Review Article

HLA Allele Type Prediction: A Review on Concepts, Methods and Algorithms

Balamurugan Sivaprakasam*, Prasanna Sadagopan

Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, INDIA.

Submission Date: 07-06-2023; Revision Date: 09-08-2023; Accepted Date: 14-09-2023.

ABSTRACT

The Human Leukocyte Antigen (HLA) gene system situated on Chromosome 6 has been the subject of extensive research, primarily due to its vital role in transplantation and its links to autoimmune, infectious, and inflammatory diseases. The classical HLA genes, including HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1, exhibit a high degree of polymorphism among individuals within a population. As many changes in the allele, computational imputation-based HLA typing is used extensively and in machine learning, it is possible through supervised learning. There are many methods available for doing HLA imputation from HLA and SNP genotype data using different methods and algorithms. The present study carefully examined the research articles and noticed that the Ensemble methods, Random Forest and Boosting algorithms are the few effective methods for HLA imputation. Attribute bagging is a technique that enhances the accuracy and stability of classifier ensembles by employing bootstrap aggregating and random variable selection. The ensemble classifier method involves two main phases. In the first phase, a collection of base-level classifiers is generated, and in the second phase, a meta-level classifier is trained to combine the outputs of the base-level classifiers. The R statistical programming language is utilized by Bioconductor software packages such as HIBAG, which are designed for the research community to impute (assign) HLA types using SNP data. In the present study, the details of different methods, software and algorithms used for HLA imputation are discussed for the non-biologists and biologists who work on HLA allele type prediction.

Keywords: HLA prediction, SNP genotype, Imputation, HIBAG, Bioconductor.

Correspondence:
Mr. Balamurugan Sivaprakasam,
Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, INDIA.

Email: sivabala76@gmail.com

INTRODUCTION

Research on HLA involves a multidisciplinary approach and can be conducted by various researchers from different fields like Immunologists, Geneticists, Transplantation researchers, Pharmacogeneticists, Bioinformaticians and Medical clinicians. Along with these, computer researchers, particularly those specializing in computational biology can contribute their expertise in HLA data analysis, allele typing algorithms development, database management, data mining, population genetics, evolutionary analysis and integration of HLA data to enhance our understanding of HLA genetics. Computer science researchers' works with HLA often face challenges in searching for articles due to their multidisciplinary nature. One of the main difficulties that arises is the specialized terminology used in HLA research including HLA allele nomenclature, immune system and terms related to genetic concepts. Therefore, the concepts relevant to HLA may be dispersed across various articles. On this aspect, in this review article, the basic concepts of HLA, methods and software algorithms for the prediction of HLA allele typing are discussed to understand easily by non-biologists.

SCAN QR CODE TO VIEW ONLINE

www.ajbls.com

DOI: 10.5530/ajbls.2023.12.29

The Human Leukocyte Antigen (HLA) system encompasses a cluster of genes that encode proteins crucial for antigen presentation to the immune system. These HLA genes are situated within the Major Histocompatibility Complex (MHC) region on chromosome 6 in humans. These genes encode the HLA proteins and they are involved in immune responses. The main classes of HLA genes are HLA class I (HLA-A, HLA-B, HLA-C), HLA class II (HLA-DP, HLA-DQ, HLA-DR) and class III. Within the HLA system, each class has multiple genes and within each gene there are numerous genetic variants known as alleles.[1] The HLA gene system with different classes is shown in Figure 1.

HLA alleles are alternative forms or variations of a specific HLA gene. They represent the different genetic sequences or variations found within a particular HLA gene. For example, the HLA-A gene has many different alleles such as HLA-A*01:01, HLA-A*01:02, HLA-A*02:01, HLA-A*02:05, HLA-A*03:01 and so on. Similarly, other HLA genes have their own sets of alleles. Alleles differ in their nucleotide sequences, resulting in differences in the proteins they encode. The few alleles of HLA-A and their corresponding protein product are shown in Figure 2. Therefore, generally all the HLA genes exhibit a high degree of polymorphism, making them one of the most diverse gene families in the human genome. Polymorphism refers to the existence of multiple alternative forms or variants of a gene within a population. The evolutionary forces acting on these loci have resulted in a significant amount of functional diversity. Due to the complexity and diversity of HLA polymorphism, comprehensive HLA typing methods are employed to accurately determine an individual's HLA genotype. The polymorphism of HLA genes has important implications in various areas of medicine and biology. In transplantation, matching the HLA types between donors and recipients is critical to minimize the risk of graft rejection. Furthermore, the presence of HLA polymorphism is linked to the susceptibility or resistance to specific diseases, as well as the diversity of immune responses and individual differences in drug responses.[2-5]

