RESEARCH ARTICLE

Editorial Process: Submission:07/15/2023 Acceptance:10/25/2023

Class I HLA Allele Predicted Restricted Antigenic Coverages for Fap2 Protein of *Fusobacterium Nucleatum* Are Associated with Colorectal Cancer Incidence

Mustapha Zeddou*

Abstract

Objective: This study investigates the association between HLA-A and -B allele diversity, *Fusobacterium nucleatum* Fap2 protein-derived antigenic coverage, and colorectal cancer (CRC) epidemiology across diverse populations. **Methods:** We examined 75 HLA-I alleles and explored 698 potential HLA-A and B-restricted Fap2-derived antigens, assessing how 21 countries may respond to these peptides based on their HLA-I distribution frequencies. Additionally, we correlated in-silico predicted Fap2 population coverage with CRC epidemiology. CRC incidence and mortality data were obtained from the Global Cancer Observatory, and HLA-A and HLA-B allele frequencies from the Allele Frequency Net Database. Binding predictions for Fap2 antigens were performed using netMHCpan4, with stringent selection criteria applied to identify relevant peptides. Population coverage was calculated using the IEDB population coverage tool, and data analysis conducted using the R programming language. **Results:** Clustering of HLA-A and -B allele frequencies partially differentiated countries with lower CRC incidence from others. Distinct patterns of Fap2 protein coverage were observed among different populations. interestingly, we found a significant inverse correlation between CRC incidence ($p = 0.0037$, $R = -0.6$) and predicted Fap2 antigen coverage, as well as CRC mortality ($p = 0.013$, $R = -0.53$). Furthermore, we identified a specific set of Fap2-derived peptides that bind to HLA supertypes, providing a global coverage of 99.04%. **Conclusion:** Our population-based study is the first to demonstrate that higher Fap2 coverage is associated with lower CRC incidence, underscoring the potential significance of Fap2-specific CD8+ T cell responses in CRC tumorigenesis.

Keywords: Colorectal cancer- *Fusobacterium Nucleatum*- Fap2 protein- microbiota dysbiosis- CD8+ T cells

Asian Pac J Cancer Prev, **24 (10)**, 3629-3636

Introduction

Colorectal cancer (CRC) is a significant global public health concern, ranking as the third most commonly diagnosed cancer and accounting for 9.4% of cancer deaths worldwide (Siegel et al., 2022; Sung et al., 2021). The development of CRC involves a complex interplay of genetic and environmental factors. Notably, the gut microbiota has emerged as a critical risk factor in CRC initiation and progression. The human microbiome, consisting of trillions of microorganisms, plays a crucial role in regulating host immunity and nutrition within the gut (Bäumler and Sperandio, 2016). However, under certain conditions, this delicate symbiotic balance can be disrupted, leading to dysbiosis, which contributes to the development of various diseases (Gilbert et al., 2018). In 2012, two independent studies reported an enrichment of Fusobacterium nucleatum (*F. nucleatum*) in colorectal carcinoma tissues, highlighting its potential involvement in CRC pathogenesis (Castellarin et al., 2012; Kostic et al., 2012) .

Numerous studies have elucidated the role of *F. nucleatum* in CRC initiation, progression, and chemoresistance. As an anaerobic bacterium commonly found in the oral cavity, *F. nucleatum* acts as a scaffold for binding with other oral colonizers (Rubinstein et al., 2013; Yu et al., 2017). Although the microbial communities in the oral cavity and the gut exhibit differences in healthy individuals, compromised oral-intestinal barrier integrity allows *F. nucleatum* to be detected in CRC tissues, suggesting its oral origin (Warren et al., 2013). The mechanisms by which *F. nucleatum* contributes to CRC involve promoting cell proliferation and metabolism (Hong et al., 2021; Rubinstein et al., 2013), establishing an inflammatory microenvironment (Kostic et al., 2013), and inhibiting the anti-tumor immune response (Dougall et al., 2017; Gur et al., 2015).

Central to *F. nucleatum's* role in CRC tumorigenesis is the Fap2 protein, a galactose adhesion hemagglutinin. Fap2 binds to galactose-N-acetyl-D-galactosamine

*Laboratory of Agro-Industrial and Medical Biotechnology, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, B.P. 523, Béni Mellal, Morocco. *For Correspondence: m.zeddou@usms.ma, mzeddou@gmail.com*

(Gal-GalNAc), a host factor overexpressed in tumors, facilitating *F. nucleatum* colonization and invasion of CRC cells. This process triggers the secretion of CXCL1 and IL-8, promoting CRC cell metastasis (Abed et al., 2016; Casasanta et al., 2020). Additionally, Fap2 interacts with T cell immunoglobulin and ITIM domain inhibitory receptor (TIGIT), inhibiting cytotoxic activity and enabling immune evasion (Dougall et al., 2017; Gur et al., 2015).

