antigen peptide or be degraded (**Zhong,1997**).

**MHC Class II Gene expression**

The highly complex pattern of expression of major histocompatibility complex class II (MHC-II) molecules determines both the immune repertoire during development and subsequently the triggering and the control of immune responses. These distinct functions result from cell type-restricted expression, developmental control and either constitutive or inducible expression of MHC-II genes , in these various situations, MHC-II gene expression is always under the control of a unique transactivator, CIITA, CIITA gene is controlled by several distinct promoters, two of which direct specific constitutive expression in dendritic cells and B lymphocytes respectively, while another mediates γ interferon-induced expression. Thus the cellular, temporal and functional diversity of MHC-II expression is ultimately controlled by differential activation of different promoters of a single transactivator gene (**Sonntag 2000; Ruemmele,1999; Mehindate,1994**).

**Nomenclature**

Developing and maintaining a nomenclature for the numerous and rapidly accumulating Class I and II alleles is a challenging problem. The nomenclature for alleles is largely based on earlier serological names since the broad HLA antigen groups were originally defined based on their reaction with antisera in complement mediated microlymphocyotoxicity assays. Antisera were often isolated from women sensitized during pregnancy to the HLA specificities encoded by paternal haplotypes. Serological specificities are referred to as HLA-A1 through A80, HLA-B7 through B81; HLA-Cw1 through Cw10; HLA-DR1 through DR18, and HLA-DQ1 through DQ9 (**Table 2.2**) (**Bodmer,1998**).

**Table 2.2 HLA loci and known alleles**

| Generic locus | Antigen or associated specificity | Number of known alleles |
|---|---|---|
| HLA-A | A1 to A80 | 489 |
| HLA-B | B7 to B81 | 830 |
| HLA-C | Cw1 to Cw10 | 266 |
| HLA-E | - | 9 |
| HLA-F | - | 20 |
| HLA-G | - | 23 |
| DRA | DR1 to DR18 | 3 |
| DRB1 | - | 463 |
| DQA1 | DQ1 to DQ9 | 34 |
| DQB1 | - | 78 |
| DPA1 | DPw1 to DPw6 | 23 |
| DPB1 | - | 125 |

**Source: http://www.ebi.ac.uk/imgt/hla/**

Once nucleic acid-based information regarding the sequence of alleles became available, a nomenclature complementary to the serological terms was devised (**Table 2.3**).

The first two digits of an allele name refer to the underlying serological specificity and the third and fourth digits indicate a specific allele sequence. For example, HLA-A*0205 and A*0210 are alleles encoding distinct polypeptides within the A2 serotype. These two alleles both encode the epitope recognized by the anti-A2 antisera but have 5 nucleotide differences elsewhere in exons 2–3 resulting in amino acid variations. Alleles within a serological group may vary from each other by a single or by several nucleotides. For Class II molecules, both the A and the B genes may contribute to antigen variability.

**Table 2.3 Nomenclature of HLA alleles**

| HLA Allele Nomenclature | Examples of alleles Comment |
|---|---|
| A*24 and A*2404 | A*24 refers to any of 33 known alleles with closely related sequences encoding Class I antigens which usually react with A24 anti-sera. A*2404 is a specific allele within this group. |
| DRB1*0801 and *0805 | Differ in exon 2 at codon 74: At this position *0801 has CTG and *0805 has GCG encoding Leu and Ala, respectively |
| A*01011 and *01012 | Differ by a silent polymorphism at codon 142: At this position *01011 has ATC and *01012 has ATT. Both encode Ile |
| B*1501101 and *1501102N | The B*1501102N (null) allele has a 10-bp deletion near the 3'end of intron 1. The mRNA is improperly spliced with a predicted truncated translated polypeptide. |

Thus a DR15 serotype may be found in an individual with one of the DRB1*15 alleles such as DRB1*1501 and DRA*0101. When necessary, a fifth digit is used to identify silent polymorphisms and the sixth and seventh digits are used to denote variation occurring outside of coding regions, such as the promoter and introns. Null allele sequences which result in either no or reduced levels of functional HLA molecules because of transcription changes, aberrant RNA splicing, and frame shift and nonsense mutations, or in frame termination codons are designated by an allele number appropriate to the group and the letter N. Individuals with null alleles may have discrepancies between serological and DNA-based typing. This problem has clinical implications since an inappropriate donor may be sought if more than one laboratory using different techniques are involved in typing a recipient and potential donors.