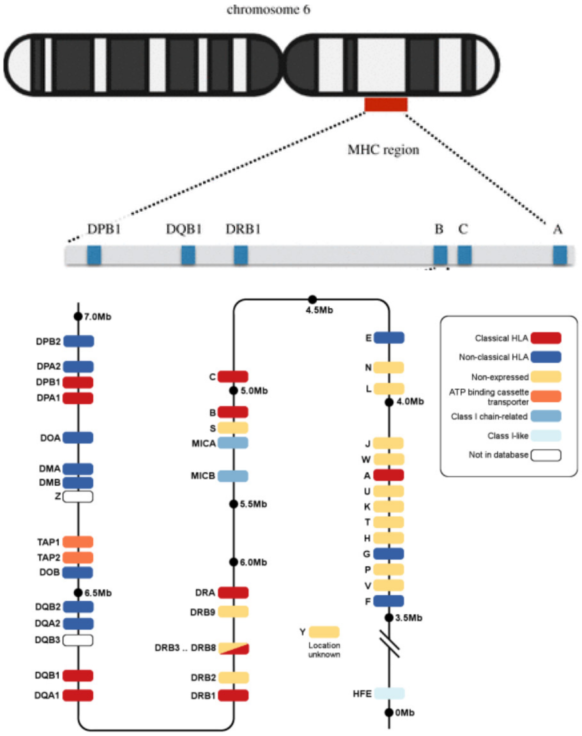

Figure 1: The genomic locations of HLA genes (loci) are depicted within the human leukocyte antigen (HLA) gene system on chromosome 6 including the classical class I and II. The position of the well-studied HLA-A, -C, -B, -DR, -DQ, -DP genes can be noticed.
(Courtesy: HLA Nomenclature @ hla.alleles.org).

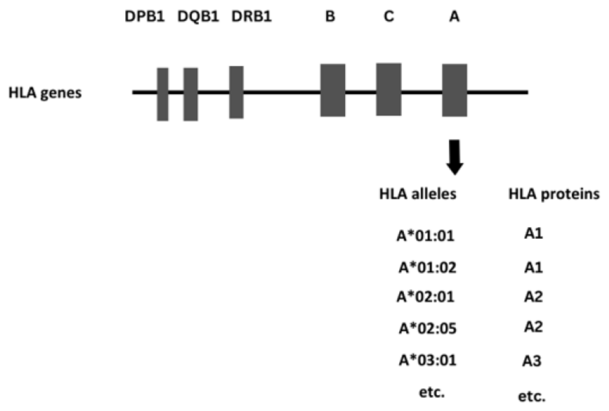

Figure 2: The alleles and corresponding proteins are shown for the HLA-A gene.

Linkage disequilibrium and Hardy-Weinberg Equilibrium in HLA typing

In HLA typing, both Linkage Disequilibrium (LD) and Hardy-Weinberg Equilibrium (HWE) play important roles in understanding the genetic characteristics of HLA alleles. LD represents the non-random association between alleles at distinct loci on the identical chromosome. In the case of HLA typing, LD is often observed between different HLA loci because of their proximity to the same chromosome and the limited recombination events that occur in these regions. LD can influence the inheritance patterns of HLA alleles, as certain alleles at different loci may be more likely to be inherited together. Understanding LD patterns is important for accurate HLA typing and haplotype inference.[6] On the other hand, Hardy-Weinberg Equilibrium describes the expected allele and genotype frequencies in an idealized population under specific assumptions, including random mating, no selection, no mutation, no migration, and large population size. In HLA typing, HWE is used as a reference for comparing observed allele frequencies with expected frequencies to assess departures from equilibrium. Deviations from HWE in HLA typing may indicate several factors, such as population substructure, selection pressures, genotyping errors, or genetic drift. Departures from HWE can be particularly informative in HLA studies, as they may indicate the presence of specific HLA alleles or haplotypes associated with diseases or traits. Both LD and HWE are essential concepts in HLA typing and population genetics, providing insights into the genetic characteristics, evolutionary dynamics, and disease associations related to HLA alleles. Understanding LD patterns helps in haplotype inference and fine mapping of HLA loci while assessing deviations from HWE allows for investigating genetic and evolutionary factors impacting HLA allele frequencies within populations.[7]

Genetic marker refers to genetic variants or markers that are physically located close to each other on a chromosome. In the context of genetic studies, adjacent markers are typically assessed to study Linkage Disequilibrium (LD), which refers to the non-random association of genetic variants within a population. Genetic markers can include Single Nucleotide Polymorphisms (SNPs) and these markers are scattered throughout the genome and can be spaced at different intervals. The concept of adjacent genetic markers is particularly relevant when studying haplotypes, which are combinations of genetic variants, inherited together on a chromosome. The LD between adjacent markers helps researchers to infer the haplotypes present in a population. By examining the LD patterns between adjacent markers, researchers can gain insights into the genetic structure, recombination rates, and historical events that have shaped the genome. LD analysis of adjacent markers is often used in population genetics, Genome-Wide Association Studies (GWAS) etc. to understand the inheritance of traits, identify disease-associated variants, or detect genomic regions under selection.[8]

SNP and Haplotype in HLA genes

A Single Nucleotide Polymorphism (SNP) denotes a variation that arises at a single nucleotide position in the DNA sequence among individuals within a population. SNPs are the most prevalent form of genetic variation and can manifest throughout the genome, including within the genes of the HLA system. Haplotype, in the context of HLA genes, refers to a set of closely linked genetic markers, including SNPs, located on the same chromosome. These markers exhibit a tendency to be inherited together as a block, primarily due to their close physical proximity. Haplotype analysis involves examining the combination of alleles or genetic markers on a chromosome to understand the patterns of inheritance and genetic variation. SNP variations and haplotypes in HLA genes are valuable for various applications in medical and population genetics. SNP tags in the HLA system are specific SNPs that serve as representative markers for a larger set of genetic variations in the HLA genes. They are used to capture genetic diversity and haplotype structure in a more manageable way for research purposes.[9] For visual representation, the SNP, haplotype and SNP tag in the HLA gene system is shown in Figure 3.