MHC class I-restricted CD8+ T cells play a crucial role in eliminating bacterial infections and conferring protective immunity against diverse bacterial species (Shepherd and McLaren, 2020). Studies have shown an inverse relationship between *F. nucleatum* abundance and CD3+ T cell density within CRC tissues. Higher levels of infiltrating CD8+ T cells have been associated with improved prognosis in CRC patients (Borowsky et al., 2021; Mima et al., 2015; Nosho et al., 2016). However, the specific antigens targeted by these immune cells remain unknown. The effectiveness of inducing a CD8+ T cell response depends on HLA-I molecule presentation, which varies among populations due to HLA-I gene polymorphism. This genetic variation has the potential to influence the epidemiological patterns of CRC. Therefore, we conducted population coverage studies considering the frequencies of 75 HLA-I alleles in different countries to investigate the landscape of 698 potential HLA-A and -B restricted peptides derived from Fap2, aiming to shed light on associations between Fap2 coverage and CRC and aid in identifying peptide candidates for vaccination strategies. CRC represents a significant global health burden, and the dysbiosis associated with *F. nucleatum* has been implicated in CRC initiation and progression. Understanding the specific antigens targeted by CD8+ T cells and the population coverage of Fap2-derived peptides is crucial for elucidating the immunological aspects of CRC and guiding future research and interventions.

Materials and Methods

Epidemiological Data

CRC epidemiological data were obtained from the Global Cancer Observatory: Cancer Today (Bray et al., 2018). This database provides estimates of cancer incidence, mortality, and prevalence in 185 countries, stratified by sex and age group for the year 2020. Supplementary File 1 provides a summary of the epidemiological data used in this study.

HLA-I collecting

The HLA-A and HLA-B allele frequencies of the populations included in the study were obtained from the Allele Frequency Net Database (The Allele Frequency Net Database - Allele, haplotype and genotype frequencies in Worldwide Populations [Internet]. [cited 2023 May 23]. Available from: http://www.allelefrequencies.net/). Datasets with fewer than 50 individuals and those derived from anthropological studies or minority ethnic groups were excluded. The allele frequencies were arranged in descending order, and cumulative allele frequencies (AF) were calculated separately for HLA-A and HLA-B. Only countries that surpassed a defined cumulative AF threshold of 0.75 (Supplementary File 2) were included in the analysis.

Binding Predictions

The primary sequence of the Fap2 protein was retrieved from the NCBI repository (galactose-inhibitable autotransporter adhesin Fap2 [Fusobacterium nucl - Protein - NCBI [Internet]. [cited 2023 Jun 12]. Available from: https://www.ncbi.nlm.nih.gov/protein/WP_273904880.1). To predict binding interactions, the entire Fap2 protein sequence was analyzed using netMHCpan4 with a peptide length of 9 (Jurtz et al., 2017). Peptides that showed complete identity and matching length with protein sequences from the Homo sapiens genome (taxid:9606) were excluded using Blastp v2.9.0. Peptides meeting specific criteria, including high probabilities of proteasomal processing (proteasome cleavage > 0.5) and efficient transport via the transporter associated with antigen processing (TAP score > 0), as determined by netCTL were selected (Larsen et al., 2005). Only strong binder peptides with a %Rank < 0.5 were considered for further analysis.

Population Coverage

Population coverage was calculated using the IEDB population coverage tool (Bui et al., 2006). Strong binder peptides meeting the criteria cited in the Binding Predictions section were included in the epitope-allele file. The user-population file was created using the same allele frequencies as for the binding predictions. The calculation option “Class I separate” was selected. The area under the curve (AUC) was calculated to estimate the degree of coverage before conducting Spearman correlation analysis.

Data and Statistical Analysis

Correlation results were considered significant when the p-value was < 0.05. Heatmaps were generated by clustering rows and columns using the complete method and 1-Pearson correlation coefficient as distance. Spearman correlations were used, unless otherwise specified. Data analysis was performed using the R environment version R.4.2.1, with the following packages utilized: tidyverse (version 2.0.0), bayestestR (version 0.13.0), pheatmap (version 1.0.12), dplyr (version 1.1.0), pROC (version 1.18.0), ggplot2 (version 3.4.1), ggpubr (version 0.6.0), gridExtra (version 2.3), stringr (version 1.5.0), and patchwork (version 1.1.2).

Results

HLA-I Allele Distribution Partially Discriminates Colorectal Cancer incidence Among Countries

The study aimed to investigate the landscape of peptides potentially presented by HLA-I molecules and their association with colorectal cancer (CRC) in different populations. A list of countries with defined HLA genetic frequencies was selected, and allele frequencies of HLA-A and HLA-B were obtained from the allele Frequency Net Database (The Allele Frequency Net Database - Allele, haplotype and genotype frequencies in Worldwide

