# *In silico* pipeline to identify tumor-specific antigens associated with frequent HLA-I alleles in the Costa Rican Central Valley Population

**Diego Morazán-Fernández**
Caja Costarricense del Seguro Social, Costa Rica
**Javier Mora**
Universidad de Costa Rica, Costa Rica
**Jose Arturo Molina-Mora** ( jose.molinamora@ucr.ac.cr )
Universidad de Costa Rica, Costa Rica https://orcid.org/0000-0001-9764-4192

**Method Article**

**Keywords:** Neopeptide, Colorectal cancer, Cancer vaccine, Single nucleotide variants, HLA, Costa Rica

**Posted Date:** May 3rd, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1609267/v1

# Abstract

Colorectal cancer is a complex disease in which uncontrolled growth of abnormal cells occurs in the large intestine (colon or rectum). The study of tumor-specific antigens (neoantigens), molecules that interact with the immune system, has been extensively explored as a possible therapy called *in silico* cancer vaccine. Cancer vaccine studies have been triggered by the current high-throughput DNA sequencing technologies. However, there is no universal bioinformatic protocol to study tumor-antigens with DNA sequencing data.

We propose a bioinformatic protocol to detect tumor-specific antigens associated with single nucleotide variants (SNVs) or “mutations” in colorectal cancer and their interaction with frequent HLA alleles (complex that present antigens to immune cells) in the Costa Rican Central Valley population. We used public data of human exome (DNA regions that produce functional products, including proteins). A variant calling analysis was implemented to detect tumor-specific SNVs, in comparison to healthy tissue. We then predicted and analyzed the peptides (protein fragments, the tumor-specific antigens) derived from these variants, in the context of their affinity with frequent alleles of HLA type I in the Costa Rican population.

We found 28 non-silent SNVs, present in 26 genes. The protocol yielded 23 strong binders peptides derived from the SNVs for frequent alleles (greater than 8%) for the Costa Rican population at the HLA-A, B, and C loci. It is concluded that the standardized protocol was able to identify neoantigens and this can be considered a first step for the eventual design of a colorectal cancer vaccine for Costa Rican patients. To our knowledge, this is the first study of an *in silico* cancer vaccine using DNA sequencing data in the context of the Costa Rican HLA alleles.


# Introduction

Colorectal cancer is a complex disease in which uncontrolled growth of abnormal cells occurs in the large intestine (colon or rectum). By year, colorectal cancer worldwide reached an incidence of > 1.8 million cases and > 8 thousand deaths, occupying the third and second cause of cancer incidence and deaths respectively (World Health Organization, 2020). Cancer vaccines and others immunotherapies have emerged as an alternative to treat this type of cancer. As summarized in Fig. 1, there is evidence that immune system cells, such as CD8 + cytotoxic lymphocytes, can recognize tumor-specific peptides or protein fragments derived from gene mutations or variants in malignant cells. In the immune response, these neopeptides are presented by cells (cancer or normal) to CD8 + cells. The antigen presentation occurs with the participation of the human major histocompatibility complex in tumor (and normal) cells defined by specific HLA alleles, and the TCR (T-cell receptor) in CD8 + cells. If the antigen is tumor-specific, the death of tumor cells is induced (Palucka & Coussens, 2016).

With the advance in high-throughput DNA sequencing technologies, cancer vaccines and tumor/ immune system interaction cells have been extensively studied (Ding, Wendl, McMichael, & Raphael, 2014; Hackl,

Charoentong, Finotello, & Trajanoski, 2016; Spencer, Zhang, & Pfeifer, 2014). Cancer vaccine studies must consider HLA alleles ("gene versions") that are frequent in the target population, the HLA type (type I or HLA-I for interaction with CD8 + cells), the peptide size (between 8–11 amino acids for HLA-I) and others (X. S. Liu & Mardis, 2017). However, the data complexity and the specific steps do not allow the application of universal protocols. Benchmark strategies are recommended for specific cases and populations (Heather & Chain, 2016; Rizzo & Buck, 2012; Su et al., 2011). Thus, this work aimed to implement a bioinformatic protocol of an *in silico* cancer vaccine to identify tumor-specific peptides in colorectal cancer considering the context of frequent HLA-I alleles from the Costa Rican Central Valley Population.