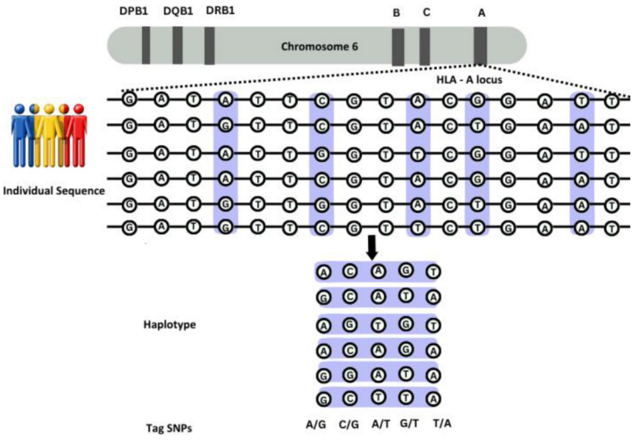

Figure 3: The organization of SNP, haplotype and SNP tag in HLA-A gene system.

Unphased and phased SNP data

There are unphased SNP data and phased SNP data are available in the databases. Unphased SNP data refers to genetic data where the phase or haplotype information of SNPs is not known or explicitly determined. Phasing SNP data refers to the process of determining the arrangement of alleles on each chromosome, specifically which alleles are inherited together on the same chromosome. The phase information is essential for understanding the haplotype structure and how specific alleles are inherited together as a unit. There are several reasons why SNP data may be unphased. It could be due to limitations in genotyping technologies, computational challenges, or lack of available parental or trio data to infer the phase accurately. Unphased SNP data can still be useful in many analyses, such as population genetics and GWAS, utilising statistical algorithms, reference panels, or haplotype imputation techniques to estimate the underlying haplotypes.[10,11]

Allelic resolution and naming of HLA alleles

The World Health Organization (WHO) nomenclature committee for factors of the HLA system is taking responsibility for standardizing and maintaining the continuously expanding list and nomenclature of HLA alleles.[8] This system provides a unique and consistent way of designating HLA alleles. The HLA allele names consist of multiple components that convey specific information about the allele. HLA typing can be performed at different levels of resolution.[12] Low-resolution typing: This provides a broad classification of HLA alleles and identifies the serological specificity or group of alleles. It typically involves identifying common alleles or groups based on limited sequence information. In the example HLA-A*02, the typing identifies the occurrence of the HLA-A*02 allele group. It provides a broad classification of the HLA-A allele but does not specify the exact subtype or sequence variations within the allele group. Intermediate-resolution typing: This provides a greater level of detail than low-resolution typing. It involves identifying specific allele groups or subtypes within the broader serological specificity. It analyzes a subset of variable positions within the HLA genes to determine more specific sequence variations. For example HLA-A*02:01, the typing identifies the specific subtype within the HLA-A02 allele group as HLA-A02:01. It indicates a more specific sequence variation within the HLA-A*02 group, but it may not provide complete information about all the possible sequence variations within the allele. In some contexts, Intermediate-resolution typing is also mentioned as low-resolution typing. High-resolution typing: High-resolution typing represents the most detailed level of HLA typing. It involves sequencing specific regions of the HLA genes to determine the precise sequence variations within alleles. In the example HLA-A*02:01:01 or HLA-A02:01:01:01, the typing specifies the exact subtype and provides more precise and detailed information on the HLA-A allele. It enables researchers and clinicians to differentiate between closely related alleles with subtle sequence variations, providing a more detailed understanding of HLA diversity and its implications in various applications such as transplantation, disease associations, and population studies.[13]

Reporting Format for HLA Assignments

HLA typing assignment refers to the procedure of identifying an individual's HLA genotype, and it is crucial for the end user to have a clear understanding of it. It is important to recognize that HLA typing assignments can be conducted at various levels of resolution, ranging from low resolution. Furthermore, the typing assignments must adhere to the WHO nomenclature for HLA system factors, which can be accessed at http://www.hla.alleles.org. When reporting HLA assignments, it is important to follow standard guidelines to ensure clear and consistent communication of the results. Initially begin by specifying the HLA gene for which the assignments are being reported, such as HLA-A, HLA-B, HLA-DRB1, etc. Next list the specific alleles that have been identified for the given HLA gene with asterisk (*) and a number representing the allele. Next indicate the level of typing that has been performed, which refers to the resolution or depth of analysis. This could include low-resolution typing, high-resolution typing, or allele-level typing. If there are any ambiguities or unresolved assignments, it is important to report them. Ambiguities refer to situations where the identified alleles cannot be definitively assigned due to overlapping patterns or inconclusive results. Finally, provide any relevant additional comments that may be important for the interpretation of the results. This could include the presence of rare or novel alleles, discrepancies, or any other notable observations.[14,15] Different types of resolutions are depicted in Figures 4 and 5.