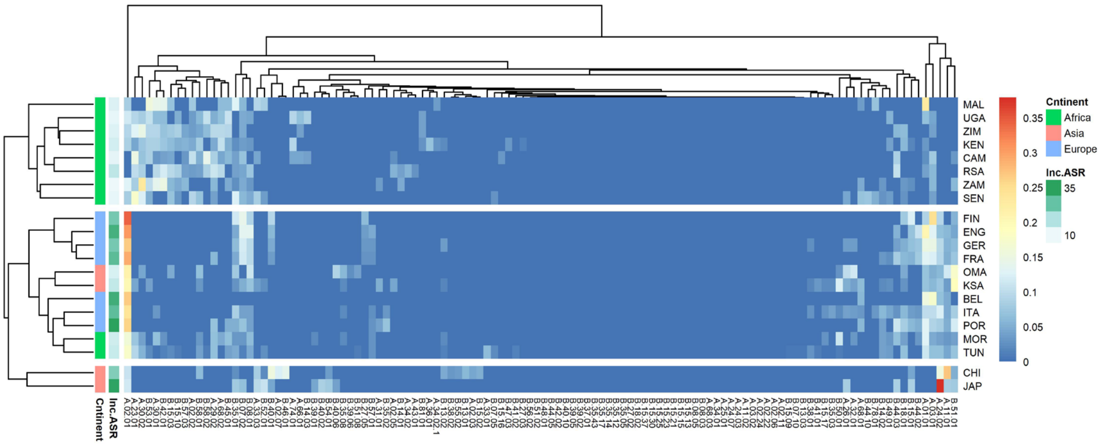

Figure 1. Unsupervised Two-Dimensional Cluster Analysis of 21 Countries by HLA-A and HLA-B Allele Frequencies. Rows represent populations, the first column represents CRC Incidence age-standardized rate (Inc.ASR) quartiles: Q1 = pale blue (lower than 10.5); Q2 = blizzard bleu (10.5 to 14.6); Q3 = light teal (23.9 to 30.1); Q4 = Irish green (greater than 30.1). The other columns represent HLA alleles, which frequency values are represented by the blue-white-yellow-red scale color bar. We used the complete method and 1-Pearson correlation coefficient as distance for clustering columns and rows.

Populations [Internet]. [cited 2023 May 23]. Available from: http://www.allelefrequencies.net/). The countries with cumulative allelic frequencies close to a threshold of 0.75 were considered, resulting in 21 countries and 75 HLA-I alleles for HLA-A and HLA-B, respectively (Supplementary File 1). The median cumulative allele frequency of the considered countries was 0.774 for HLA-A and 0.761 for HLA-B. The number of HLA-A alleles ranged from 4 to 12 (median = 8), and the number of HLA-B alleles ranged from 8 to 27 (median = 13).

CRC-related epidemiological data were retrieved from The Global Cancer Observatory (Supplementary File 2). We investigated whether HLA-I allele frequency distribution among populations could shape CRC epidemiological features. We chose to focus on Cancer incidence, which is the number of new cancer cases arising in a specific population over a period. We used Incidence age-standardized rate (Inc.ASR), as Standardization is essential when comparing several populations that differ concerning age. Age has a strong influence on the risk of cancer. We performed an unsupervised hierarchical clustering to assess whether HLA allele frequency distribution among countries could explain CRC incidence. We used HLA-A and HLA-B allele frequencies for each country. We divided countries into quartiles depending on their Inc.ASR (Figure 1). As expected, results showed that geographical location highly shapes the clustering. Cluster 1 contains exclusively African countries (green, MAL, UGA, ZIM, KEN, CAM, RSA, ZAM, SEN), corresponding to 80% of all analyzed African countries. (Figure 1). Cluster 3, also homogenous regarding the geographic appurtenance, contains only Asian countries (pink, CHI, JPN), corresponding to 50% of the analyzed Asian countries. We observed that countries forming cluster 1 belong to the first two quartiles (Q1: pale bleue and Q2 blizzard bleue, low Inc.ASR). Thus, although the clustering of HLA frequencies reflects the continental distribution of the country, it also seems to separate the first quartiles (low incidence) from the other groups. The most widespread HLA-I alleles in Q1 are HLA-A*.23.01, HLA-A*.30.02 and HLA-B*.51.01, while in Q4 (irish green, higher Inc.ASR) are HLA-A*.24.02, HLA-A*.03.01, HLA-A*.11.01 and HLA-A*.02.01. We obtained the same result when we used the Cancer-mortality age-standardized rate as epidemiological feature. Indeed, the clustering of HLA frequencies appears to

