W. Helmberg
M. Feolo

13th IHWS Shared Resources Joint Report

B2. dbMHC

Authors' address
Wolfgang Helmberg, Michael Feolo
Center Drive 45, NCBI/NLM/NIH

Bethesda MD 20892

Correspondence to:
Wolfgang Helmberg
e-mail: helmberg@ncbi.nlm.nih.gov

HLA 2003
Vol. ■: 000–000

HLA 2002
■-■

Introduction

The dbMHC is a permanent and public data resource maintained at the NCBI (Fig. 1) and is made possible through collaboration with the International Histocompatibility Working Group (IHWG; http://www.ihwg.org). The 13th International Histocompatibility Workshop (IHWS) provided an initial bolus of clinical and genetic data for dbMHC, a resource designed to provide an open platform where the HLA community will be able to submit, edit, view, and exchange MHC-related data. The dbMHC also hosts a variety of tools to support genotyping and sequence analysis of HLA, KIR, and various immune response gene polymorphisms. Current tools available at the dbMHC consist of an interactive Alignment Viewer for HLA and related genes, an MHC microsatellite database (dbMHCms), a sequence interpretation site for Sequence Based Typing (SBT), and a Primer/Probe reference database for HLA typing. We are presently in the process of creating a comprehensive database that will house a wide variety of human MHC data including:

- Detailed SNP mapping data of the HLA region
- KIR gene and haplotype data
- Extended HLA haplotype data
- Global HLA diversity/anthropology data
- Hematopoietic stem cell transplantation data
- HLA-related disease association data
- Peptide binding prediction data

All data are fully integrated with other NCBI resources as well as with the IHWG web site and are linked to the IMmunoGeneTics HLA (IMGT/HLA) database where appropriate. This article is a description of current dbMHC resources and will discuss plans for future content and development.

The NCBI database is currently accepting online submis-

Figure 1. Start page of dbMHC as www.ncbi.nih.gov/mhc.

sions for the Primer/Probe/Mix component of dbMHC. Submissions may include typing data from any of the following systems: Sequence Specific Oligonucleotide probe hybridization (SSO), Sequence Specific Primers (SSP), SSO and/or SSP mixes, HLA typing kits as containers for SSO/SSP reagents, and Sequencing Based Typing (SBT). dbMHC allows submitters to edit their submissions online at any time.

Sequence and Genotyping Resources

The following section is a summary of the main resources and tools dealing with sequences and genotyping. Detailed information can be obtained from The NCBI Handbook (1).

Alignment Viewer

The Alignment Viewer (Fig. 2) is designed to display precompiled allele sequence alignments in an aligned or FASTA format for selected loci. It offers an interactive display of the alignment, where a user can select alleles, highlight SNPs within the alignment, and change between a codon display and a display of nucleotides in blocks of 10 (a "decade"). The user can also switch the display from one where the complete sequence can be viewed to one where only the differences between selected sequences and a preselected reference sequence are displayed. All sequences displayed in the Alignment Viewer can be downloaded and formatted as an alignment, FASTA, or XML file. If the alignment option is selected, it is up to the user to define how the alignment should be organized. The alignment option allows the user to download the entire alignment, or just a section of it, and also allows the user to specify how the sequence will be displayed (e.g., in groups of 10 nucleotides or amino acids).

The allele database is based on the IMGT/HLA database and is updated in parallel with the IMGT/HLA database (2, 3). Each allele is directly linked to the detailed information of the IMGT/HLA database.