Methods of HLA typing

Classical HLA genotyping methodologies, specifically Sequence-Based Typing (SBT), are widely recognized as the benchmark for HLA typing. SBT involves sequencing the exons of HLA genes to accurately identify specific alleles, providing high-resolution and comprehensive HLA allele information. However SBT is time-consuming and expensive as its accuracy and precision make it essential for precise HLA matching in transplantation. Alternative genotyping methods make predictions and offer cost-effective and efficient options but may provide lower resolution. Therefore, researchers and clinicians should consider the specific requirements of their studies or applications when selecting the appropriate HLA genotyping approach.[14]

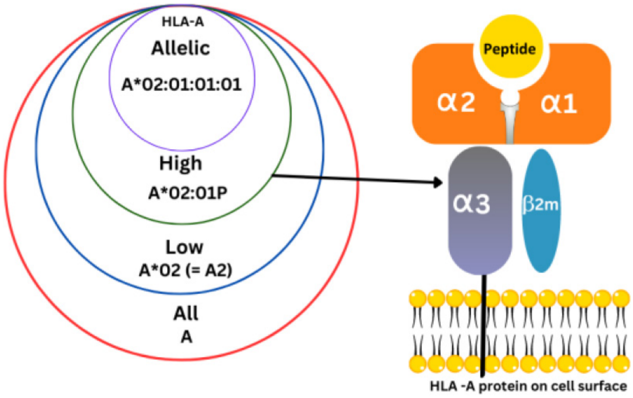

Figure 4: The general HLA typing resolution arrangement and the responsible HLA-A antigen binding protein.

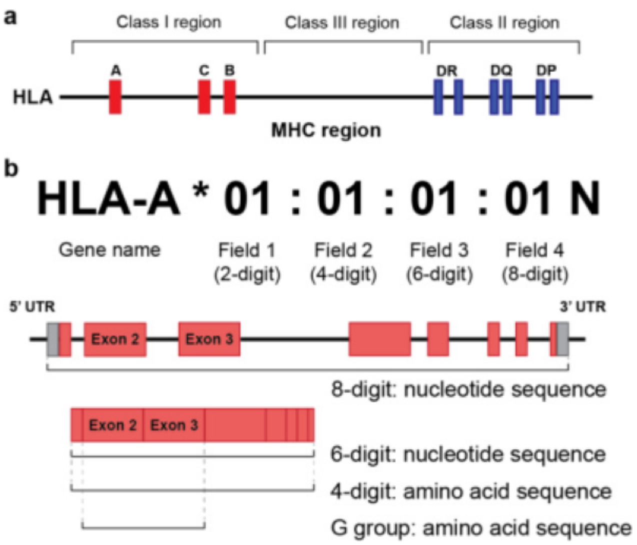

Figure 5: HLA typing resolution. (Courtesy: www.hla.alleles.org).

Reliability of HLA typing

Confirmatory HLA typing is a laboratory technique used to determine the specific genetic makeup of HLA genes in an individual and this high-resolution typing helps in organ transplantation by ensuring better matching between donor and recipient. The term "confirmatory typing" has become less clear because typing practices and matching criteria have changed. To ensure better clarity, it is recommended to replace the term "confirmatory typing" with "Verification typing" and "extended typing." Verification typing focuses on validating initial results, while extended typing offers a more comprehensive characterization of an individual's HLA profile. All three methods contribute to improving the accuracy and reliability of HLA typing, ultimately supporting successful transplantation outcomes.[15]

Databases on HLA

The HLA allele frequencies for the disease associations and population data are available in different resources. NCBI ClinVar aggregates information about relationships between genetic variation and human health.[16] HapMap data, also known as a Haplotype Map, is a valuable tool for researchers seeking to identify genes and genetic variations that impact health and disease.[17] It can be accessed at http://www.hapmap.org.[17] The HapMap and 1000 genome project provides many HLA genotype data in different formats for the researchers for analyses. The Allele Frequency Net Database (allelefrequencies.net), bethematchclinical.org, http://hla.alleles.org, Wellcome Trust Case Control Consortium (WTCCC) and http://ashi-hla.org resources also have HLA data.[18,19]

The interpretation of HLA typing results often involves referencing and comparing the identified alleles with known information stored in specialized databases. The commonly used databases for HLA typing interpretation are the IMGT/HLA Database and the Allele Registry. The IMGT/HLA Database, maintained by the International ImMunoGeneTics Information System® (IMGT®), is a comprehensive and authoritative resource for HLA sequences, alleles, and nomenclature. It provides a repository of validated and curated HLA allele sequences, along with associated information such as exon boundaries, polymorphisms, and population frequencies.[20] To date, over 32,000 alleles have been identified for the class-I and class-II genes [IPD-IMGT/HLA Database (ebi.ac.uk)]. The IMGT/HLA Database is extensively utilized by laboratories and researchers to interpret HLA typing outcomes. Additionally, the Immunogenetics and Transplantation Laboratory at the Anthony Nolan Research Institute operates the Allele Registry, which serves as another valuable resource for interpreting HLA typing results. It contains a curated collection of HLA allele sequences, including updates on new alleles, allele nomenclature, and other relevant information. The Allele Registry functions as a resource for HLA typing reference and supports the identification and characterization of novel or rare alleles. These databases serve as vital references for accurate and reliable interpretation of HLA typing results, aiding in the determination of allele specificity, population frequencies, and compatibility assessments for transplantation procedures. It is important to note that these databases are regularly updated as new alleles are discovered and characterized. Therefore, it is important for laboratories and researchers to ensure they are using the most up-to-date versions of these databases for accurate interpretation of HLA typing results.