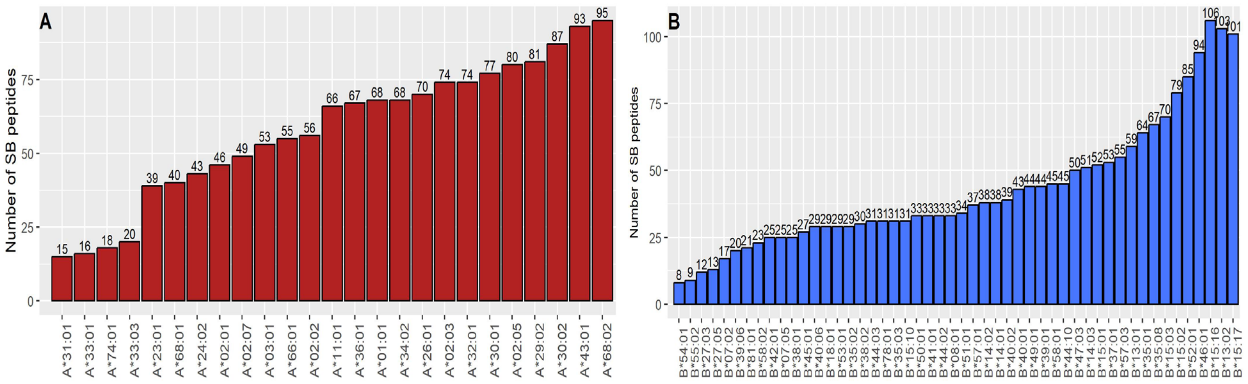

Figure 2. Number of Unique Strong Binder (SB) Peptides for the 75 Analyzed HLA-I Alleles. Global view of the predicted binding spectra of the Fap2 protein-derived SB peptides for 25 HLA-A (A), and 50 HLA-B alleles (B)

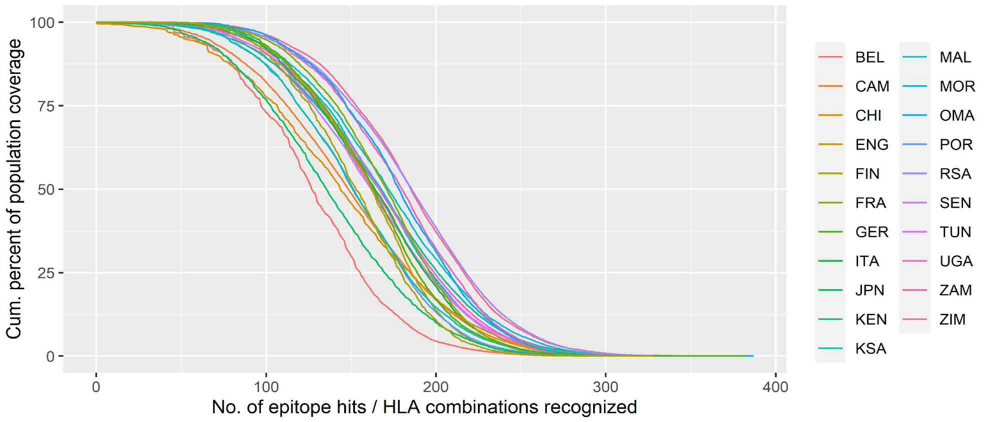

Figure 3. Population Coverage for SB Peptides Derived from Fap2 Protein According to the HLA-I Set of the Present Study. Cumulative population coverage considering the number of Fap2-derived epitope-HLA allele combinations (%). Each population is represented by a line. The abbreviations are not sorted by AUC but are alphabetically ordered in the legends.

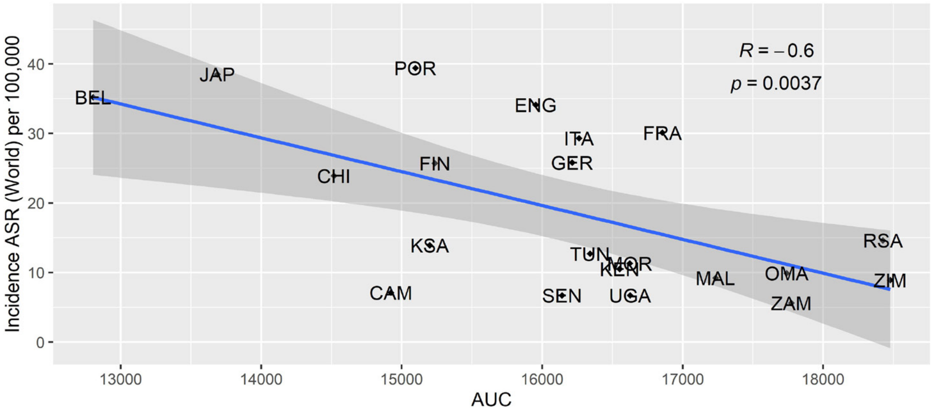

Figure 4. Correlation between Fap2-Derived SB Peptides Coverage of Each Population and CRC Incidence. Spearman correlations between AUC calculated from the population coverage considering the entire Fap2 protein and CRC incidence.

separate the first quartiles (low mortality) from the other groups (data not shown). This result underlines the possible implication of HLA alleles in the differential response to CRC among countries. Given the role of the Fap2 protein of *F. nucleatum* in the onset and development of CRC, we postulated that studying the binding spectra of each HLA-I allele for each Fap2-derived peptide might help better understand epidemiological data. Therefore, we decided to perform Fap2 protein-derived peptides binding prediction to HLA-I alleles.