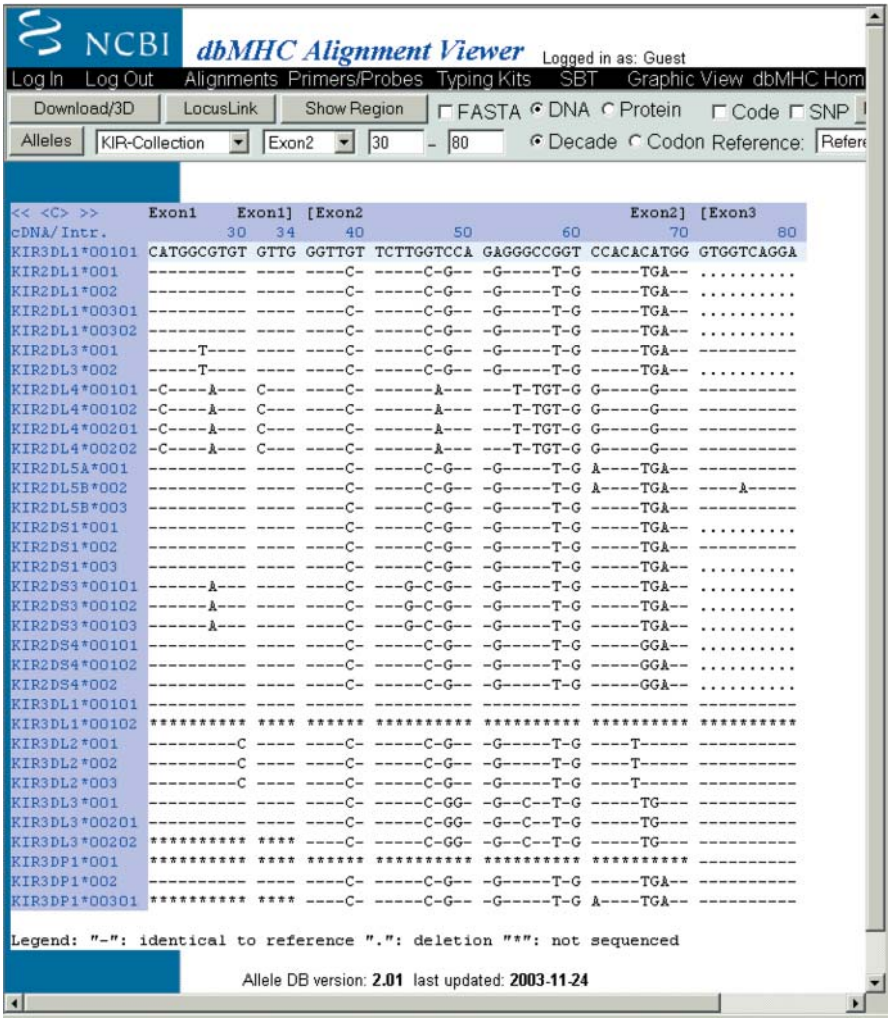

Figure 2. The Alignment Viewer, displaying KIR sequences.

Sequencing Based Typing (SBT) Interface

The SBT interface is designed to analyze sequence information provided by the user and will furnish the user with an interpretation of the sequence that will include potential allele assignments as well as a display of the query sequence aligned to existing alleles and parsed into exons and introns. Users can submit sequences to the SBT interface by copying and pasting, or by uploading them from a file. Sequences may be submitted either as haploid strands or heterozygote sequences. The current format allows input of text or FASTA formatted sequences.

Submitted sequences are aligned to reference locus sequences residing in the dbMHC allele database based on a user-defined degree of nucleotide mismatch. After analysis, the SBT interface will display exons, introns, and untranslated regions for each sequence. Allele assignments are listed according to level of sequence homology. Mismatched nucleotide positions are listed separately in the Alignment Viewer, which is located in the lower frame of the SBT interface. The SBT interface is capable of analyzing one or more sequences for a single locus. The allele database is based on the IMGT/HLA database and is updated in parallel with the IMGT/HLA database.

Integration of Three-dimensional Models with Cn3D

The dbMHC links to a three-dimensional (3D) display to allow visualization of individual amino acid positions on a number of HLA molecules. Additionally, this tool allows the positions of amino acid differences between selected alleles to be highlighted.

