Biology Internal Assessment

Inferring the possibility of ankylosing spondylitis occurrence based on the clade of HLA-B27 phylogenetic tree for the European population

1. Introduction

Human leukocyte antigen (HLA) is the human version of the major histocompatibility complex (MHC) and regulates the immune system by encoding cell-surface proteins that interact with T-lymphocytes [1]. HLA is divided into three classes: class I (HLA-A, HLA-B, and HLA-C), class II (HLA-D), and class III, which incorporates cytokines and complement proteins [1].

Due to its crucial role in the immune system, the HLA system is a primary factor in transplant rejections [1]. Therefore, HLA matching is essential to ensure the recipient's transplant will be accepted [1].

The HLA system is also known for its correlation with autoimmune diseases [1]. For example, HLA-A3 is associated with hemochromatosis [2], HLA-B27 with ankylosing spondylitis (AS) [3], and HLA-B8 with Graves' disease [4].

However, the following will focus only on HLA-B27 and AS for two reasons. Firstly, the author is HLA-B27 positive and have ankylosing spondylitis, and researching this topic will aid his understanding of the disease. Secondly, the mechanism of the disease is unknown, and the presence of HLA-B27 does not necessarily indicate AS [3]. An estimated 8% of the general population possesses an HLA-B27 allele but does not develop AS [5]. Furthermore, there are HLA-B27 alleles that do not show a correlation with the disease [5]. Hence, it is believed that only specific alleles are associated with AS [5].

Consequently, the aim of this investigation is to determine whether HLA-B27 alleles present in the European population (the closest region to the researcher) can be categorized in a way that the presence of a patient's antigen in a certain group becomes a reliable indicator of AS. For this purpose, phylogeny inference tools will be employed, drawing upon the researcher’s expertise in computer science. Moreover, this approach aligns with the researcher’s interest in bioinformatics.

Importantly, molecular phylogeny has consistently yielded promising results since its initial application by Woese and Fox in 1977 [6]. They used it to demonstrate that cellular life can be categorized into three large relatedness groups: eukaryotes, eubacteria, and archaebacteria [6].

For these reasons, the investigation aims to address the research question**: To what extent can the occurrence of ankylosing spondylitis be inferred based on the clade of the HLA-B27 phylogenetic tree for the European population?** It is hypothesized that the presence of a patient's HLA-B27 subtype in certain clades may serve as a reliable biomarker of AS to a significant extent.

2. Methods and tools

2.1. Sequences source

The nucleotide sequence of HLA-B27 alleles were obtained from the IPD-IMGT/HLA database, using the IPD-IMGT/HLA Allele Query Tool [7]. The query to obtain the sequences for Europeans looked as follows:

and(startsWith(name,"B*27"),startsWith(cell_entries.ancestry,"European"))

Table 1. Query for retrieving sequences.

The sequences were then downloaded as Genomic FASTA. The total of 32 sequences was downloaded [8].

2.2. Pre-processing

Before alignment, the file with the sequences was modified to clarify the display of alleles in the phylogenetic tree. The pre-processing was done with a custom Python script, using a Biopython library. The code is hosted online [8]. The script modified the head of each sequence. For instance, "HLA00220|B*27:01|2991 bp" was transformed to "B*27:01 HLA00220|B*27:01|2991 bp" such that each leaf would display only the allele name (e.g. "B*27:01"), excluding its ID.

2.3. Alignment

The sequences were aligned using MAFFT - a multiple sequence alignment software [9]. More specifically, 7.520 version was used. Although MAFFT is dedicated to unix-like operating system (i.e. Linux distributions and macOS versions) [9], it was run on Windows 11 with Windows Subsystem for Linux (WSL) installed. WSL ran Ubuntu 22.04.3 LTS release.

For high accuracy, the Smith-Waterman algorithm was chosen with 16 iterative refinement cycles [9]. Large gaps in the alignment were allowed to avoid any inconsistencies [9]. The output was stored in FASTA format [8]. The whole MAFFT command is reported below.

"/usr/bin/mafft" --ep --localpair --maxiterate 16 --inputorder "european_alleles_head.txt" > "european_alleles_head_aligned.txt"

Table 2. MAFFT command for sequence alignment.

2.4. Phylogenetic tree construction

Two phylogenetic trees from the aligned sequences were created with Python scripts, using Biopython library. The scripts were partially based on the Biopython documentation for the Bio.Phylo module [10], with some modifications that removed unnecessary node names for clarity of the tree. The code is hosted online [8].