## Methods

General workflow of the analysis is shown in Fig. 2.

# Sequence data of tumor samples and pre-processing

Raw data from chromosome 1 of 7 exomes from human colorectal cancer (4 primary tumor and 3 from liver metastasis) and normal mucosa samples (all from the same patient) were downloaded from NCBI Bioproject PRJNA271316 CHET9 experiment (sequenced by Ilumina® technology, access: www.ncbi.nlm.nih.gov/bioproject/PRJNA271316). The exome reads were downloaded in Fastq format for subsequent analyzes. Quality control was done using FastQC, and trimming of reads was done with Trimmomatic, including filtering by quality and adapters removal (Blankenberg et al., 2010).

# Mapping to human genome and Variant Calling analysis

To map reads to the human reference genome (version hg19), the BWA program (Li & Durbin, 2009) was used with default parameters. Samtools (Li et al., 2009) was used for extra data manipulation as usual. Picard (http://broadinstitute.github.io/picard/) was used to evaluate the quality of the mapping. VarScan2 program (Koboldt et al., 2012)was used to detect SNVs with quality Q higher than 20 in all samples; other parameters were used by default in the deduplication and re-calibration steps. To select mutations present in all the tumor samples, VCFlib (https://github.com/vcflib/vcflib) was used. Subsequently, the SNPeff program (Cingolani, Platts, et al., 2012) was used to evaluate and compare the variants between samples. SnpSift (Cingolani, Patel, et al., 2012) was used to filter all those SNVs that were in the normal mucosa sample to finally obtain the tumor exclusive variants. Finally, these mutations were annotated with ANNOVAR (K. Wang, Li, & Hakonarson, 2010), obtaining the description of the SNVs and their possible effects on genes. The Atlas repository of colorectal cancer was used to see the involvement of the genes with selected SNVs and the possible metabolic or signaling pathways affected (Chisanga et al., 2016).

# Putative tumor-specific peptides (neoantigens) identification

Selected tumor-specific variants were analyzed with MuPeXi package (Mette, Morten, & Hadrup, 2017), to generate the peptides (8–11 amino acid residues) to the different alleles of HLA type I and the possible priorities of related mutant peptides. We chose those HLA type I alleles with frequencies greater than 8% for the Costa Rican Central Valley population (Arrieta-Bolaños et al., 2011), using Allele Frequency Net Database (http://www.allelefrequencies.net/default.asp). Due to HLA data had a resolution of two-digits, HLA supertypes (four-digits) were considered.

Mutant peptides were analyzed with the NetMHCpan program (Nielsen & Andreatta, 2016), to establish the affinity of these peptides to the respective alleles of HLA type I utilizing the rank percentile (% rank). The 0.5% threshold was used for strong binders. Finally, the characterization of peptides was done using the Protparam tool of the ExPasy server (Gasteiger et al., 2005), which provides physicochemical properties.

Table 1

**Strong binding peptides for most frequent HLA alleles of the Costa Rican population.** Candidates are derived from the selected SNVs in the primary and metastatic tumor samples (pI: Isoelectric point, GRAVY: Grand average of hydropathy).