**The origin of diversity of MHC alleles**

The major histocompatibility complex (MHC) loci are known to be highly polymorphic in humans, mice and certain other mammals, with heterozygosity as high as 80-90%. Four different hypotheses have been considered to explain this high degree of polymorphism:

(1) A high mutation rate

(2) Gene conversion or inter locus genetic exchange

(3) Over dominant (balancing) selection

(4) Frequency-dependent selection.

The distribution of the pattern of sequence polymorphism in human and mouse class I genes provides evidence for four co-ordinate factors that contribute to the origin and sustenance of abundant allele diversity that characterizes the MHC in the species. These include: (a) a gradual accumulation of spontaneous mutational substitution over evolutionary time but not an unusually high mutation rate; (b) selection against mutational divergence in regions of the class I molecule involved in T cell receptor interaction and also in certain regions that interact with common features of antigens; (c) positive selection pressure in favor of persistence of polymorphism and heterozygosity at the antigen recognition site; and (d) periodic intragenic (interallelic) and more rarely, intergenic, recombination within the class I genes.

Evolutionary interplay between mutation and recombination varies with MHC locus, and even for subregions of the same gene **(Parham & Ohta 1996; Hughes, 1988; Hughes, 1989)**. For example, phylogenetic inferences based on the exon 2 region of HLA-DRB loci are complicated by selection and recombination (gene conversion). Noncoding region analysis may help clarify patterns of allele evolution usually with contrasting results to those obtained from coding region analyzes **(Hickson, 1997)**. The main source for the variability in the HLA gene sequences is point mutation but the mutation rate is by no means higher in the MHC than elsewhere in the genome **(Lawlor, 1988, Parham ,1995)**. Because of trans species polymorphism, accumulation of point mutations over millions of years results in extensive polymorphism. In contrast, gene conversions have produced at

least 80 new class I alleles since the separation of the Homo lineage and the rate of conversion is much higher than that of point mutation **(Little 1999; Marsh 2000)** **(Figure 2.9)**.