HLA-I Binding Patterns of Fap2-Derived Peptides across Diverse Populations

We utilized the netMHCpan4 tool (Jurtz et al., 2017) to predict the binding affinities of 33 HLA-A and 73 HLA-B alleles for the potential peptides derived from the Fap2 protein. To ensure high-quality binding predictions, we applied multiple filters. Firstly, we retained only strong binders (SB) with a % rank <0.5. Secondly, we used the BLASTp tool to eliminate peptides that matched 100% identity with sequences in the human proteome. Lastly, we included peptides meeting the criteria for proteasomal processing (proteasome cleavage >0.5) and efficient transport via the transporter associated with antigen processing (TAP score > 0), as predicted by netCTL. Through this stringent filtering process, we obtained a total of 698 peptides, with 362 peptides binding to HLA-A and 336 peptides binding to HLA-B.

Among the HLA-A alleles, HLA-A68.02 demonstrated the highest capacity for presenting a greater number of SB peptides, with 95 unique peptides. It was followed by HLA-A43.01, which presented 93 unique peptides, and HLA-A30.02, which presented 87 unique peptides. In contrast, HLA-A31.01, a specialist allele, was expected to present the lowest number of peptides, with only 15 unique peptides (Figure 2A).

Notably, HLA-A43.01, one of the most generalist HLA-A alleles for the Fap2 protein, was found to be specific to RSA. Conversely, HLA-A31.01 and HLA-A33.01, two highly specialized HLA-A alleles, were country-specific, being associated with KSA and Por, and TUN, respectively. The most generalist HLA-A allele, HLA-A68.02, was exclusively found in countries falling within the first two quartiles (Q1, Q2), with a median allele frequency of 0.0731 and a mean of 0.0747.

In terms of HLA-B alleles, the highest capability for presenting a greater number of SB peptides was observed in HLA-B15.16 (106 unique peptides, specific to CAM), HLA-B13.02 (103 unique peptides), and HLA-B15.17 (101 unique peptides). Conversely, HLA-B54.01 (8 unique peptides, specific to CHI and JAP) and HLA-B*55.02 (9 unique peptides, specific to CHI) were predicted to present the lowest number of peptides (Figure 2B).

Distinct Patterns of Fap2 Protein Coverage Are Observed Among Populations

To assess the immune response potential within different populations, we analyzed the HLA-restricted strong binder (SB) epitopes derived from the Fap2 protein. These epitopes possess sufficient affinity for binding to HLA molecules. By considering the frequencies of HLA alleles, we estimated the proportion of individuals in each population predicted to mount a response, known as population coverage. We employed the IEDB population coverage tool to calculate this coverage, which provided us with the number of epitopes/HLA combinations required to achieve a specified rate of coverage for each population. Two parameters were considered in assessing population coverage: the minimum number of epitopes/HLA combinations necessary to cover 90% of the population (MNEC) and the Area Under the Curves (AUC). The MNEC analysis revealed substantial cumulative coverage across the countries, ranging from 71 in CHI to 124 in ZIM, with an average of 100.95 ± 15.37 (Figure 3) and (Supplementary File 3). A comparison of AUC values highlighted varying levels of coverage among the populations studied. The lowest coverage was observed in BEL and JAP, with AUC values of 12,804.96 and 13,688.9, respectively. In contrast, ZIM and RSA exhibited higher levels of coverage, with AUC values of 18,478.12 and 18,426.12, respectively (Figure 3). These findings indicate distinct patterns of Fap2 protein coverage among populations, reflecting the variability in HLA allele frequencies and epitope presentation capacities.

Colorectal Cancer Epidemiological Parameters Are Associated With Fap2 Antigen Coverage

To investigate the relationship between population coverage of Fap2 epitopes and the occurrence of CRC in different countries, we analyzed the Area Under the Curve (AUC) for each country and its corresponding CRC incidence and mortality rates. To account for the influence of age on cancer risk, we utilized age-standardized rates (Inc.ASR for cancer incidence and MOR.ASR for mortality). Our analysis revealed a significant inverse correlation between the calculated coverage of each country and CRC incidence (Inc.ASR) (p=0.0037, R= -0.6) (Figure 4). Similar findings were observed when considering mortality (MOR.ASR) as the epidemiological data (p=0.013, R= -0.53) (Data not shown). However, when focusing on countries within the first quartiles (Q1 and Q2), no significant correlation was found between the HLA alleles predicted to present strong binder (SB) peptides and Inc.ASR (p=0.92, R= 0.031) or MOR.ASR (p=0.52, R= -0.21) (Data not shown).

To explore the peptides implicated in the observed correlation between population coverage for Fap2 peptides and CRC epidemiological data, we selected the top ten major histocompatibility complex-I restricted epitopes for both HLA-A and HLA-B considering the selection criteria mentioned in the Binding Predictions section. Some of these top binder (TB) peptides were associated with HLA alleles present in Q4 (Figure 1). However, we found no significant correlation between the calculated population coverage of these TB peptides and Inc.ASR (p=0.23, R=-0.27) or MOR.ASR (p=0.18, R= -0.31) (Data not shown).