Matching for tissue transplantation

To facilitate allogeneic transplantation, it is essential for every laboratory to supply a detailed account of the matching status of HLA assignments between a potential donor and a patient, whether they are related or unrelated individuals.[21] HLA haplotypes that are Identical by Descent (IBD) indicate haplotypes inherited from a common ancestor without any recombination events. These haplotypes are shared by individuals with a direct blood relationship, such as siblings or parents and children. Another type of matching involves related donors who show compatibility with the patient's tested HLA loci based on segregation within the family. HLA identity for all loci tested means that two individuals have an exact match of HLA alleles across multiple HLA loci. In some families, it may not be possible to confirm Identity by Descent (IBD) through segregation analysis due to factors such as unavailable family members, complex family structures, multiple ancestral contributions, or recombination events. Alternative methods to estimate the probability of IBD based on HLA allele frequencies are population-based statistical analyses, matching algorithms and comparisons with reference databases. The overall goal is to minimize graft rejection risk and improve transplantation outcomes. With the availability of IBD information, the algorithms and software tools are developed for HLA allele type imputation.[22]

Software used in HLA Prediction

Leslie *et al.* (2008) introduced a new statistical approach for HLA allele prediction. They utilized a database of SNP haplotypes containing known HLA alleles and employed an Identity-by-Descent (IBD) model based on approximate coalescent models to develop their LDMhc algorithm. For SNP selection, they implemented a leave-one-out cross-validation scheme.[22] Approximate coalescent models are computational methods used in population genetics to study the genealogical history and genetic diversity of populations. They provide a simplified framework for simulating and analyzing genetic data by modelling the process of genetic coalescence. These models make calculations more manageable by incorporating assumptions and approximations, such as constant population size or neglecting recombination. By using approximate coalescent models, researchers can gain insights into population history, genetic structure, and evolutionary processes. However, it's important to validate these models against empirical data and consider their limitations. The LDMhc algorithm, also known as the "Linkage Disequilibrium-based Multi-locus Haplotype Construction" algorithm, is a computational method used for haplotype inference from genotype data. It is specifically designed for multi-locus genetic data where Linkage Disequilibrium (LD) patterns exist between adjacent genetic markers. The LDMhc algorithm leverages LD patterns to infer haplotypes, which are combinations of genetic variants inherited together on a chromosome. Haplotypes are important in genetic studies as they provide insights into the inheritance patterns of genetic variants and can help identify disease associations or population differences. The LDMhc algorithm utilizes LD measures, such as D' and r-squared, to quantify the extent of association between adjacent genetic markers. It then employs an iterative algorithm that assigns haplotypes to individuals based on the observed LD patterns and the compatibility of haplotypes across individuals. The algorithm aims to find a set of haplotypes that best explains the observed genotype data while satisfying LD constraints. By inferring haplotypes from genotype data using the LDMhc algorithm, researchers can gain a better understanding of the underlying genetic variation and haplotype structure within a population. This information is useful in various genetic analyses, including disease association studies, population genetics, and understanding the evolutionary history of populations. It's important to note that the LDMhc algorithm is just one of many algorithms available for haplotype inference, and its performance and suitability may vary depending on the specific dataset and research question.

In a subsequent study, Dilthey *et al.* (2010) developed the integrated software HLA*IMP, which incorporates a two-step approach. In the first step, haplotype frequencies are estimated from the SNP genotypes using a statistical model called LDMhc.[23] In the second step, HLA*IMP (Web service now discontinued) utilizes attribute bagging and expectation propagation to impute the HLA alleles based on the haplotype frequencies obtained in the first step. It has limitations such as bias due to the reference panel used, the inability to detect novel rare alleles and the possibility of imputation errors leading to incorrect predictions.[23]

BEAGLE is an alternative method for HLA imputation, specifically designed to infer HLA genotypes and haplotypes from unphased SNP data. It utilizes a statistical algorithm that leverages patterns of Linkage Disequilibrium (LD) to impute missing genotypes and infer the underlying haplotypes. It employs a Markov Chain Monte Carlo approach to estimate the most likely haplotype configurations based on the observed SNP data and LD information. It utilizes reference panels or databases of known HLA haplotypes to improve accuracy and impute missing or unobserved genotypes. It has certain limitations and it relies on the availability and quality of reference panels, and its performance may vary depending on the specific population dataset.[24]

SNP2HLA performs the imputation of amino acids, HLA alleles, and SNPs in the MHC region based on SNP genotype data.[25] Though it is valuable in studying the association between HLA alleles and diseases, it has certain limitations including reference panel bias, ambiguity in imputed alleles and population-specific effects. HLA-Check is an additional tool that evaluates HLA data using SNP information, and its license explicitly permits unrestricted use and modification to suit individual requirements.[26]

HIBAG utilizes SNP genotype data to impute HLA classical alleles, relying on a training set of HLA genotype data. Rather than necessitating access to extensive training sample datasets, it offers published parameter estimates to the research community. HIBAG synergistically integrates attribute bagging, an ensemble classifier method, with haplotype inference for SNPs and HLA types. The incorporation of attribute bagging techniques, such as bootstrap aggregating and random variable selection, effectively improves the accuracy and stability of classifier ensembles.[27] HLA-IMPUTER (Web service now discontinued) is a software tool used for imputing or predicting classical HLA alleles from genotype data. It is designed to fill in missing or unobserved HLA allele information based on genetic markers, typically SNPs.[28]

Deep*HLA is a deep learning-based method used for predicting HLA alleles from Next-Generation Sequencing (NGS) data. It employs deep neural networks to analyze the sequence reads and predicts the corresponding HLA alleles. Deep*HLA has shown promising results in terms of accuracy and speed in HLA typing from NGS data. It is specifically designed to handle the complexities and challenges associated with high-throughput sequencing technologies.[29] CookHLA is a software tool used for HLA allele calling and genotyping from NGS data. It utilizes statistical models and computational algorithms to process the sequencing reads and determine the most likely HLA alleles present in the sample. CookHLA incorporates various steps, such as read alignment, variant calling, and HLA allele inference, to provide accurate and reliable HLA genotyping results. It is commonly used in research studies and clinical settings to analyze HLA data obtained from NGS platforms.[30]

From the review, the present study noticed that in developing an application for predicting HLA allele type, the R statistical programming language, the HIBAG package and their supplementary software tools are important and readily available for the research community.[31] It is supported through the recent comparison study that HIBAG is still one of the best imputation methods.[32] The HIBAG software is freely accessible and publicly available as an R/Bioconductor package, which can be obtained from the following link: http://www.bioconductor.org/packages/HIBAG. It is compatible with Windows operating systems, allowing for easy installation and execution and it can also be installed in other operating systems like Linux and Mac OS. Packages found under Bioconductor software packages are used for imputing HLA types using SNP data.