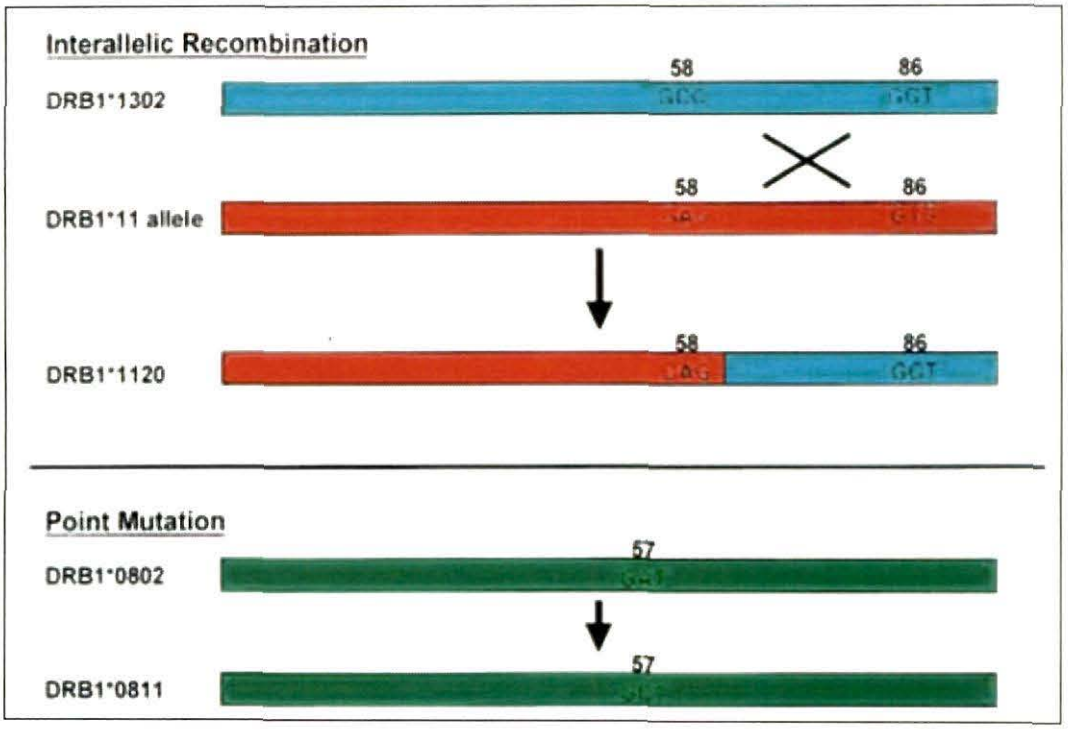



**Figure 2.9 Mechanism of generation of new alleles**

Alleles arise from existing alleles through several postulated mechanisms **(Marsh & Parham,2000).** DRB1*1120 likely arose via interallelic recombination between DRB1*1302 and a *DRB1**11 allele **(Cizman & While,1996)** DRB1*0811 probably is derived from DRB1*0802 by a point mutation **(Williams,1994)**.

**Nature of Class I and II Gene Polymorphism**

Multiple alleles are found within most of the known serotypes, although a few serotypes (an example is DR9, *DRB1**09012) are accounted for solely by a single allele. For example the B35 serotype has more than 39 alleles. HLA allele frequencies exhibit ethnic variation, with some alleles found widely distributed among populations and others almost exclusively within a particular ethnic group. The number of different phenotypes that are possible from all combinations of the known HLA alleles is greater than the earth's population. However, the Class I and

II loci reside on a relatively small region of chromosome 6 and specific haplotypes were apparently present at high frequencies in founding populations or were selected for or against by infectious organisms. In this setting, linkage disequilibrium results in a significant overrepresentation of certain haplotypes **(Alper 1992) (Figure 2.10).**

