Université de Montréal

Proteogenomic analyses of colorectal cancer reveal tumor-specific antigens across microsatellite status

By

Jenna Cleyle

Institute of Research in Immunology and Cancer, Department of Molecular Biology

Thesis submitted for obtaining the rank of Master of Science in Molecular Biology, Systems Biology option

January 2022

Université de Montréal

Institute of Research in Immunology and Cancer, Department of Molecular Biology

---

This thesis entitled

Proteogenomic analyses of colorectal cancer reveal tumor-specific antigens across microsatellite status

Presented by

Jenna Cleyle

Has been evaluated by a jury composed of the following people:

Réjean Lapointe

Président-rapporteur

Pierre Thibault

Directeur de recherche

Moutih Rafei

Membre du jury

Résumé

Le cancer colorectal est la deuxième cause de décès par cancer au monde. Il a été démontré que la thérapie par inhibition du point de contrôle immunitaire traite efficacement le cancer colorectal avec microsatellites instables, mais la majorité des tumeurs ne répondent pas bien à ce traitement. La recherche s'est ainsi tournée vers des stratégies immunothérapeutiques, qui pourraient activer une réponse des lymphocytes T cytotoxiques contre des antigènes spécifiques aux tumeurs. Ces antigènes sont mieux identifiés en utilisant la spectrométrie de masse, qui permet l'échantillonnage et le séquençage directs de ces peptides. Alors que les quelques antigènes spécifiques aux tumeurs identifiés à date sont dérivés de régions codantes du génome, des découvertes récentes indiquent qu'une grande proportion d'antigènes spécifiques aux tumeurs proviennent de régions prétendument non codantes. Ici, nous avons utilisé une nouvelle approche protéogénomique pour identifier les antigènes tumoraux dans une collection de lignées cellulaires dérivées du cancer colorectal et d'échantillons de biopsie. L'utilisation de bases de données personnalisées sur le cancer en tandem avec des analyses de spectrométrie de masse a permis d'identifier plus de 30 000 peptides uniques associés au CMH I et 19 antigènes spécifiques aux tumeurs dans des tumeurs avec microsatellites stables et instables, dont plus de deux tiers provenaient de régions non codantes. Ces découvertes pourraient bénéficier le développement de vaccins à base de cellules T, dans lesquels les cellules T sont amorcées contre ces antigènes pour cibler et éradiquer les tumeurs. Un tel vaccin pourrait être utilisé avec les thérapies existantes d'inhibition des points de contrôle immunitaire, pour traiter efficacement divers sous-types de cancer colorectal avec des pronostics différents. Les études futures devraient inclure une évaluation rigoureuse de l'immunogénicité de ces peptides, ainsi que l'optimisation des formulations spécifiques de vaccins anticancéreux pour traiter le plus efficacement possible le cancer colorectal.

Mots-clés : cancer colorectal, antigène spécifique aux tumeurs, spectrométrie de masse, protéogénomique, immunothérapie du cancer

Abstract

Colorectal cancer is the second leading cause of cancer death worldwide. Immune checkpoint inhibition therapy has been shown to effectively treat microsatellite unstable colorectal cancer, but the majority of tumors do not respond well to this treatment. Research has thus turned to immunotherapeutic strategies, which could activate a cytotoxic T cell response against tumor-specific antigens. Such antigens are best identified using mass spectrometry, which allows the direct sampling and sequencing of these peptides. While the few tumor-specific antigens identified to date are derived from coding regions of the genome, recent findings indicate that a large proportion of tumor-specific antigens originate from allegedly noncoding regions. Here, we employed a novel proteogenomic approach to identify tumor antigens in a collection of colorectal cancer-derived cell lines and biopsy samples. Using personalized cancer databases in tandem with mass spectrometry analyses resulted in the identification of over 30 000 unique MHC I-associated peptides and 19 tumor-specific antigens in both microsatellite stable and unstable tumors, over two-thirds of which were derived from non-coding regions. These findings could benefit the development of T cell-based vaccines, in which T cells are primed against these antigens to target and eradicate tumors. Such a vaccine could be used with existing immune checkpoint inhibition therapies, to effectively treat varying subtypes of colorectal cancer with differing prognoses. Future studies should include rigorous evaluation of the immunogenicity of these peptides, as well as the optimization of the specific cancer vaccine formulations to most effectively treat colorectal cancer.

Keywords : colorectal cancer, tumor specific antigen, mass spectrometry, proteogenomics, cancer immunotherapy