HLA Genotype Imputation with Attribute Bagging (HIBAG)

HIBAG is a software tool specifically developed for conducting HLA typing based on genotype data. It employs imputation methods to deduce HLA types from SNP genotype data, offering a cost-effective solution for HLA typing. By incorporating reference databases of HLA alleles, such as the IMGT/HLA database, HIBAG leverages the linkage disequilibrium patterns between HLA alleles and nearby SNPs to impute missing HLA genotypes. It is compatible with various genotyping platforms (Affymetrix, Illumina etc.) and can handle different formats of genotype data. HIBAG is a method available as a regularly updating Bioconductor package developed for HLA typing using SNP genotype data It also supports the typing of multiple HLA loci, including HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, and HLA-DRB1. It enables simultaneous typing of these loci by incorporating haplotype information and linkage disequilibrium patterns between them. HIBAG incorporates reference databases of HLA alleles, such as the International ImMunoGeneTics (IMGT)/HLA database, to perform the imputation. These databases contain extensive information about known HLA alleles, their haplotypes, and associated SNPs. HIBAG offers the flexibility to create population-specific models for imputation. It considers the allele frequency distribution and linkage disequilibrium patterns specific to the target population, which can improve the accuracy of HLA typing predictions. HIBAG incorporates quality control measures to assess the reliability of imputed HLA types and provides confidence estimates or imputation scores to indicate the accuracy of the predictions. HIBAG provides confidence estimates or imputation scores to indicate the reliability of the predicted HLA genotypes, allowing users to evaluate the accuracy of the results. Importantly, it is available as open-source software, allowing researchers to access and modify the code to suit their specific needs. The open-source nature promotes collaboration and enables the incorporation of new advancements and improvements.[31] The unique aspect of HIBAG is that it doesn't require access to huge training sample datasets; rather, researchers can utilise the published parameter estimates which are available on the website http://www.biostat.washington.edu/bsweir/HIBAG/.

Algorithm behind HIBAG

In machine learning, ensemble techniques involve combining multiple models to make predictions rather than relying on a single model. One such ensemble technique is bagging, which is the underlying principle behind the Random Forest algorithm. Bootstrap Aggregating (Bagging) is a technique that involves generating unique subsets of the training data by randomly selecting samples with replacements. Each subset is used to train an individual base model, with decision trees being the base models in the case of Random Forest. During the training of a Random Forest, each decision tree is trained on a distinct subset of the training data, generated through random sampling with replacement. Consequently, certain instances may occur multiple times within a subset, while others may be omitted. Moreover, during each split in the decision tree, only a random subset of features is considered, adding further randomness and diversity to the ensemble. When making predictions, the Random Forest combines the predictions from all the decision trees using a majority vote. In classification tasks, the final prediction is determined by selecting the class with the highest number of votes among the decision trees. In regression tasks, the average prediction from the decision trees is taken. By employing the bagging technique and aggregating the predictions of multiple decision trees, Random Forest can achieve higher accuracy, robustness against overfitting, and improved generalization compared to individual decision trees. The diversity introduced by random sampling and feature selection helps the ensemble to capture different aspects of the data, leading to more reliable predictions.[27,31]

Work Flow for the supplication development

HIBAG is one of the packages in Bioconductor available for the research community, in which SNP data is employed for inferring HLA types. Researchers who have published population-based models can use HIBAG rather than needing access to huge training sample datasets. Also, the models can be created using the same Bioconductor packages. The database like the 1000 Genomes Project website makes the HLA genotypes of study participants available to the public.[30] The classifier ensembles can increase accuracy and stability through the attribute bagging technique. This approach uses its own 30% randomised data together with 70% of the training datasets of SNP and HLA genotype data. The 30% of data that was omitted is used as a validation set. Additionally, it lowers variance and aids in preventing overfitting. The classifier models are trained with several classifiers using the Bagging method. HIBAG functions contain the training algorithm and along with the user-defined and R functions can develop an ensemble model, generally, 100 separate classifiers can be generated by default, Nevertheless, this procedure is time-consuming and typically requires approximately four days to complete on a single core.[31] In order to reduce the time, the available population-wise pre-fit classifiers can be used as models for the HLA type prediction. To demonstrate the method of application development, the standard HLA and SNPs datasets of 1000 genomes project are used from the ImmPute project.[32] The pre-fit classifiers of broad and population-specific data are publicly available for the research community in the databases and can be used as models. This application consists of two main components. The first component involves downloading a pre-fitted classifier from the HIBAG website, applying it to an SNP dataset, and evaluating its foretelling performance. In the second component, the accuracy of the imputation can be assessed. The accuracy of the imputation is influenced by several factors, such as the evolutionary closeness of the sample to the training set, the sample size of the training data, the frequency of the HLA allele, and the density of SNPs in the HLA region. The workflow, depicted in Figure 6, includes prediction stages that utilize the input sample, model data, as well as the Bioconductor HIBAG and R functions.