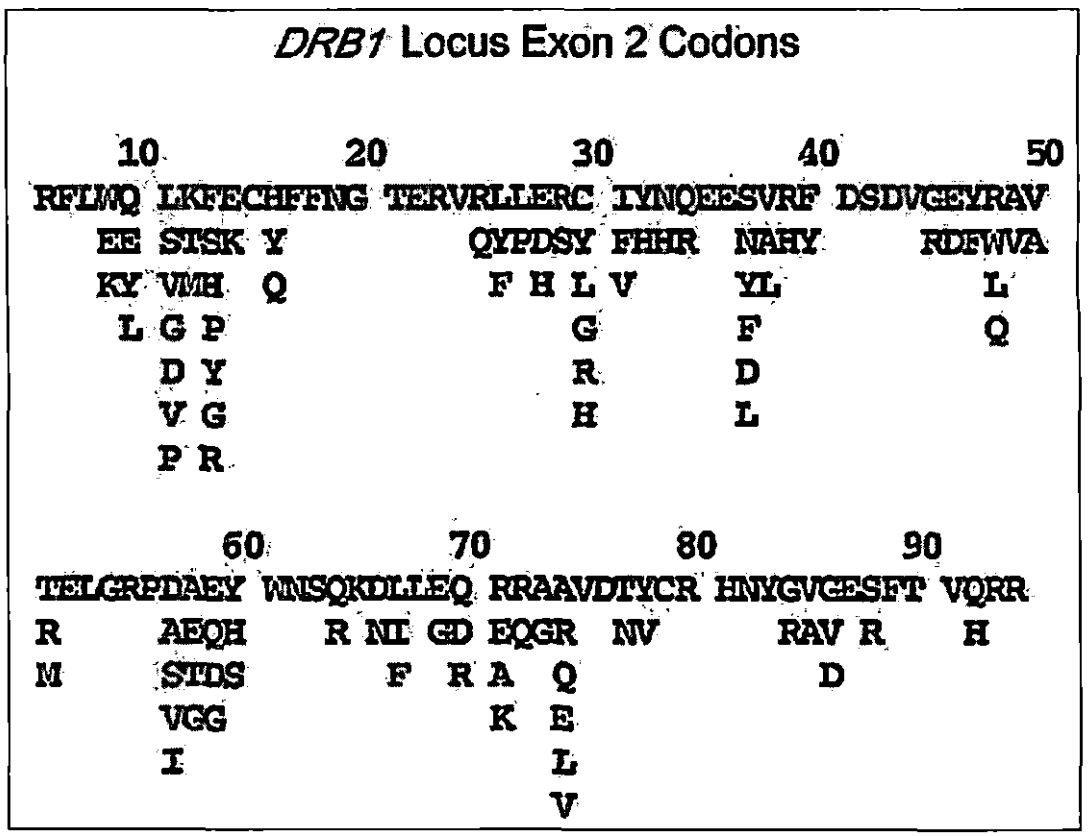



**Figure 2.10 DRB1 Gene polymorphism**

Single letter amino acid codes are shown for DRB1 exon 2 codons 6–94. About half of the positions are invariant while the remainder displays polymorphism with a few codons encoding as many as seven different amino acids. For example, all *DRB1* alleles have glycine encoded by position 20 while alleles may encode glycine, valine, or aspartic acid at codon 86. *DRB1* alleles arise through the many possible combinations of these polymorphisms.

This fact makes it possible for organizations such as the National Marrow Donor Program to have a fairly high likelihood of finding a well matched donor for a Caucasian recipient with 3–4 million potential volunteer donors in its database **(Kernan,1993).**

Although HLA gene loci are closely linked, meiotic recombination may occur between loci. For example, the crossover rate is approximately 0.8% between the *A* and *B* loci and virtually zero between the *DRB1* and *DQB1* loci. The great majority of the polymorphism found in the Class I and II genes occurs in the exons that encode the α-1 and α-2 (Class I, exons 2–3) and the α-1 and β-1 (Class II, exon 2) domains which bind processed peptides **(Little,1999)**.

Some nucleotide positions in these exons are invariant, others may have two or three or even all four of the possible bases as possibilities. Thus, some codons are constant while others display varying degrees of variability. The polymorphic exons are relatively short in length (about 250 nucleotides), they can easily be amplified in the PCR for molecular diagnostic studies.

**Mechanisms maintaining the extreme polymorphism of the MHC**

**1. Pathogen driven mechanism**

Pathogen-driven selection favors genetic diversity of the MHC through both heterozygote advantage (over dominance) and frequency-dependent selection (Potts WK1993). Selection is thought to favor rare MHC genotypes, since pathogens are more likely to have developed mechanisms to evade the MHC-dependent immunity encoded by common MHC genotypes. Six molecular models of pathogen-driven selection have been presented **(Potts 1995)**.

A. Pathogen Evasion Models

Escape of a single T-cell clone recognition

Escape into holes in the T-cell repertoire produced by T-cells energized by pathogen variants

Escape into holes in the T-cell repertoire induced by self-tolerance

Escape of MHC presentation

B. Host-Pathogen Interactions:

Heterozygote advantage

Pathogens bearing allo-MHC antigens

MHC associations with specific infectious diseases have been difficult to demonstrate. The best known one is malaria in humans **(Hill, 1991)**. Since most

infectious agents have multiple epitopes which MHC has to deal **(Fienberg,1970)**, Rather than resistance of specific heterozygous genotypes to specific. In human heterozygote advantage have been reported for a specific genotype in HIV infection **(Carrington,1999)** and in HBV infection **(Thursz,1999; Thio,1999)**

**2. Non-pathogen driven mechanisms**

MHC is exploited to discriminate against genetic similarity at highly polymorphic loci to avoid inbreeding. MHC-based disassortative mating would produce heterozygous progeny at least at the MHC which would result in increased fitness **(Ober,1997).** Progeny derived from MHC-dissimilar parents would have high fitness because of reduced levels of inbreeding depression and increased resistance to infectious diseases due to high MHC heterozygosity. This selection contributes to the high levels of genetic polymorphism observed at the MHC loci.

**Evidence for Selection on MHC alleles**

1. One important feature of the MHC genes is that the ratio of non-synonymous (replacement) to synonymous (silent) substitutions (dn /ds ratio) is very high in the codons encoding the antigen recognition site of polymorphic class II molecules compared to other codons **(Hughes, 1994).** This pattern is evidence that the polymorphism at the antigen recognition sites is maintained by over dominant selection of which the most common form is heterozygote advantage. This kind of selection has been noted for all expressed DRB genes including DRB3 and DRB4 **(Klein, 1991).** This feature and the others such as (1) an extremely large number of alleles; (2) ancient allelic lineages that pre-date contemporary species (trans-species evolution) and; (3) extremely high sequence divergence of alleles make the MHC a unique system in the whole genome.
2. The expected number of alleles under neutrality is far lower than the number of MHC alleles observed in natural population which indicates that some form of balancing (diversifying) selection is also acting **(Hedrick, 1983, 1994, 1996).** For a subdivided population over a large range of migration

rates, it appears that the number of self-incompatibility alleles (or MHC-alleles) observed can provide a rough estimate of the total number of individuals in the population but it underestimates the neutral effective size of the subdivided population **(Schierup,1998, 2000, 2002).**