| Alleles | Peptides | pI | Acidic nature | GRAVY | Lenght (AA) |
|---|---|---|---|---|---|
| A*02:01 | KLVGFSLDV | 5.84 | acidic | 1.133 | 9 |
| | LLIHCDAYL | 5.80 | acidic | 1.356 | 9 |
| | YLHSPMYFFI | 6.74 | neutral | 0.76 | 10 |
| | ILLIHCDAYL | 5.08 | acidic | 1.67 | 10 |
| A*03:01 | AMNILEINEK | 4.53 | acidic | -0.14 | 10 |
| | GSSFYALEK | 6.00 | neutral | -0.256 | 9 |
| A*24:02 | AYLHSPMYFF | 6.78 | neutral | 0.49 | 10 |
| | AYLHSPMYFFI | 6.78 | neutral | 0.855 | 11 |
| | CDAYLHSPMYF | 5.08 | acidic | 0.1 | 11 |
| | DAYLHSPMYF | 5.08 | acidic | -0.14 | 10 |
| | DAYLHSPMYFF | 5.08 | acidic | 0.127 | 11 |
| | RWYSTPSSYL | 8.59 | basic | -0.89 | 10 |
| | YLHSPMYFF | 6.74 | neutral | 0.344 | 9 |
| B*07:02 | GPLSSEKAAM | 6.00 | neutral | -0.17 | 10 |
| B*35:01 | DAYLHSPMY | 5.08 | acidic | -0.467 | 9 |
| B*40:01 | AEINILEI | 3.80 | acidic | 1.075 | 8 |
| | EEKLVGFSL | 4.53 | acidic | 0.278 | 9 |
| | LEEKLVGFSL | 4.53 | acidic | 0.63 | 10 |
| | RELTHLREKL | 8.75 | basic | -1.24 | 10 |
| | SEKAAMNIL | 5.52 | acidic | 0.233 | 9 |
| B*44:02 | AEINILEI | 3.80 | acidic | 1.075 | 8 |
| C*03:02 | KAAMNILEI | 6.00 | neutral | 0.822 | 9 |
| C*04:01 | AYLHSPMYF | 6.78 | neutral | 0.233 | 9 |
| | RWYSTPSSYL | 8.59 | basic | -0.89 | 10 |
| | WYSTPSSYL | 5.52 | acidic | -0.489 | 9 |

| Alleles | Peptides | pI | Acidic nature | GRAVY | Lenght (AA) |
|---|---|---|---|---|---|
| | YLHSPMYFF | 6.74 | neutral | 0.344 | 9 |
| C*06:02 | WYSTPSSYL | 5.52 | acidic | -0.489 | 9 |
| C*07:01 | WYSTPSSYL | 5.52 | acidic | -0.489 | 9 |

# Results And Discussion

Identification of neoantigens or tumor-specific antigens was described at the end of the last century (Emens et al., 2017). DNA sequencing technologies and bioinformatic protocols have triggered the understanding of the interactions of tumor cells with the immune system, including the design of cancer vaccines. This strategy is not suitable for all cancer types due to a relatively high mutation rate is expected to generate new antigens. In colorectal cancer, melanoma, and other tumor types, a high number of neoantigens is associated with patient response to immune therapies, making them suitable for cancer vaccines.

In this context, to detect possible variants with the potential to induce tumor-specific peptides in samples of colorectal tumors, a bioinformatic protocol was implemented with samples of human colorectal cancer for chromosome 1. We identified 395 SNVs shared by all tumor samples (and absent in the normal mucosa sample). See the *Supplementary file* for the whole list of SNVs.

Despite this, only 28 non-silent SNPs in exonic regions were identified (see *Supplementary file* for details). They were selected to predict neoantigens. These variants were distributed in 26 genes, most of them related to cell cycle control. Interestingly, multiple of these genes, including NOTCH, OBSCN, HSPG2, and NBPF, have been demonstrated to be expressed in colorectal cancer (Andries et al., 2015; Liao et al., 2018; Shriver et al., 2015). Further analyses are required to assess the role of these specific genes in the development of the disease and the immune response, including studies of driving genes (important for the tumor survival), transcriptomic/proteomic profiling, epigenetic mechanisms, or other genomic variants.