The 3D display can be accessed via the Alignment Viewer or the SBT Interface. After a locus is selected, the button Download/3D opens a pop-up window, allowing a download of the Cn3D file.

To use this feature, the user must install the NCBI's Cn3D viewer application. A link to the installation page for this program for Windows, Macintosh, and UNIX platforms is provided. All HLA gene 3D structures available on the Cn3D viewer have been retrieved from the Molecular Model Database (MMDB). At present, X-ray crystallographic models of HLA-A, B, C, DR, DQ, DM, and MICA molecules are available.

Each model lists the allelic variants on which it is based. The models are used to highlight the amino acid positions; highlighting does not substitute amino acids and will not produce a conformational change in the model.

The CN3D feature of dbMHC has been implemented to also work in close interaction with the dbMHC Alignment Viewer. If the user specifies a set of alleles within a locus, all amino acid differences within this set can be highlighted. Information on each individual substitution can be viewed in the "Style-Annotation" section of Cn3D. It is also possible to highlight a contiguous stretch (e.g., exon 2) in the model.

MHC Microsatellite Database (dbMHCms)

The dbMHCms database contains many of the known microsatellites (Msat) across the HLA region and is designed to search for descriptive information, when available, about these highly variable short tandem repeats (STRs). Users will be able to extract MHC Msat information describing physical location, informativity, known allele length and number, heterozygosity, allelic motifs, and primer sequences from the dbMHCms database.

The dbMHCms web page is intended to provide a web-based resource for cataloging and disseminating information on markers that could be used to provide evidence for genetic linkage and/or in studies of association between regional MHC polymorphism and disease susceptibility. This resource was developed by NCBI in collaboration with A. Foissac, M. Salhi, and A. Cambon-Thomsen,who provided the original reviews on MHC microsatellites in a series of updates (4–6). dbMHCms will be enhanced to include a very detailed set of STR marker data developed by the IHWG. These data will integrate/stan-