3. The large number of alleles showing a relatively even distribution is against neutrality expectations and indicates that diversifying, and not simply directional, selection operates in contemporary populations.
4. The observed deficiency of homozygotes in humans **(Black, 1997, 1981; Hedrick,1990)** indicates that selection favors heterozygotes, because of high immune response. When the amino acid heterozygosities per site for HLA-A and -B loci were determined, for the 54 amino acid sites thought to have functional importance, the average heterozygosity per site was 0.301. Sixteen positions have heterozygosities greater than 0.5 at one or both loci and the frequencies of amino acids at a given position are very even, resulting in nearly the maximum heterozygosity possible. High heterozygosity is concentrated in the peptide-binding sites, whereas the sites that interact with the T-cell receptor have lower heterozygosity. Overall, these results indicate the importance of some form of balancing selection operating at HLA loci, maybe even at the individual amino acid level **(Hedrick,1996).**
5. The observed linkage disequilibrium among tightly linked MHC genes suggests that the strength of selection is uneven within the MHC **(Apanius, 1997).**
6. Studies in West Africa showed that resistance against malaria is HLA-B53 associated and this is the reason for an increased frequency of B53 in that area. The selection differential for HLA-B*5301 is estimated to be 0.028 **(Hill 1991). Apanius *et al.*, (1997)** suggested that MHC haplotypes can confer resistance to infectious diseases that outweighs the deleterious effects from autoimmunity. Another hypothesis for the maintenance of autoimmune-predisposing MHC haplotypes is that these alleles protect against initial infection, but the pathogen triggers autoimmunity through molecular mimicry.

**Allelic diversity and MHC evolution**

MHC gene families are found in essentially all vertebrates, though the gene composition and genomic arrangement vary widely. Chickens, for instance, have one of the smallest known MHC regions (19 genes), though most mammals have an MHC structure and composition fairly similar to that of humans. Gene duplication is almost certainly responsible for much of the genic diversity. In humans, the MHC is littered with many pseudogenes.

One of the most striking features of the MHC, particularly in humans, is the astounding allelic diversity found therein, and especially among the nine classical genes. In humans, the most conspicuously-diverse loci, HLA-A, HLA-B, and HLA-DRB1, have 489, 830 and 463 known alleles respectively this allelic diversity is truly exceptional in the human genome. The MHC gene is the most polymorphic in the genome. One remarkable featureof HLA Loci is that many of these alleles are quite ancient. It is often the case that an allele from a particular HLA gene is more closely related to an allele found in chimpanzees than it is to another human allele from the same gene.

Phylogenetically the marsupial MHC lies between eutherian mammals and the minimal essential MHC of birds, although it is closer in organization to non-mammals. It's Class I genes have amplified within the Class II region, resulting in a unique Class I/II region.

The allelic diversity of MHC genes has created fertile grounds for evolutionary biologists (**Arnaiz-Villena *et al.*, 1999**). The most important task for theoreticians is to explain the evolutionary forces that have created and maintained such diversity. Most explanations invoke balancing selection, a broad term that identifies any kind of natural selection in which no single allele is absolutely most fit. Frequency-dependent selection and heterozygote advantage are two types of balancing selection that have been suggested to explain MHC allelic diversity.