The first script created a consensus tree, that is, a summary of the set of trees [11]. The majority rule was applied that chooses the majority split (case that occurs in more than 50% of the tree) for each leaf [11]. The bootstrap values were also calculated to determine how often a taxon corresponds to the taxa of the tree computed for the whole tree (bootstrap values are the measure of confidence) [11]. The length of branches was calculated with the percentage identity that looks for the number of matches between two sequences [11]. The computed tree was stored in the phyloxml format, which is a xml-based format for storing phylogenetic trees and supports all of the above-mentioned elements [8].

The second script did not make a consensus tree, i.e. a series of trees, but a single tree. The bootstrap values were not calculated. However, similarly, percentage identity was computed to

obtain distance matrix that holds valued of relatedness between sequences [11]. Distance matrix was then used by the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm to provide clustering (common ancestors) [11]. Again, the output was stored in the phyloxml format [8].

2.5. Ethical and environmental consideration

The research involves merely the analysis of data that is publicly and freely available on the web from a trusted source (IPD-IMGT/HLA database is directed by a reputable European Molecular Biology Laboratory (EMBL) [12]). The researcher themselves gathered no physical data, that is, they did not perform DNA sequencing and antigen typing on any individual to obtain HLA-B27 alleles for the investigation. Therefore, the research pose no ethical issues regarding Laboratory Safety Guidelines [13].

One might raise concerns about the carbon footprint associated with using a laptop. However, based on a median estimate, the carbon footprint resulting from using a laptop for four years, eight hours a day, is approximately 61.5 kgCO2eq [14]. Even if we assume that the entire research process, including literature review, experimentation, and report writing, took a total of 24 hours (the estimate is exaggerated), the resulting carbon footprint of around ~ 0.126 kgCO2eq ($\frac{24}{8\times365\times4} \times 61.5\ \mathrm{kgCO2eq} \approx 0.126$) is considered negligible.

3. Results and data analysis

The bootstrap tree was generated with a Python script based on the Biopython library documentation. The code is hosted online [8]. The tree visualization is given in the Fig.1.

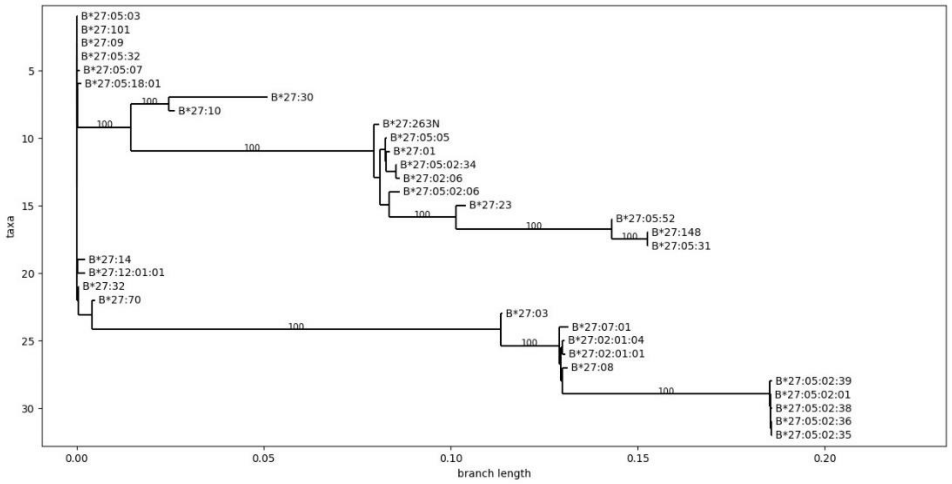

Figure 1. Phylogenetic tree of the alleles with bootstrap values.

The bootstrap values for all the nodes are equal to 100, ensuring a high level of correctness of the taxa of the tree. Some bootstrap values for certain nodes have been remove from the Fig.1 for simplicity and clarity of the depiction.

The second tree was visualized with the same script. It provides identical allele relations, however, it is more readable, which will be an asset in the discussion. It is given in the Fig.2.