In our search for potential vaccine peptides, we focused on HLA alleles expressed in countries within quartiles Q1 and Q2, while excluding Q4. This selection process yielded 293 candidate peptides with strong binding potential (SB peptides). Estimating world population coverage using these 293 peptides, we found a coverage rate of 20.57%. Interestingly, coverage percentages varied across specific geographical regions. South Africa (56.27%), West Africa (55.58%), and North Africa (46.59%) displayed the highest coverage, while Northeast Asia (6.1%), Southeast Asia (2.17%), and Europe (11.61%) showed lower coverages. Central Africa (50.57%), South Asia (18.28%), Southwest Asia (18.28%), and North America (24.17%) exhibited intermediate coverage levels. From the initial pool of 293 SB peptides, we further selected five peptides with the highest binding strength. Together, these five peptides achieved the same level of coverage (20.57%) across the global population (Supplementary File 4).

HLA supertypes refer to groups of HLA molecules sharing similar characteristics in terms of their peptide-binding properties (Sette and Sidney, 1999). To assess the potential coverage of HLA supertypes within the global population, we focused on nine HLA supertypes: HLA-A1, -A2, -A3, -A24, -B7, -B27, -B44, -B58, and -B62. Using seven SB peptides that demonstrated binding affinity to eight of these HLA supertypes, we estimated the populational coverage on a global scale. Our analysis revealed an impressive average world population coverage of 99.04%. This indicates that the combination of these seven peptides has the potential to activate immune responses in a significant proportion of the global population.

Discussion

In this study, we investigated the varying susceptibility observed in different countries based on their HLA-I allele profiles and how it relates to CRC epidemiology. Through the utilization of multiple in-silico approaches, we predicted the population's immune response to the Fap2 protein derived from *F. nucleatum* and examined its correlation with CRC incidence and mortality rates. Our findings suggest that favorable outcomes may depend, in part, on the population's ability to present Fap2 protein-derived peptides effectively.

To better represent each country, we utilized HLA-I allele frequencies obtained from blood donor registries, prioritizing this data over anthropology studies. However, the unavailability of HLA-C gene data in many countries prevented its inclusion in our analysis. The HLA region on chromosome 6p21.3 contains numerous genes, resulting in the transmission of HLA-I haplotypes being influenced by linkage disequilibrium (LD) (Marrack and Kappler, 1986). While LD correction typically requires haplotype frequencies instead of allele frequencies (AF), this information is often not available for many countries. Nonetheless, studies suggest that the impact of LD on population coverage calculations is considered negligible (Bui et al., 2006).

T lymphocyte recognition of pathogen-derived epitopes is restricted to a specific major histocompatibility complex (MHC) (Zinkernagel and Doherty, 1974). Thus, an epitope will induce a response only in individuals expressing an HLA molecule able to bind that particular epitope. The extreme polymorphism of HLA molecules has led to the discovery of over a thousand different HLA allelic variants ("HLA Informatics Group | Anthony Nolan [Internet]. [cited 2023 May 23]. Available from: https://www.anthonynolan.org/clinicians-and-researchers/scientists-and-researchers/hla-informatics-group,"). having dramatically different frequency distributions among ethnicities (Imanishi: HLA 1991, Proceedings of the Eleventh Internati.. 2023). HLA-binding predictions have become valuable tools in the development of T-cell epitope-based vaccines, enabling the selection of numerous epitopes specific to different HLA alleles. Population coverage studies are also essential in the context of genetically diverse human populations. Our study provides insights into the association between Fap2 antigen coverage and CRC incidence, highlighting the potential role of Fap2-specific CD8+ T cells in conferring immunity against CRC. This finding aligns with previous research that emphasizes the critical role of CD8+ T lymphocytes in the immune response against CRC (Titu et al., 2002; Tran et al., 2016).

Numerous studies have established a strong correlation between CRC and the lifestyles and diets prevalent in developed countries, which account for over 63% of all CRC cases (Janout and Kollárová, 2001). However, our study revealed intriguing findings regarding countries in quartiles Q1 and Q2, primarily situated in Africa and characterized by lower- to middle-income levels, as they exhibited the lowest incidences of CRC (Figure 1). These observations suggest that African countries may possess a certain level of protection against CRC due to lifestyle factors that promote less dysbiosis when compared to Western countries. Western diets have been described to contribute to dysbiosis of the intestinal microbiota, a condition closely associated with CRC (Wang and Fang, 2023). Additionally, recent research has identified *F. nucleatum* as a potential pathogen involved in cancer development, being enriched in tumor tissues of CRC patients where it colonizes and invades using the Fap2 adhesin (Abed et al., 2016). To gain further insights into the relationship between Western lifestyle and CRC, studies examining cancer risk among non-Western migrants to Europe are of great significance. These studies consistently indicate that migrants originating from non-Western countries have a lower likelihood of developing cancers linked to Western lifestyles, including CRC (Arnold et al., 2010). By considering these studies in conjunction with the findings from our study, we propose an additional explanation for the relatively lower incidence of CRC in African countries. In addition to the well-established role of Western lifestyles and diets in CRC pathogenesis, our study suggests that better Fap2 coverage in African populations may contribute to enhanced immune control of *F. nucleatum* invasion. Thus, our findings support the hypothesis that the observed lower CRC incidence in African countries may be associated with their improved ability to mount immune responses against *F. nucleatum* invasion, thanks to favorable Fap2 coverage. (Figure 4). These findings have implications for clinicians and researchers in understanding the role of Fap2-specific CD8+ T cell responses and the potential significance of Fap2 antigen coverage in CRC tumorigenesis. Furthermore, they highlight the need for continued exploration of the intricate interactions between lifestyle, microbiota, pathogens, and CRC occurrence across diverse populations. That said, it is essential to acknowledge the inherent limitations and potential factors contributing to the observed differences in our study findings, as it ensures a comprehensive understanding of the research landscape. Factors such as CRC data quality and coverage disparities between Western countries and those with low- to middle-income levels can influence the generalizability of our results. Further research is needed to deeply investigate the complex interactions between lifestyle, microbiota, pathogens, and CRC incidence in different populations.