Regarding the SNVs identification, we used VarScan2 to call variants (mutations). This software has been shown to have a high sensitivity (81–100%) in the variant calling analysis, in particular with cancer data (Spencer, Tyagi, et al., 2014; Xu, 2018). In addition, the use of a normal tissue sample to filter variants has been advised by several studies to reduce false positives (such as germline mutations) (Bae, Kim, & Kang, 2016; Y. Wang et al., 2018), as we did here. Confirmation strategies are required to validate identified variants (for example using Sanger technology to re-sequence genomic regions). Also, other approaches to call variants can be used to assess other DNA data, to deal with complexity, other sequencing technologies, and biological variability (Z. K. Liu, Shang, Chen, & Bian, 2017).

On the other hand, for HLA of the Costa Rican Central Valley population, information only had a two-digits resolution. Thus, we used the HLA supertypes (A*01, A*02, A*03, A*24 and B*07) or the representative allele of the group to predict interactions (four-digits resolution as required). The length of tumor-specific peptides was considered between 8–11 amino acids, which is the known peptide size that can be presented by HLA-I. It has been suggested that simultaneous testing of peptides between 8 and 11 amino acid residues is advisable in peptide prediction for putative cancer vaccines, due to the "core" or the common sequence (8 amino acids) may have greater versatility in different fragments (Luo et al., 2015). This allowed us to test several possible peptides of the same variant and to improve the search for possible neoantigens.

Considering the above, 23 candidate peptides were found to have affinity against the most prevalent Costa Rican Central Valley HLA I alleles (Table 1). Some peptides were identified to be presented by different alleles of the HLA, such as AYLHSPMYF (two different loci HLA A*2402 and HLA * C0401), however, most peptides are predicted to be presented in only one allele.

Regarding the number of peptides and haplotypes, the HLA-A locus was the most significant, contributing to most of the high-affinity peptides for the antigenic presentation (Fig. 3 and Table 1). For example, up to nine peptides can be presented in the case of haplotype A*2402/B*3501, in which the first allele is responsible for the presentation of nine peptides.

We also analyzed the physicochemical nature of the peptides (Table 1). Many of them are acidic peptides and, based on the hydrophobicity index, most of them have a slight hydrophobicity. Some “core” peptides are found in several peptides, such as AYLHSPMY which is present as part of seven candidate peptides of different HLA alleles (HLA-A*24:02, B*35:01, and C*04:01).

In the context of the selection of peptides, the algorithms use metrics such as IC50 value (concentration necessary to displace 50% of the classical peptide bound to an HLA allele) (Giannakis et al., 2016; Karasaki et al., 2017). However, it has been shown that it can yield a greater number of false positives in the detection of neopeptides, since some alleles can bind a significant percentage of random peptides (Nielsen & Andreatta, 2016). This motivated the present investigation to use only peptides with strong affinities (rank < 0.5%).

Other biologic factors should be considered to identify possible neoantigens and not included in our study. For example, protein degradation is possible at the endoplasmic reticulum (Fleri et al., 2017). In addition, peptides affinity must be validated using *in vitro* assays with antigen-presenting cells of the Costa Rican population, for example, the Enzyme-linked Immunospot Assay (ELISPOT) (Kato et al., 2018).

# Conclusion

In summary, our approach was able to find 23 candidate neoantigens in all samples of colorectal cancer to eventually create a vaccine. The contribution of this study is the development of the first bioinformatic

protocol to detect tumor-specific antigens in the context of the HLA alleles of the Costa Rican population. We hope that this study establishes a useful basis for future studies with therapeutic uses for cancer vaccines against colorectal tumors or other cancer types.

# Declarations

### Author contributions

D.M.F. and J.A.M.M. participated in the conception and design of the study. J.A.M.M. implemented and standardized the bioinformatics pipelines. D.M.F. processed all data using the pipelines. D.M.F., J.A.M.M., and J.M. participated in the interpretation of the results. D.M.F. drafted the manuscript. All authors reviewed and approved the final manuscript.