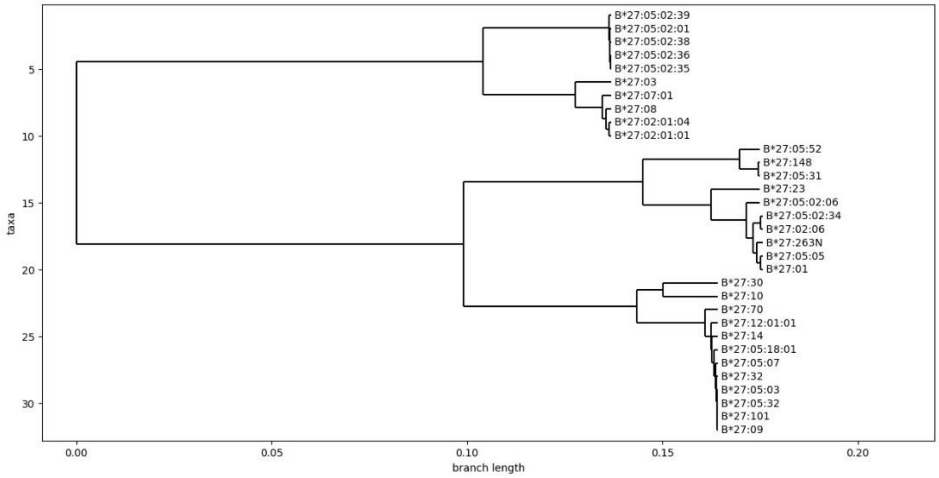

Figure 2. Phylogenetic tree without bootstrap values.

4. Discussion

Several attempts have been made to construct phylogenetic trees for HLA to aid the HLA typing process [15], [16], including separate trees for HLA-B variants [15]. However, there has been limited research specifically on constructing phylogenetic trees for HLA-B27 subtypes and assessing their relationship with AS in the European population. Two studies have presented phylogenetic trees of HLA-B27 subtypes—one on worldwide susceptibility to AS [17] and another with an emphasis on geographic diversity [3]. Unfortunately, comparing these trees with the one shown in Figure 2 is challenging for several reasons. Firstly, the authors do not provide the text files for the trees, sequence alignments, or alleles themselves; they only supply graphical representations. Since one cannot obtain precise numerical data from graphics, a direct comparison of the trees is difficult. Additionally, some data used by these authors is not publicly accessible, rendering the replication of their results impossible. Secondly, even if the data were available, it may not be directly applicable, as the authors did not focus on the European population, necessitating the removal of alleles from outside Europe to avoid affecting the alignment and structure of the phylogenetic tree. The removal, however, could be done only on the file with the alleles, not their alignment, which further complicates the issue. Therefore, the comparison of the results with the results from the literature will be omitted due to the lack of sufficient data for such comparison.

Out of the 17 major alleles (HLA-B27 plus two digits), only 7 have been clinically linked or shown to be unrelated to AS in the European population [18], [19], [20]. The region is a crucial factor because different ethnic groups may share the same allele, but its association with the disease can vary [18]. For example, while B*27:03 is linked to AS in the European population, it is not associated with the disease in the African population [19]. Hence, despite studies establishing the B*27:10 allele's association with the disease in the Chinese population [21], it cannot be considered related to the disease in the European population due to the aforementioned considerations.

Within these constraints, 4 alleles are significantly correlated with the disease (B*27:07, B*27:05, B*27:03 , B*27:02) [18], [20], and 3 have either weak or no correlation (B*27:09, B*27:08, and B*27:01) [18]. The remaining alleles, for which the relation to AS has not been established, include

B*27:101, B*27:30, B*27:263N, B*27:23, B*27:148, B*27:12, B*27:32, and B*27:70, B*27:10, B*27:14.

Figure 2 illustrates several major clades, each will be now numbered for ease of reference. Clade 1, the most distant one, positioned between taxa 20 and 30 with a branch length of 0.17, includes four B*27:05 subtypes (e.g., B*27:05:03, B*27:05:32, B*27:05:07, B*27:05:18:01) and eight other HLA-B27 subtypes (e.g., B*27:30, B*27:10, B*27:70, B*27:12:01:01, B*27:14, B*27:32, B*27:101, B*27:09). Clade 2, located around taxa 15 with a branch length of 0.17, comprises five B*27:05 subtypes (e.g., B*27:05:52, B*27:05:02:06, B*27:05:02:34, B*27:05:05) and five other HLA-B27 subtypes (e.g., B*27:148, B*27:23, B*27:02:06, B*27:263N, B*27:01). Clade 3, positioned between taxa 10 and 5, includes five HLA-B27 subtypes (e.g., B*27:03, B*27:07:01, B*27:08, B*27:02:01:04, B*27:02:01:01). Clade 4, situated at the highest level, consists solely of B*27:05 subtypes (e.g., B*27:05:02:39, B*27:05:02:01, B*27:05:02:36, B*27:05:02:38, B*27:05:02:35).