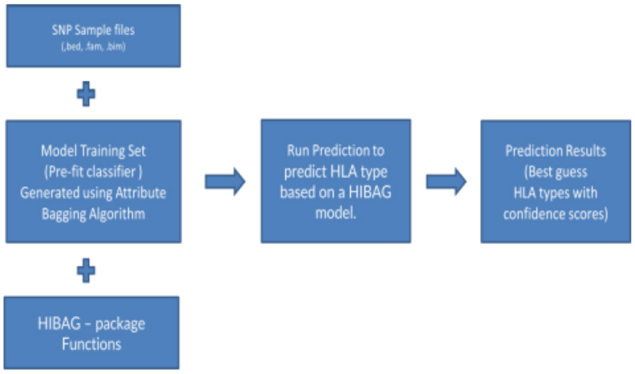
SNP Sample files (.bed, .fam, .bim)
Model Training Set (Pre-fit classifier) Generated using Attribute Bagging Algorithm
HIBAG - package Functions
Run Prediction to predict HLA type based on a HIBAG model.
Prediction Results (Best guess HLA types with confidence scores)

Figure 6: Workflow of HLA prediction using R programming and Bioconductor package HIBAG.

CONCLUSION

In statistics, imputation is the method of changing missing data with substituted values. There are many methods and software available for HLA imputation. When choosing a method and software for HLA imputation, several factors should be considered. First, define the specific goals of the study, whether it is association studies, population genetics, or clinical applications, because, different methods and software may have specific strengths for different research objectives. Second, one should consider the type and amount of genetic data one has as HLA imputation can be performed using genotyping arrays, whole-genome sequencing, or targeted sequencing data. Third, the choice of reference panel is important for imputation accuracy. Fourth, one should assess the imputation accuracy and resolution of different methods by referring to benchmarking evaluations and studies. Hence, it is essential for researchers to conduct an up-to-date and comprehensive evaluation of methods, software, and algorithms in order to select a dependable approach for analyzing HLA typing. In this study, our main focus was to systematically compare the available HLA typing algorithms and published software tools. When computer science researchers interested in HLA research can collaborate with experts from genetics, immunology, or bioinformatics fields, they can provide valuable guidance to help in understanding the specialized terminology. These approaches will help researchers broaden their understanding of the field and enable them to contribute to HLA research from a computer science perspective. We believe that this work can support computer and biological researchers in choosing the best HLA typing method for their data. This, in turn, may help to develop more powerful tools in the future. The study noticed that the Ensemble methods, Random Forest and Boosting algorithms are the few among the effective methods for HLA imputation. Also, the present study may continue to develop an application for the prediction of HLA typing using the above-mentioned Bioconductor package HIBAG and its related algorithms. Further, the results will be published somewhere else.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

HLA: Human Leukocyte Antigen; **MHC:** Major histocompatibility complex; **LD:** Linkage disequilibrium; **HWE:** Hardy-Weinberg Equilibrium; **GWAS:** Genome-wide association studies; **SNP:** Single Nucleotide Polymorphism; **WHO:** World Health Organization; **SBT:** Sequence-based typing; **WTCCC:** Wellcome Trust Case Control Consortium; **HIBAG:** HLA Genotype Imputation with Attribute Bagging; **NGS:** Next-generation sequencing; **IMGT:** ImMunoGeneTics; **Bagging:** Bootstrap Aggregating.

SUMMARY

The Human Leukocyte Antigen (HLA) gene system is involved in transplantation and association with autoimmune, infectious and inflammatory diseases. They have highly polymorphic alleles among the individuals in a population. As with many changes in the allele, computational imputation-based HLA typing is possible which is more cost-effective and time-consuming than the regular sequencing method. There are many methods available for doing HLA imputation from HLA and SNP genotype data. The present study noticed from the literature that the Ensemble methods, Random Forest and Boosting algorithms are the few among the effective methods for HLA imputation. The Bioconductor package HIBAG is freely available for the research community for imputing (assigning) HLA types using SNP data and it uses the R statistical programming language. On this aspect, in this review article, the basic concepts of HLA, methods and software algorithms for the prediction of HLA allele typing are discussed to understand easily by non-biologists.