### Declaration of Competing Interest

The authors declare that there is no conflict of interest.

### Material availability

Processed data is found in the Supplementary material.

### Funding

This work was funded by Vicerrectoría de Investigación – Universidad de Costa Rica, with the Project "C1163 proNGS 2.0: Protocolos operativos estandarizados de análisis de datos moleculares obtenidos por NGS o afines y de algoritmos de inteligencia artificial en modelos biológicos (2021-2023)".

# References

1. Andries, V., Vandepoele, K., Staes, K., Berx, G., Bogaert, P., Isterdael, G., … Roy, F. (2015). NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. BMC Cancer, *15*(1). https://doi.org/10.1186/S12885-015-1408-5
2. Arrieta-Bolaños, E., Maldonado-Torres, H., Dimitriu, O., Hoddinott, M. A., Fowles, F., Shah, A., … Madrigal, J. A. (2011). HLA-A, -B, -C, -DQB1, and -DRB1,3,4,5 allele and haplotype frequencies in the Costa Rica Central Valley Population and its relationship to worldwide populations. *Human Immunology*, *72*(1), 80–86. https://doi.org/10.1016/J.HUMIMM.2010.10.005
3. Bae, J. M., Kim, J. H., & Kang, G. H. (2016). Molecular Subtypes of Colorectal Cancer and Their Clinicopathologic Features, With an Emphasis on the Serrated Neoplasia Pathway. Archives of Pathology & Laboratory Medicine, *140*(5), 406–412. https://doi.org/10.5858/ARPA.2015-0310-RA
4. Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., & Team, G. (2010). Manipulation of FASTQ data with galaxy. Bioinformatics, *26*(14), 1783–1785. https://doi.org/10.1093/bioinformatics/btq281