The clade most directly linked to AS is Clade 4, as it exclusively consists of B*27:05 subtypes. However, this information holds little practical value since this possibility could be deduced without constructing the tree. Even if combined with Clade 3, which includes B*27:03, B*27:02, and B*27:07 alleles, to form a larger clade, the potential association with AS remains ambiguous as Clade 3 contains the B*27:08 allele, which has weak link to the disease.

Clade 2 does not exhibit a clear correlation with AS. Despite the presence of B*27:05 and B*27:02 alleles, many HLA-B27 subtypes within this group have not been definitively associated or disassociated with the disease, possibly due to the rarity of these alleles. Moreover, the presence of the B*27:01 allele, which has been proven not to be linked with the disease, further complicates the interpretation.

Clade 1 comprises many alleles, including B*27:10, B*27:14, and B*27:09, which are not strongly associated with AS. This suggests that it could be considered the "safe" clade, indicating that a patient's antigen presence in this clade may imply a lower risk of developing AS. However, this assumption is quickly disproved by the presence of a significant number of B*27:05 subtypes within this clade. Additionally, the lack of clinical characterization for several alleles, such as B*27:32, further challenges this assumption.

5. Evaluation

5.1 Strengths

A strength of this investigation and the selected scientific approach is its high reproducibility. All results presented can be easily replicated by anyone, as the computations required for aligning sequences and constructing phylogenetic trees are not overly complex and can be performed on standard computing devices such as laptops. This accessibility enhances the trustworthiness of the data, as individuals with basic computational skills can independently verify the findings presented in the study.

Additionally, the reliability of the research is enhanced by the reputable source of the genetic sequences used. EMBL is a well-established institution renowned for its expertise in molecular research [12]. Given its credibility within the scientific community, the data and tools provided by EMBL can be considered trustworthy. Moreover, being a European institute [12], EMBL's resources may be particularly advantageous for studying genetic variations within the European population.

5.2 Weaknesses and further research

A limitation of the study is its reliance on only one sequence alignment algorithm, the Smith-Waterman algorithm. Alternative algorithms, such as the Needleman-Wunsch algorithm [9], could potentially produce different sequence alignments and consequently different phylogenetic trees, which may reveal additional insights into the relationship between HLA-B27 clades and AS. Exploring different alignment methods requires further investigation to ensure robustness and reliability of the findings.

Furthermore, the study is limited by the scarcity of clinical data regarding the association between specific HLA-B27 subtypes and AS. Out of the 17 subtypes considered, only 7 have been clinically linked or shown to be unrelated to AS in the European population [18], [19], [20]. This limited clinical data is primarily due to the rarity of these alleles in the European population [18], highlighting the need for additional research on more prevalent HLA-B27 subtypes to better understand their relationship with AS.

Another weakness is the study's focus solely on the European population, which may not fully reflect the diversity of genetic variations present in today's globalized world. Given the increasing levels of globalization and migration, individuals from diverse ethnic backgrounds interact and intermarry [22], leading to a mixing of genetic alleles across populations. Therefore, future research should investigate the extent to which globalization and migration influence the distribution of HLA-B27 alleles in the European population.

6. Conclusion

The investigation has revealed that the occurrence of ankylosing spondylitis cannot be reliably inferred based solely on the clade of the HLA-B27 phylogenetic tree for the European population. While there may be some correlation between certain HLA-B27 clades and the prevalence of ankylosing spondylitis, the extent of this association is negligible. Ankylosing spondylitis is a complex disease influenced by multiple genetic and environmental factors [18], and relying solely on HLA-B27 clades for inference would overlook many other contributing factors.

7. Bibliography

[1] J.-M. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, R. A. Levy, and R. Cervera, "AUTOIMMUNITY From Bench to Bedside," 1st ed., vol. 368, J.-M. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, R. A. Levy, and R. Cervera, Eds., Bogota: Macmillan, 2013.

[2] J. C. Barton James C. and Barton, "Autoimmune Conditions in 235 Hemochromatosis Probands with HFEC282Y Homozygosity and Their First-Degree Relatives," *J Immunol Res*, vol. 2015, p. 453046, Oct. 2015, doi: 10.1155/2015/453046.