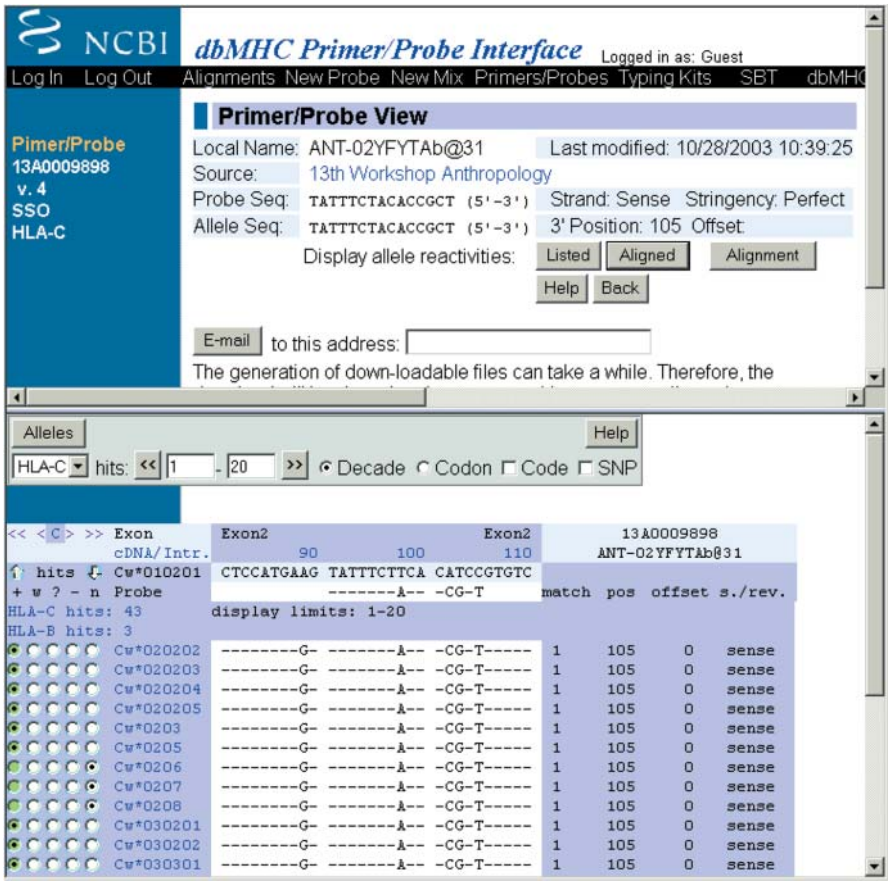

Figure 3. **The Primer/Probe Interface. An SSO probe of the anthropology typing kit of the IHWG for the locus HLA-C is displayed.** The alignment or probes with reacting alleles can be seen in the *lower section* of the browser.

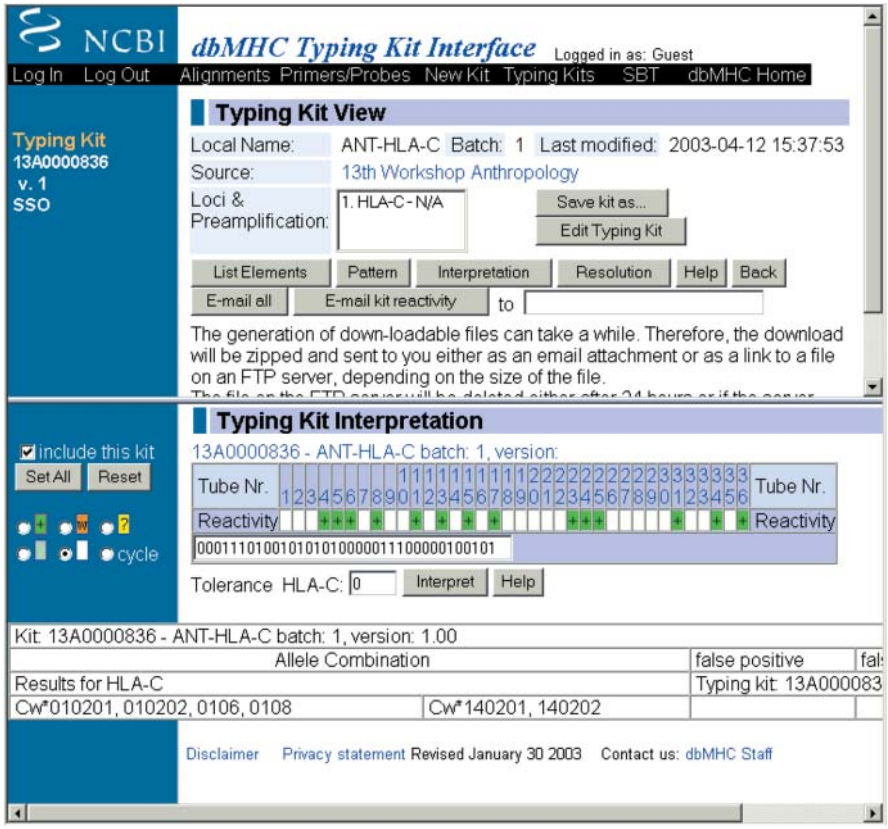

Figure 4. Typing Kit Interface. The HLA-C typing kit of the Anthropology project has been selected. The *lower section* of the screen displays the interpretation interface. The reactivity pattern is displayed as a string of "0" and "1" and visualized with "+". The current pattern can be interpreted as Cw*010201/010202/0106/0108 – CW*140201/140202.

dardize microsatellite data across many IHWG projects through a common microsatellite allele nomenclature derived from the direct sequencing of several of the IHWG Msat reference cell lines as described elsewhere in this volume.

Primer/Probe/Typing Kit Database

The Primer/Probe database can be used as a reference resource for probe design and is intended to provide necessary tools for the exchange of DNA typing data based on the above technologies. This database is composed of a Primer/Probe Interface and a Typing Kit Interface. Each interface has a series of interactive frames that allow the user to access different database functions.