**Population studies**

Molecular HLA-typing has proved to be an invaluable tool in studying the evolutionary origin of human populations (**Arguello *et al.*, 1998; Arnaiz-Villena *et***

***al.*, 1999; Luo *et al.*, 1999; Albis- Camps & Blasczyk 1999**). This information, in turn, contributes to the understanding of cultural and linguistic relationships and practices among and within various ethnic groups. This has become possible due to the continual discovery of new HLA alleles using DNA technology which has increased the power of HLA to distinguish individuals (**Arguello *et al.*, 1998; Arnaiz-Villena *et al.*, 1999; Luo *et al.*, 1999; Albis- Camps & Blasczyk 1999**). It has been postulated that gene conversion events are the main mechanisms for distributing and reshuffling sequences among alleles. In addition reciprocal recombination and point mutations have been suggested to be responsible for the generation of alleles over evolutionary time **(Mason & Parham 1998; Hughes & Nei 1992).**

Some indigenous populations (e.g. groups of Native Americans or from Papua New Guinea) show a very restricted diversity of alleles at DRB1 as well as other HLA loci **(Inman & Rudin 1997).** The extensive polymorphism of the major histocompatibility complex (MHC) genes in humans and the differential allelic distribution in ethnic populations of varied origin has been major focus of immuno genetic research. The presence of certain alleles with high frequency only in specific populations (e.g. A36, A43 African Americans) and the strong linkage disequilibrium between HLA neighboring loci, demonstrates that certain combinations of contiguous alleles (HLA haplotypes) show a characteristic frequency or are distinctive in certain living populations **(Browning & McMichael 1996).** The wide range of allelic diversity and the conserved combinations of different alleles are used as genetic markers and anthropological data is based on the information supplied by population studies **(Direskeneli 2000).** Population studies indicate that there are many alleles and DR-DQ haplotypes that appear to be specific for given ethnic group. The existence of ancestral haplotypes implies conservation of large chromosomal segments.

The extra ordinary power of this small segment of the human genome clusters population in a manner expected from linguistic, anthropological and archaeological evidences. Sequences reveal a dramatic level of diversity. Species specific residues i.e. residues that identify a MHC molecule as belonging to a

particular species, are extremely rare. In contrast, species unique residues i.e. residues that were not characteristic of a species are unique to individuals of that species appear somewhat more frequently. The allele frequency distributions and patterns of variability within the molecule, suggest strong selective forces acting on class II loci and I. The form of selection is unknown and potential selective mechanisms should be examined in the light of classical population genetic theory, which states that it is very difficult to maintain so many alleles even with strong balancing selection. Thus, HLA variability at the level of DNA is useful in unraveling the evolutionary relationships between populations and in investigating the evolutionary forces which shaped the genetic profiles of contemporary populations **(Inman & Rudin,1997).**

The HLA polymorphisms have been created because of balancing selections, which maintain a few allelic lines over very long period **(Harpending,1999).** The extensive variation in HLA markers makes the system highly useful for determining genealogical relationships between populations. **Monsalve.(1999)** have successfully compared the relationship between linguistic and genetic data in Native Americans and Asian populations. They have concluded that gene flow and genetic drift are important factors in shaping the genetic landscape of Native American populations. The results are most congruent with the single migration model. In addition the understanding of the events contributing to MHC class II evolution requires comparison of species that are very closely related. The distribution of alleles in different populations can be used to construct a matrix of genetic distances between populations and a phylogenetic tree or un rooted network in order to examine the historical / evolutionary relationships between these groups.

The high polymorphism, tight linkage among loci, and the random association of alleles make the system of particular interest from the perspective of population genetics. Information on the dynamic evolutionary forces that have acted on a locus can be inferred from the number and distribution of alleles that it carries. The major histocompatibility complex (MHC) is unique in the number of highly polymorphic loci spread over such a small chromosomal region. This creates a context for interpreting HLA region variation in both evolutionary and in clinical

terms. The extensive allelic variation among the HLA class I and class II genes distinguishes these as the most polymorphic coding sequence. The distribution and dispersal of certain HLA Class II alleles in different continents of world is shown below **Figure 2.12**.