Vaccination strategies employing cytotoxic T-lymphocyte (CTL) epitopes offer a promising approach for immunization compared to delivering whole proteins as immunogens. However, the application of CTL epitope-based vaccines is limited by the restriction of epitope recognition to specific cognate HLA alleles. The identification of HLA supertypes has significantly improved the ability to immunize a diverse ethnic population using a narrow selection of peptides. HLA supertypes enable the design of candidate vaccines comprising a small number of peptides, each presented by multiple HLA-I alleles with high frequencies across different populations (Sidney et al., 2008). In our study, we focused on the strong binding peptides presented by HLA alleles expressed in countries with low cancer incidence (Q1 and Q2), while absent in Q4. This selection yielded 293 peptides, resulting in an average calculated world coverage of 20.57%. It is important to note that this relatively low coverage level was anticipated, considering that world coverage calculations encompass populations that do not express the selected HLA alleles. Additionally, we prioritized HLA-I supertypes and identified seven SB peptides that bind to eight prominent HLA supertypes (Supplementary File 5). These seven vaccine-candidate peptides demonstrated an average world coverage of 99.04%. In a related study, Padma et al. attempted to design vaccine candidates for therapeutic intervention in human CRC by selecting nine MHC-I epitopes of Fap2 using the VaxiJen tool (Padma et al., 2023). Notably, none of these epitopes are included in the list of SB epitopes that meet the IEDB criteria established in our study. It is worth mentioning that such disparities are observed even within studies employing the same computational tool, let alone when different tools are utilized (Sohail et al., 2021).

As the utility of in-silico tools for epitope prediction gains recognition, these discrepancies underscore the necessity for further research and systematic experimental validation to enhance their accuracy and reliability.

Our study utilized a populational approach to investigate the association between Fap2 antigen coverage, CRC incidence, and Fap2-specific CD8+ T cell responses. The findings demonstrated diverse patterns of Fap2 protein coverage among different populations, with a negative correlation observed between CRC incidence and predicted Fap2 antigen coverage. This suggests a potential role for Fap2-specific CD8+ T cell responses in CRC tumorigenesis. The results highlight the importance of understanding the immunological aspects of CRC and provide valuable insights for future research and interventions in this field. Further investigations are warranted to explore the intricate interactions between lifestyle, microbiota, pathogens, and CRC occurrence across diverse populations, with a focus on the specific antigens targeted by immune cells in CRC patients.

Author Contribution Statement

I, Dr Mustapha ZEDDOU, am the sole author of this study. I conceived the research idea, conducted data collection, analysis, and interpretation independently. As the sole contributor, I take full responsibility for all aspects of this manuscript..

Acknowledgements

Funding statement

The present study benefited from in-house financial support from the Laboratory of Agro-Industrial and Medical Biotechnology, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Béni Mellal, Morocco.

Study Approval

The research committee of the Laboratory of Agro-Industrial and Medical Biotechnology, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, approved this project.

Ethical approval

Ethical approval was not required for this study as it involved open-source dataset analysis. The research did not encompass human subjects, animal experimentation, or any confidential or sensitive information.

Data availability

The data is accessible upon request to the author.

Conflicts of interests

The authors declare no conflict of interest.

References

Abed J, Emgård JE, Zamir G, et al (2016). Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe*, **20**, 215-25.

Arnold M, Razum O,Coebergh JW (2010). Cancer risk diversity in non-western migrants to Europe: An overview of the literature. *Eur J Cancer*, **46**, 2647-59.

Bäumler AJ, Sperandio V (2016). Interactions between the microbiota and pathogenic bacteria in the gut. *Nature*, **535**, 85-93.

Borowsky J, Haruki K, Lau MC, et al (2021). Association of Fusobacterium nucleatum with Specific T-cell Subsets in the Colorectal Carcinoma Microenvironment. *Clin Cancer Res*, **27**, 2816-26.

Bray F, Ferlay J, Soerjomataram I, et al (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, **68**, 394-424.

Bui HH, Sidney J, Dinh K, et al (2006). Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, **7**, 153.

Casasanta MA, Yoo CC, Udayasuryan B, et al (2020). Fusobacterium nucleatum host-cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration. *Sci Signal*, **13**.