Submissions to and downloads from this database can be performed using the XML data exchange format described in http://www.ncbi.nlm.nih.gov/mhc/xml/schema/.

Primer/Probe Interface

The Primer/Probe database (Fig. 3) is intended to provide a comprehensive and standardized characterization of individual typing primers and probes, as well as primer and/or probe mixes, and encompasses the following technologies:

1. Sequence Specific Oligonucleotide probe hybridization (SSO)
2. DNA amplification using Sequence Specific Primers (SSP)
3. Sequencing Based Typing (SBT) protocols

The Primer/Probe Interface provides information on individual reagents used for the typing of MHC or MHC-related loci. Users can access this information through the multiple functional frames within the interface. The words "Primer/Probe" represent SSOs, SSPs, SSP mixes, SSO mixes, and their combinations, which include nested PCR or SSO hybridization on group-specific amplifications. Nested PCR and group-specific amplification are both based on a primary amplification with an SSP mix.

Reagent Allele Reactivity Prediction

With each update of the allele database, all allele reactivities of primers and probes submitted to dbMHC are recalculated with an allele reactivity prediction tool. The allele reactivity prediction tool for primers and probes available through dbMHC presents calculations based solely on sequence similarity and can therefore only offer suggestions for the possible allele reactivities of individual primers and probes.

Users should not consider this tool to be 100% accurate. A reliable prediction of primer/probe allele reactivities requires complete sequence information, as well as reaction data that include annealing temperature and magnesium concentration. Because these data are not consistently available to dbMHC, we are unable to offer more than direct sequence comparison for prediction of primer/probe allele reactivity.

Typing Kit Interface

The Typing Kit Interface contains information on individual typing kits used for typing MHC or MHC-related loci (Fig. 4). Users can access this information through multiple functional frames within the interface. Typing kits contained within the Typing Kit Interface consist of SSOs, SSO mixes, or SSP mixes. Elements of typing kits may interact with unamplified DNA, pre-amplified DNA, or several distinct groups of pre-amplified DNA within one locus (e.g., two different amplifications of a certain exon using a distinct polymorphism). The elements of each typing kit will react in characteristic patterns with individual alleles. Users can employ these patterns to determine allelic variants or groups of allelic variants within a locus.

IHWG/IHWS Data-Related Resources

Data collected by the IHWG, initially during the 13th IHWS, will be displayed as separate project-specific resources. The aim of these resources is to provide access to individual raw data, as opposed to simple calculated summaries of data. Users will be able to download data and test their hypotheses against these datasets. Periodic updates to the existing data, as well as new data, will be provided by the IHWG though 2005 and the 14th IHWS.

Anthropology/Human Diversity Data Resource (Available December 2003)

The data generated by the IHWG Anthropology/Human Diversity Project are the first to be made available at the dbMHC

***Figure 5.* Anthropology/Human HLA diversity project.** The upper part of the allele frequencies of HLA-A for the regions Europe, North Africa, and North-East Asia is displayed. The number of typed alleles (2n) is displayed for each region. The A02 serogroup is opened for a detailed view. The striking difference between A*0205 and A*0206 allele frequencies in Europe and North-East Asia is illustrated. For each allele the total number of alleles is listed, and the frequency is added in parentheses.