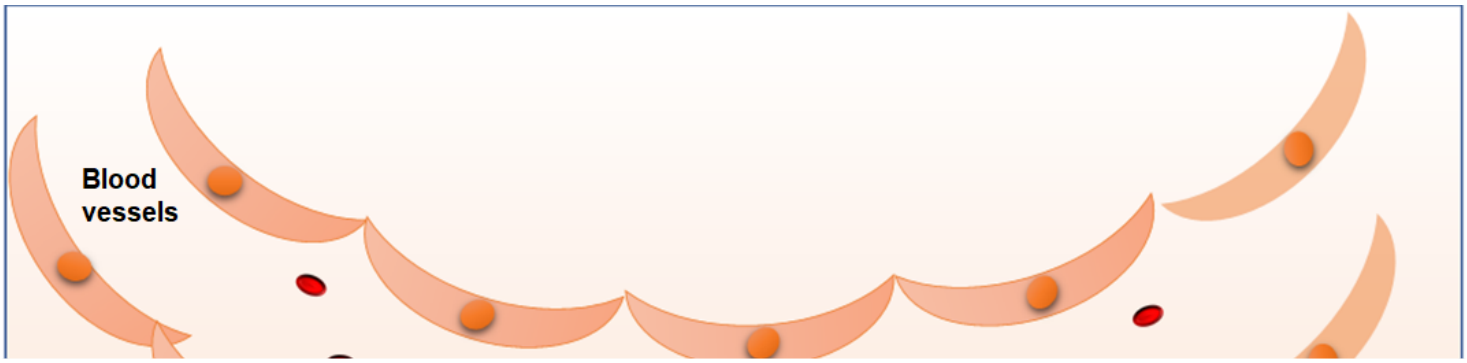
5. Chisanga, D., Keerthikumar, S., Pathan, M., Ariyaratne, D., Kalra, H., Boukouris, S., … Mathivanan, S. (2016). Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. Nucleic Acids Research, *44*(D1), D969–D974. https://doi.org/10.1093/NAR/GKV1097
6. Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., & Ruden, D. M. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics, *3*(315), 1–9. https://doi.org/10.3389/fgene.2012.00035
7. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, *6*(2), 80–92. https://doi.org/10.4161/fly.19695
8. Ding, L., Wendl, M. C., McMichael, J. F., & Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. Nature Reviews Genetics, *15*(8), 556–570. https://doi.org/10.1038/nrg3767
9. Emens, L. A., Ascierto, P. A., Darcy, P. K., Demaria, S., Eggermont, A. M. M., Redmond, W. L., … Marincola, F. M. (2017). Cancer immunotherapy: Opportunities and challenges in the rapidly evolving clinical landscape. European Journal of Cancer (Oxford, England: 1990), *81*, 116–129. https://doi.org/10.1016/J.EJCA.2017.01.035
10. Fleri, W., Paul, S., Dhanda, S. K., Mahajan, S., Xu, X., Peters, B., & Sette, A. (2017). The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. Frontiers in Immunology, *8*(MAR). https://doi.org/10.3389/FIMMU.2017.00278
11. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Wlaker (Ed.), *The Proteomics Protocols Handbook* (pp. 571–608). https://doi.org/10.1385/1592598900
12. Giannakis, M., Mu, X. J., Shukla, S. A., Qian, Z. R., Cohen, O., Nishihara, R., … Garraway, L. A. (2016). Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Reports, *15*(4), 857–865. https://doi.org/10.1016/J.CELREP.2016.03.075
13. Hackl, H., Charoentong, P., Finotello, F., & Trajanoski, Z. (2016). Computational genomics tools for dissecting tumour–immune cell interactions. Nature Reviews Genetics, *17*, 441–458. https://doi.org/10.1038/nrg.2016.67
14. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, *107*(1), 1. https://doi.org/10.1016/J.YGENO.2015.11.003
15. Karasaki, T., Nagayama, K., Kuwano, H., Nitadori, J. ichi, Sato, M., Anraku, M., … Nakajima, J. (2017). An Immunogram for the Cancer-Immunity Cycle: Towards Personalized Immunotherapy of Lung Cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer, *12*(5), 791–803. https://doi.org/10.1016/J.JTHO.2017.01.005
16. Kato, T., Matsuda, T., Ikeda, Y., Park, J. H., Leisegang, M., Yoshimura, S., … Nakamura, Y. (2018). Effective screening of T cells recognizing neoantigens and construction of T-cell receptor-engineered T cells. Oncotarget, *9*(13), 11009. https://doi.org/10.18632/ONCOTARGET.24232

17. Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, *22*, 568–576. https://doi.org/10.1101/gr.129684.111.568
18. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
19. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
20. Liao, W., Li, G., You, Y., Wan, H., Wu, Q., Wang, C., & Lv, N. (2018). Antitumor activity of Notch-1 inhibition in human colorectal carcinoma cells. Oncology Reports, *39*(3), 1063. https://doi.org/10.3892/OR.2017.6176
21. Liu, X. S., & Mardis, E. R. (2017). Applications of Immunogenomics to Cancer. Cell, *168*(4), 600–612. https://doi.org/10.1016/j.cell.2017.01.014
22. Liu, Z. K., Shang, Y. K., Chen, Z. N., & Bian, H. (2017). A three-caller pipeline for variant analysis of cancer whole-exome sequencing data. Molecular Medicine Reports, *15*(5), 2489–2494. https://doi.org/10.3892/MMR.2017.6336
23. Luo, H., Ye, H., Ng, H. W., Shi, L., Tong, W., Mattes, W., … Hong, H. (2015). Understanding and predicting binding between human leukocyte antigens (HLAs) and peptides by network analysis. BMC Bioinformatics, *16*(13), 1–9. https://doi.org/10.1186/1471-2105-16-S13-S9/FIGURES/5
24. Mette, A., Morten, B., & Hadrup, S. R. (2017). MuPeXI: prediction of neo – epitopes from tumor sequencing data. Cancer Immunology, Immunotherapy, *66*(9), 1123–1130. https://doi.org/10.1007/s00262-017-2001-3
25. Nielsen, M., & Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Medicine, *8*(1), 1–9. https://doi.org/10.1186/s13073-016-0288-x
26. Palucka, A. K., & Coussens, L. M. (2016). The Basis of Oncoimmunology. Cell, *164*(6), 1233–1247. https://doi.org/10.1016/j.cell.2016.01.049
27. Rizzo, J. M., & Buck, M. J. (2012). Key principles and clinical applications of "next-generation";DNA sequencing. *Cancer Prevention Research (Philadelphia, Pa.)*, *5*(7), 887–900. https://doi.org/10.1158/1940-6207.CAPR-11-0432
28. Shriver, M., Stroka, K. M., Vitolo, M. I., Martin, S., Huso, D. L., Konstantopoulos, K., & Kontrogianni-Konstantopoulos, A. (2015). Loss of giant obscurins from breast epithelium promotes epithelial-to-mesenchymal transition, tumorigenicity and metastasis. Oncogene, *34*(32), 4248–4259. https://doi.org/10.1038/ONC.2014.358
29. Spencer, D. H., Tyagi, M., Vallania, F., Bredemeyer, A. J., Pfeifer, J. D., Mitra, R. D., & Duncavage, E. J. (2014). Performance of common analysis methods for detecting low-frequency single nucleotide

variants in targeted next-generation sequence data. The Journal of Molecular Diagnostics: JMD, *16*(1), 75–88. https://doi.org/10.1016/J.JMOLDX.2013.09.003

30. Spencer, D. H., Zhang, B., & Pfeifer, J. (2014). Single Nucleotide Variant Detection Using Next Generation Sequencing. In J. Kulkarni, S.; Pfeifer (Ed.), *Clinical Genomics* (pp. 109–127). https://doi.org/10.1016/B978-0-12-404748-8.00008-3
31. Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., & Shi, L. (2011). Next-generation sequencing and its application in molecular diagnostics. Expert Review of Molecular Diagnostics, *11*(3), 1–16. https://doi.org/http://dx.doi.org/10.1586/erm.11.3
32. Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, *38*(16), 1–7. https://doi.org/10.1093/nar/gkq603
33. Wang, Y., Liping, G. U. O., Feng, L., Zhang, W., Xiao, T., Xuebing, D. I., … Zhang, K. (2018). Single nucleotide variant profiles of viable single circulating tumour cells reveal CTC behaviours in breast cancer. Oncology Reports, *39*(5), 2147–2159. https://doi.org/10.3892/OR.2018.6325
34. World Health Organization. (2020). Colorectal cancer statistics. Retrieved April 25, 2022, from https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf
35. Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal, *16*, 15–24. https://doi.org/10.1016/J.CSBJ.2018.01.003

# Figures




## Figure 1

**Conceptualization of the biological meaning of the neoantigens identification in cancer vaccines**. The tumor-specific antigens (neoantigens) are originated from variants (mutations) in the genes of tumor cells, which are selected as candidate peptides for the vaccine. Peptides that are predicted to be presented by MCH type I molecules to cytotoxic CD8+ cells (HLA-antigen-TCR), are kept for the subsequent analyses. If the antigen is an endogenous antigen (non-modified and also present in normal cells), it is excluded. The eventual signaling in the immune response will induce death of tumor cells. Own elaboration.

## Figure 2

**Bioinformatic pipeline to identify neoantigens in primary and metastatic tumor samples of colorectal cancer.** The protocol includes three main steps: pre-processing, variant calling and neoantigens (peptides)

identification.