Castellarin M, Warren RL, Freeman JD, et al (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*, **22**, 299-306.

Dougall WC, Kurtulus S, Smyth MJ, Anderson AC (2017). TIGIT and CD96: new checkpoint receptor targets for cancer immunotherapy. *Immunol Rev*, **276**, 112-20.

galactose-inhibitable autotransporter adhesin Fap2 [Fusobacterium nucl - Protein - NCBI [Internet]. [cited 2023 Jun 12]. Available from: https://www.ncbi.nlm.nih.gov/protein/WP_273904880.1.

Gilbert JA, Blaser MJ, Caporaso JG, et al (2018). Current understanding of the human microbiome. *Nat Med*, **24**, 392-400.

Gur C, Ibrahim Y, Isaacson B, et al (2015). Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity*, **42**, 344-55.

HLA Informatics Group | Anthony Nolan [Internet]. [cited 2023 May 23]. Available from: https://www.anthonynolan.org/clinicians-and-researchers/scientists-and-researchers/hla-informatics-group.

Hong J, Guo F, Lu SY, et al (2021). *F. nucleatum* targets lncRNA ENO1-IT1 to promote glycolysis and oncogenesis in colorectal cancer. *Gut*, **70**, 2123-37.

Imanishi HLA (1991). Proceedings of the Eleventh Internati.. (2023). ... - Google Scholar [Internet]. [cited 2023 May 23]. Available from: https://scholar.google.com/scholar_lookup?title=HLA+1991:+Proceedings+of+the+Eleventh+International+Histocompatibility+Workshop+and+Conference&author=T+Imanishi&author=T+Akaza&author=A+Kimura&author=K+Tokunaga&author=T+Gojoubori&publication_year=1992&.

Janout V, Kollárová H (2001). Epidemiology of colorectal cancer. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, **145**, 5-10.

Jurtz V, Paul S, Andreatta M, et al (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol*, **199**, 3360-8.

Kostic AD, Chun E, Robertson L, et al (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe*, **14**, 207-15.

Kostic AD, Gevers D, Pedamallu CS, et al (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res*, **22**, 292-8.

Larsen MV, Lundegaard C, Lamberth K, et al (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*, **35**, 2295-303.

Marrack P, Kappler J (1986). The antigen-specific, major histocompatibility complex-restricted receptor on T cells. *Adv Immunol*, **38**, 1-30.

Mima K, Sukawa Y, Nishihara R, et al (2015). Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. *JAMA Oncol*, **1**, 653-61.

Nosho K, Sukawa Y, Adachi Y, et al (2016). Association of Fusobacterium nucleatum with immunity and molecular alterations in colorectal cancer. *World J Gastroenterol*, **22**, 557-66.

Padma S, Patra R, Sen Gupta PS, et al (2023). Cell Surface Fibroblast Activation Protein-2 (Fap2) of Fusobacterium nucleatum as a Vaccine Candidate for Therapeutic Intervention of Human Colorectal Cancer: An Immunoinformatics Approach. *Vaccines (Basel)*, **11**.

Rubinstein MR, Wang X, Liu W, et al (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe*, **14**, 195-206.

Sette A, Sidney J (1999). Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, **50**, 201-12.

Shepherd FR, McLaren JE (2020). T Cell Immunity to Bacterial Pathogens: Mechanisms of Immune Control and Bacterial Evasion. *Int J Mol Sci*, **21**, 1.

Sidney J, Peters B, Frahm N, Brander C, Sette A (2008). HLA class I supertypes: a revised and updated classification. *BMC Immunol*, **9**, 1.

Siegel RL, Miller KD, Fuchs HE, Jemal A (2022). Cancer statistics, 2022. *CA Cancer J Clin*, **72**, 7-33.

Sohail MS, Ahmed SF, Quadeer AA, McKay MR (2021). In silico T cell epitope identification for SARS-CoV-2: Progress and perspectives. *Adv Drug Deliv Rev*, **171**, 29-47.

Sung H, Ferlay J, Siegel RL, et al (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, **71**, 209-49.

The Allele Frequency Net Database - Allele, haplotype and genotype frequencies in Worldwide Populations [Internet]. [cited 2023 May 23]. Available from: http://www.allelefrequencies.net/.

Titu LV, Monson JR,Greenman J (2002). The role of CD8(+) T cells in immune responses to colorectal cancer. *Cancer Immunol Immunother*, **51**, 235-47.

Tran E, Robbins PF, Lu YC, et al (2016). T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med*, **375**, 2255-62.

Wang N, Fang JY (2023). Fusobacterium nucleatum, a key pathogenic factor and microbial biomarker for colorectal cancer. *Trends Microbiol*, **31**, 159-72.

Warren RL, Freeman DJ, Pleasance S, et al (2013). Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*, **1**, 16.

Yu T, Guo F, Yu Y, et al (2017). Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell*, **170**, 548-63.

Zinkernagel RM, Doherty PC (1974). Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature*, **248**, 701-2.