[3] M. A. Blanco-Gelaz, A. López-Vázquez, S. García-Fernández, J. Martínez-Borra, S. González, and C. López-Larrea, "Genetic variability, molecular evolution, and geographic diversity of HLA-B27," *Hum Immunol*, vol. 62, no. 9, pp. 1042–1050, 2001, doi: https://doi.org/10.1016/S0198-8859(01)00299-3.

[4] Y. Orhan, A. Azezli, M. Çarin, F. Aral, E. Sencer, and S. Molvalılar, "Human lymphocyte antigens (HLA) and Graves' disease in Turkey," *J Clin Immunol*, vol. 13, no. 5, pp. 339–343, 1993, doi: 10.1007/BF00920242.

[5] NHS, “Ankylosing spondylitis - causes.”

[6] N. R. Pace, J. Sapp, and N. Goldenfeld, “Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1011–1018, 2012, doi: 10.1073/pnas.1109716109.

[7] J. Robinson *et al.*, “KIR Nomenclature in Non-Human Species,” *Immunogenetics*, 2018.

[8] M. Konat, “HLA-B7 Phylogeny.” [Online]. Available: https://github.com/undeMalum/hla-b27-phylogeny

[9] K. Katoh, “About,” MAFFT.

[10] P. Cock, “Phylo - Working with Phylogenetic Trees,” Wiki Documentation.

[11] J. H. Degnan, M. DeGiorgio, D. Bryant, and N. A. Rosenberg, “Properties of Consensus Methods for Inferring Species Trees from Gene Trees,” *Syst Biol*, vol. 58, no. 1, pp. 35–54, Feb. 2009, doi: 10.1093/sysbio/syp008.

[12] EMBL, “About EMBL.”

[13] J. A. Kaufman, “Laboratory Safety Guidelines,” *The Laboratory Safety Institute*. 2014. [Online]. Available: https://www.labsafety.org/wp-content/uploads/woocommerce_uploads/LSGE-booklet-NFAB.pdf

[14] Circular Computing, “WHAT IS THE CARBON FOOTPRINT OF A LAPTOP?” 2024. [Online]. Available: https://circularcomputing.com/news/carbon-footprint-laptop/

[15] V. Ramsuran *et al.*, “Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes,” *The Journal of Immunology*, vol. 198, no. 6, pp. 2320–2329, Feb. 2017, doi: 10.4049/jimmunol.1601679.

[16] K. Srivastava, K. R. Wollenberg, and W. A. Flegel, “The phylogeny of 48 alleles, experimentally verified at 21 kb, and its application to clinical allele detection,” *J Transl Med*, vol. 17, no. 1, p. 43, 2019, doi: 10.1186/s12967-019-1791-9.

[17] S. Gonzalez-Roces *et al.*, “HLA-B27 polymorphism and worldwide susceptibility to ankylosing spondylitis,” *Tissue Antigens*, vol. 49, no. 2, pp. 116–123, 1997, doi: https://doi.org/10.1111/j.1399-0039.1997.tb02724.x.

[18] S. González *et al.*, “High variability of HLA-B27 alleles in ankylosing spondylitis and related spondyloarthropathies in the population of Northern Spain,” *Hum Immunol*, vol. 63, pp. 673–676, Feb. 2002, doi: 10.1016/S0198-8859(02)00404-4.

[19] N. García-Medel *et al.*, “Peptide Handling by HLA-B27 Subtypes Influences Their Biological Behavior, Association with Ankylosing Spondylitis and Susceptibility to Endoplasmic Reticulum Aminopeptidase 1 (ERAP1),” *Molecular & Cellular Proteomics*, vol. 13, no. 12, pp. 3367–3380, Dec. 2014, doi: 10.1074/mcp.M114.039214.

[20] M. T. Fiorillo, M. Maragno, R. Butler, M. L. Dupuis, and R. Sorrentino, “CD8+ T-cell autoreactivity to an HLA-B27–restricted self-epitope correlates with ankylosing spondylitis,” *J Clin Invest*, vol. 106, no. 1, pp. 47–53, Feb. 2000, doi: 10.1172/JCI9295.

[21] L. Yi *et al.*, “Profiling of hla-B alleles for association studies with ankylosing spondylitis in the chinese population,” *Open Rheumatol J*, vol. 7, pp. 51–54, Aug. 2013.

[22] United Nations, “Globalization of Migration: What the Modern World Can Learn from Nomadic Cultures,” *Migration*, vol. L, no. 3, Sep. 2013.