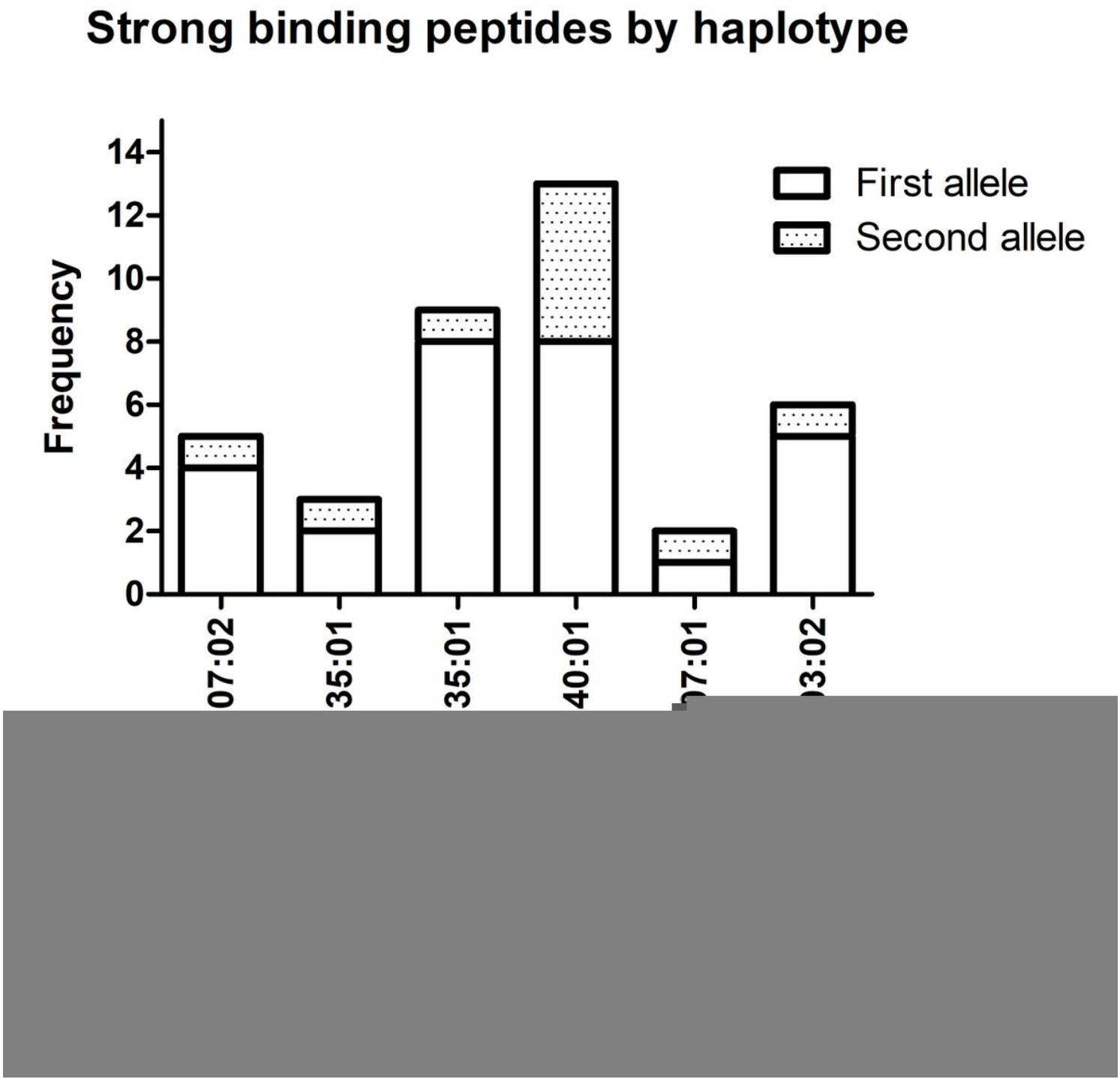



**Figure 3**

**Number of strong binding peptides for different HLA I haplotypes from Costa Rican population.** Peptides are derived from the selected SNVs in the primary and metastatic tumor samples of the study.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile.xlsx