REFERENCES

1. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010;125(2);Suppl 2:S3-S23. doi: 10.1016/j.jaci.2009.12.980, PMID 20176265.
2. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018;18(5):325-39. doi: 10.1038/nri.2017.143, PMID 29292391.
3. Liu B, Shao Y, Fu R. Current research status of HLA in immune□related diseases. Immun Inflam Dis. 2021;9(2):340-50. doi: 10.1002/iid3.416, PMID 33657268.
4. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14(1):301-23. doi: 10.1146/annurev-genom-091212-153455, PMID 23875801.
5. Crux NB, Elahi S. Human leukocyte antigen (HLA) and immune regulation: how do classical and non-classical HLA alleles modulate immune response to human immunodeficiency virus and hepatitis C virus infections? Front Immunol. 2017;8:832. doi: 10.3389/fimmu.2017.00832, PMID 28769934.
6. Heijmans CMC, de Groot NG, Bontrop RE. Comparative genetics of the major histocompatibility complex in humans and nonhuman primates. Int J Immunogenet. 2020;47(3):243-60. doi: 10.1111/iji.12490, PMID 32358905.
7. Almawi WY, Nemr R, Finan RR, Saldhana FL, Hajjej A, HLA-A. HLA-A, -B, -C, -DRB1 and -DQB1 allele and haplotype frequencies in Lebanese and their relatedness to neighboring and distant populations. BMC Genomics. 2022;23(1):456. doi: 10.1186/s12864-022-08682-7, PMID 35725365.
8. Sanchez-Mazas A. A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. Swiss Med Wkly. 2020;150(150):w20214. doi: 10.4414/smw.2020.20214, PMID 32297957.
9. Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, *et al*. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum Mol Genet. 2019;15;28(12):2078-92. doi: 10.1093/hmg/ddy443, PMID 30590525.
10. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, *et al*. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J. 2018;18(3):367-76. doi: 10.1038/tpj.2017.7, PMID 28440342.
11. Gowda M, Ambardar S. Comparative analyses of low, medium and high-resolution HLA typing technologies for human populations. J Clin Cell Immunol. 2016;07(2):1-8. doi: 10.4172/2155-9899.1000399.
12. Do MD, Le LGH, Nguyen VT, Dang TN, Nguyen NH, Vu HA, *et al*. High-resolution HLA typing of HLA-A, -B, -C, -DRB1, and -DQB1 in Kinh Vietnamese by Using Next-Generation Sequencing. Front Genet. 2020;11:383. doi: 10.3389/fgene.2020.00383, PMID 32425978.

13. Vince N, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, *et al*. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. Genet Epidemiol. 2020;44(7):733-40. doi: 10.1002/gepi.22334, PMID 32681667.
14. Jaramillo A, Hacke K, editors. The human leukocyte antigen system: nomenclature and DNA-based typing for transplantation. Human leukocyte antigens – updates and advances Sevim Gönen; 2023. doi: 10.5772/intechopen.1001105.
15. Nunes E, Heslop H, Fernandez-Vina M, Taves C, Wagenknecht DR, Eisenbrey AB, *et al*. Definitions of histocompatibility typing terms. Blood. 2011;118(23):e180-3. doi: 10.1182/blood-2011-05-353490, PMID 22001389.
16. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, *et al*. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42((Database issue)):D980-5. doi: 10.1093/nar/gkt1113, PMID 24234437.
17. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. Genome Res. 2005;15(11):1592-3. doi: 10.1101/gr.4413105, PMID 16251469.
18. Faviel FF, Gonzalez-Galarza, McCabe A, Jones AR, Middleton D, *et al*. A snapshot of human leukocyte antigen (HLA) diversity using data from the Allele Frequency Net Database. Hum Immunol. 2021;82(7):496-504. doi: 10.1016/j.humimm.2020.10.004, PMID 33755549.
19. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, *et al*. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience. 2017;6(7):1-8. doi: 10.1093/gigascience/gix038, PMID 28531267.
20. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. Nucleic Acids Res. 2020;48(D1):D948-55. doi: 10.1093/nar/gkz950, PMID 31667505.
21. Fürst D, Neuchel C, Tsamadou C, Schrezenmeier H, Mytilineos J. HLA matching in unrelated stem cell transplantation up to date. Transfus Med Hemother. 2019;46(5):326-36. doi: 10.1159/000502263, PMID 31832058.
22. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008;82(1):48-56. doi: 10.1016/j.ajhg.2007.09.001, PMID 18179884.
23. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP – an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011;27(7):968-72. doi: 10.1093/bioinformatics/btr061, PMID 21300701.
24. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210-23. doi: 10.1016/j.ajhg.2009.01.005, PMID 19200528.
25. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, *et al*. Imputing amino acid polymorphisms in human leukocyte antigens. PLOS ONE. 2013;8(6):e64683. doi: 10.1371/journal.pone.0064683, PMID 23762245.
26. Jeanmougin M, Noirel J, Coulonges C, Zagury JF. HLA-check: evaluating HLA data from SNP information. BMC Bioinformatics. 2017;18(1):334. doi: 10.1186/s12859-017-1746-1, PMID 28697761.
27. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, *et al*. HIBAG–HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014;14(2):192-200. doi: 10.1038/tpj.2013.18, PMID 23712092.
28. Shen JJ, Yang C, Wang YF, Wang TY, Guo M, Lau YL, *et al*. HLA-IMPUTER: an easy-to-use web application for HLA imputation and association analysis using population specific reference panels. Bioinformatics. 2019;35(7):1244-6. doi: 10.1093/bioinformatics/bty730, PMID 30169743.
29. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, *et al*. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat Commun. 2021;12(1):1639. doi: 10.1038/s41467-021-21975-x, PMID 33712626.
30. Cook S, Choi W, Lim H, Luo Y, Kim K, Jia X, *et al*. Accurate imputation of human leukocyte antigens with CookHLA. Nat Commun. 2021;12(1):1264. doi: 10.1038/s41467-021-21541-5, PMID 33627654.
31. HLA genotype imputation with attribute bagging (HIBAG). Version 1.34.1; 2022.
32. Nanjala R, Mbiyavanga M, Hashim S, de Villiers S, Mulder N. Assessing HLA imputation accuracy in a West African population. bioRxiv [preprint]. 2023. doi: 10.1101/2023.01.23.525129, PMID 36747714.

Cite this article: Sivaprakasam B, Sadagopan P. HLA Allele Type Prediction: A Review on Concepts, Methods and Algorithms. Asian J Biol Life Sci. 2023;12(2):206-15.