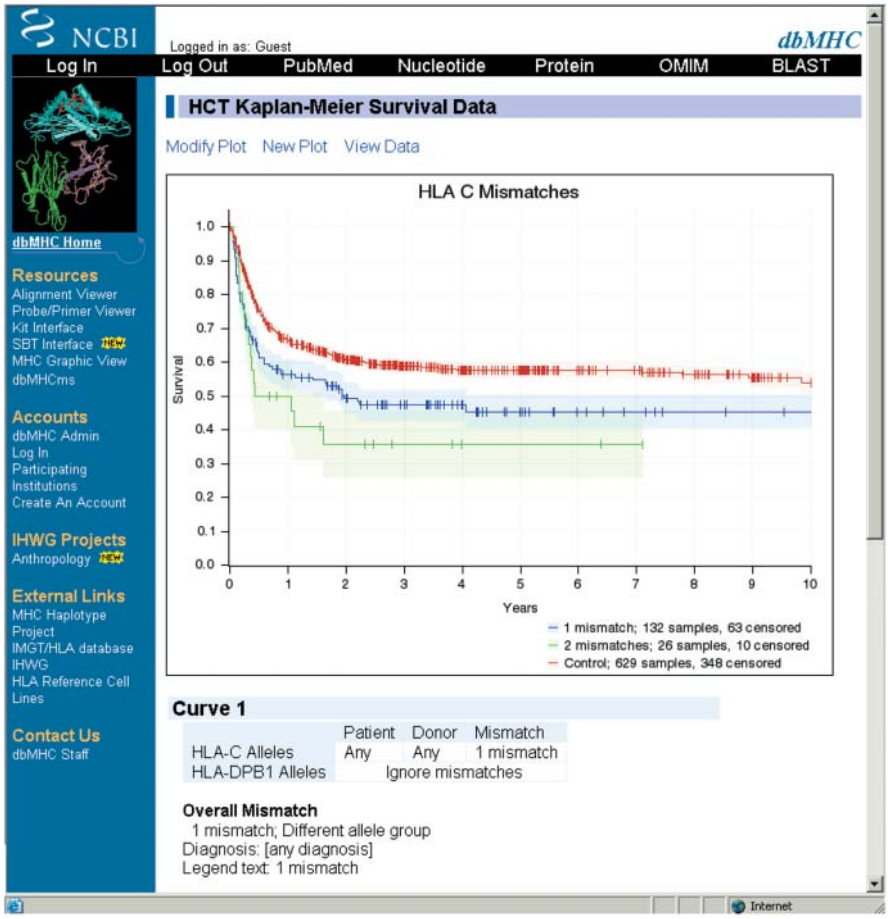

***Figure 6.* Hematopoietic Stem Cell Transplantation project.** Shown is the Kaplan-Meier Survival Curve for recipient-donor pairs with a simple HLA locus mismatch at HLA-C. All cases are serologically matched (matched for the first 2 digits) and matched at the allele level for HLA-A, B, DRB1, and DQB1. Matching for DPB1 was not considered. Censored data points are visualized as tick marks.

web site at NCBI, at http://www.ncbi.nlm.nih.gov/mhc/ihwg.cgi? (Fig. 5).

The anthropology dataset comprises class I (A, B, C) and class II (DRB1, DQA1, DQB1, DPA1, DPB1) genotypes for over 12,000 individuals from 95 populations generated by 32 participating laboratories. In contrast with other allele frequency databases, dbMHC has been implemented to be able to store and provide individual HLA genotypes together with raw typing data, when available. Allele frequencies are based on the dataset submitted and are recalculated each time upon retrieval. At the web site, data can be displayed or downloaded either as individual typing results or as cumulative allele frequencies. Users can select subsets of data by population, global region, HLA locus, individual allele, and/or generating laboratory. Multiple selections are supported.

Allele frequencies can be sorted either by allele name or allele frequencies within a particular population or region. Individual allele frequencies are also summed over each main allele serotype or supertype (equivalent to A*02) (Comment: HLA supertypes do not correspond to HLA serogroups, but it seems like the two concepts are being used interchangeably here. Given my understanding of the way the data is presented on dbMHC, supertype should be replaced with serogroup in this section). This allows comparisons of the overall frequency of a given supertype. A hierarchical view is available for display of frequency data as nodes for each supertype, and each node can be expanded to list individual allele frequencies of this group. Alleles that have not been observed are not listed. All query results can be downloaded as a file, either as a tab-delimited text file or as an XML file. Downloads of individual records contain the submitting center as well as the citation of the publication from which the record is a part.