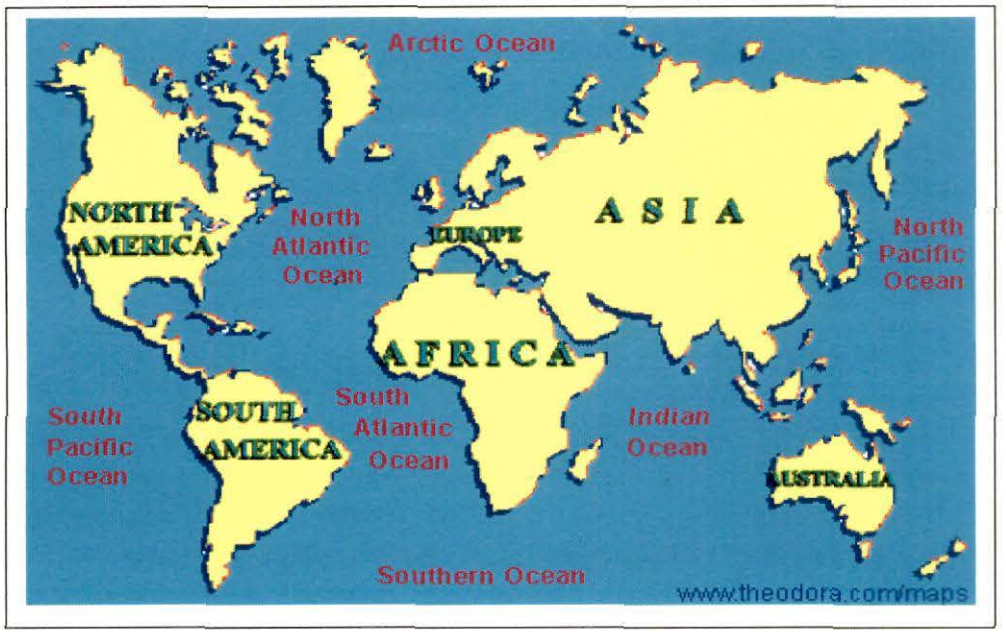



**Figure 2.11 Map of Southeast Asia, Australia and the Pacific Ocean, showing the approximate locations of the populations**

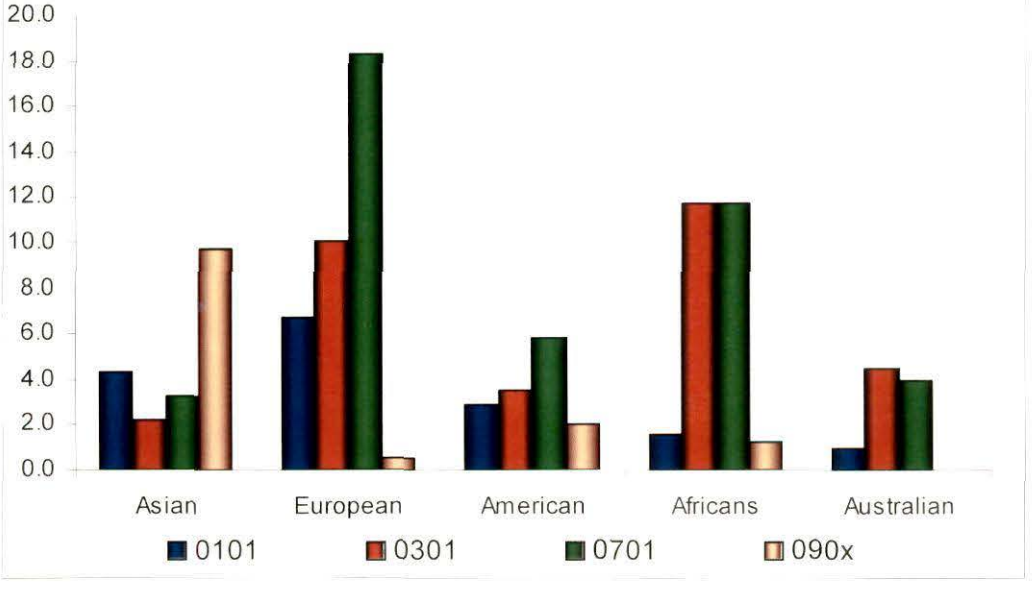



**Figure-2.12: Distribution of HLA Class II alleles in different continents of world**

It is known that HLA Class II is highly polymorphic as at DRB1 locus 463 alleles, at DQA1 locus 34 alleles and at DQB1 locus 78 alleles have been identified worldwide. The world map as per different continents is shown in **Figure 2.11.** Various alleles of class II are widely distributed in different continents of worlds like Asians, Europeans, Africans, Americans and Australians (**Figure 2.12**).

The present study has been designed to keep the enormous diversity of Indian populations and to decipher the divergence of various populations at HLA class II loci. As there are hundreds of alleles and enormous diversity, many of these alleles are quite ancient, the origin, similarity and the diversity of the Indian subcontinent can be well studied by using this highly polymorphic genetic marker.