Hematopoietic Stem Cell Transplantation Data (Available February 2004)

Data of the stem cell transplantation component are currently restricted to unrelated donor-recipient pairs. All pairs are typed at high resolution (allele level) for at least HLA-A, B, C, DRB1, DQB1, DQA1, and some for DPB1.

Several steps have been taken to ensure anonymity of the submitted HSCT data. A single record consists of all data avail-

able for a donor-recipient pair. The order of donor-recipient pairs has been randomized so that no link to a particular transplant center can be made. However, if requested by a transplant center, dbMHC is prepared to store and display this transplant center for each of its transplants.

Data can be grouped by patient age, sex, diagnosis, degree of mismatch, and vector (2-allele, graph rejection, or GvHD) of mismatch. The clinical end point of the current dataset is time of survival. The dataset can be expanded to include additional clinical data.

The resource provides information about the allele distribution of donors and recipients; it also informs whether and how often a donor and recipient match has occurred for any specific genotype within the dataset. Future updates of the HSCT data will include severity of GvHD, relapse, year of transplant, and non-relapse mortality.

Users can generate and display online Kaplan-Meier estimates for survival for user-defined subsets of recipients (Fig. 6). All data can be downloaded, either as summarized data or as individual donor-recipient pairs, grouped according to the selection of the user.

Integration with Other Resources

NCBI Links

The dbMHC provides a set of internal links to other NCBI resources through either the black "quick link" bar located at the top of the dbMHC homepage or through the rotating HLA molecule located at the top left-hand corner of the dbMHC homepage. These links will take the user to the following NCBI resources:

PubMed, Nucleotides, Protein, Online Mendelian Inheritance in Man (OMIM), Blast, and Molecular Modeling database (MMdb).

The dbMHC Alignment Viewer provides locus-level linkage to NCBI's LocusLink and to dbSNP at the individual SNP level. The Graphic View page provides a hyperlink-enabled schematic representation of the MHC region that can be used as an alternate display of the dbMHC homepage. Locus-level links from the Graphic View page include:

dbSNP, Map View, LocusLink, Online Mendelian Inheritance in Man (OMIM), Nucleotide, Protein, PubMed, Structure,

Following the initial printing, this document, *The 13th IHWS Proceedings* will also be available through the NCBI Books resource providing online access of subjects, chapters, tables, and figures herein.

References

1. Kitts A, Feolo M, Helmberg W. The Major Histocompatibility Complex database, dbMHC. In: McEntyre J, editor. The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct [cited 2003 June]. Available from: http://www.ncbi.nlm.nih.gov/entrez.fcgi?db=Books.
2. Robinson J, Bodmer JG, Malik A, Marsh SGE. Development of the International Immunogenetics HLA Database. *Hum Immunol* **59** (Suppl 1), 17, 1998.
3. Robinson J, Marsh SGE, Bodmer JG. The IMGT/HLA Database. *Eur J Immunogenet* **26**, 75, 1999.
4. Foissac A, Salhi M, Cambon-Thomsen A. Microsatellites in the HLA region: 1999 update. *Tissue Antigens* **55** (6), 477–509, 2000.
5. Foissac A, Cambon-Thomsen A. Microsatellites in the HLA region: 1998 update. *Tissue Antigens* **52** (4), 318–352, 1998.
6. Foissac A, Crouau-Roy B, Faure S, Thomsen M, Cambon-Thomsen A. Microsatellites in the HLA region: an overview. *Tissue Antigens* **49** (3 Pt 1), 197–214, 1997.
7. E Mickelson, A Damodaran, P He, M Malkki, A Smith, E Petersdorf, et al. Standardization of Microsatellite Data: Comparative Analysis of Results from 13th IHWS Participating Laboratories and Establishment of a Microsatellite Reference Panel and Standard Nomenclature. This volume.