Table of contents

Résumé 3

Abstract 5

Table of contents 7

Table list 11

Figure list 13

Acknowledgements 19

Chapter 1 – Introduction 21

1.1 Colorectal cancer 21

1.1.1 Colorectal cancer epidemiology 21

1.1.2 Colorectal cancer diagnoses and treatments 23

1.1.3 Colorectal cancer tumorigenesis 23

1.1.4 Molecular subtypes of colorectal cancer 25

1.1.4.1 Chromosomal instability 25

1.1.4.2 Microsatellite instability 25

1.1.4.3 CpG island methylator phenotype 26

1.1.4.4 Familial CRC 27

1.2 Immune system 28

1.2.1 Innate immune system 28

1.2.2 Adaptive immune system 28

1.2.3 MHC I peptide presentation 31

1.2.4 T cell receptor rearrangement 34

1.2.5 T cell selection 34

1.2.6 T cell activation 35
1.2.7 Immune checkpoints 35
1.2.8 Immune system and cancer 36
1.2.9 Tumor microenvironment 36
1.2.10 Adaptive immune response to cancer 38
1.2.11 Immune checkpoint inhibition 39
1.3 Analyzing MAPs by mass spectrometry 40
1.3.1 Peptide quantification 41
1.3.2 Databases for mass spectrometry analyses 42
1.4 Identification of tumor antigens 44
1.4.1 TAAs 44
1.4.2 Mutated TSAs 45
1.4.3 Aberrantly expressed TSAs 46
1.4.4 TSAs in CRC 47
1.5 Research Objectives 49
1.6 Thesis Overview 50
Chapter 2: Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability 51
2.1 Abstract 52
2.2 Introduction 53
2.3 Experimental Procedures 55
2.3.1 Samples 55
2.3.1.1 Cell lines 55
2.3.1.2 Primary tissues 55

2.3.2 RNA extraction and sequencing 56
2.3.2.1 RNA extraction 56
2.3.2.2 RNA sequencing 56
2.3.2.3 Bioinformatic analyses 56
2.3.3 Transcriptomics 57
2.3.3.1 HLA genotyping 57
2.3.3.2 Microsatellite instability detection 57
2.3.3.3 Differential expression analysis 57
2.3.3.4 Transcriptome analysis of tissue samples 58
2.3.3.5 Mutation profiles and genetic variant annotation 58
2.3.4 Database generation 58
2.3.5 Isolation of MAPs 59
2.3.6 TMT labeling 59
2.3.7 Mass spectrometry analyses 60
2.3.8 MAP identification 60
2.3.8.1 MAP source gene analysis 61
2.3.9 Quantification of MAP coding sequences in RNA-Seq data 61
2.3.10 Determination of MAP source transcripts 61
2.3.11 Identification of TSA candidates 62
2.3.11.1 Intertumoral sharing 63
2.3.11.2 Immunogenicity prediction 63
2.3.12 TSA validation and relative quantification with synthetic peptides 63
2.3.12.1 Validation of TSA peptide candidates 63
2.3.12.2 Relative quantification 64

2.3.13 Data analysis and visualization 65
2.3.14 Experimental Design and Statistical Rationale 65
2.4 Results 67
2.4.1 Immunopeptidomic analyses using a proteogenomic approach 67
2.4.2 Transcriptomic analyses reveal heterogeneity between MSI and MSS samples 75
2.4.3 Immunopeptidomic analyses highlight the diversity of CRC antigens 81
2.4.4 Identification of tumor-specific and tumor-associated antigens in CRC 88
2.4.5 RNA expression of putative tumor-specific and tumor-associated antigens 95
2.4.6 Cancer specificity and predicted immunogenicity of TSAs and TAAs 99
2.5 Discussion 106
2.6 Acknowledgments 111
2.7 Data Availability 112
Chapter 3 – Conclusion and Perspectives 113
3.1 Conclusion 113
3.2 Perspectives 114
3.2.1 Proteogenomics approach 114
3.2.2 Immunogenicity and T cell reactivity 115
3.2.3 Cancer vaccines 116
3.2.4 Remaining questions 119
Bibliography 123

Table list

Table 1. – Description of CRC-derived cell lines 72
Table 2. – Description of primary tumor and matched NAT 73
Table 3. – MSISensor-pro results for CRC primary tissues..... 75
Table 4. – Biological relevance of TSA source genes in CRC 91
Table 5. – Biological relevance of TAA source genes in CRC..... 93
Table 6. – Justifications for exclusion of Löffler et al. 2018 tumor antigens 95
Table 7. – Relative quantification ratios of validated tumor antigens in CRC 101

Figure list

Figure 1. – Adenoma-carcinoma sequence of colorectal cancer. 24
Figure 2. – MHC structures. 30
Figure 3. – T cell receptor loci 31
Figure 4. – MHC Class I peptide presentation pathway 33
Figure 5. – Tumor microenvironment. 38
Figure 6. – Schematic of mass spectrometry analysis. 41
Figure 7. – Standard mTSA/neoantigen identification 46
Figure 8. – Types of tumor antigens 48
Figure 9. – Proteogenomic workflow for the discovery of tumor-specific antigens (TSAs) in both colorectal cancer (CRC)-derived cell lines and primary tumor samples 68
Figure 10. – TA identification flowchart 70
Figure 11. – Upset plot of HLA alleles 74
Figure 12. – Transcriptomic profile of primary tumor/normal adjacent tissue CRC biopsies 77
Figure 13. – Transcriptomic profile of CRC-derived cell lines and GO term analysis of MSI and MSS primary tissue samples 78
Figure 14. – ssGSEA analysis of immune infiltration in CRC tissues and mutation profile of all samples 80
Figure 15. – Immunopeptidomics of CRC-derived cell lines and tissues 82
Figure 16. – Overview of unique and shared MAPs in CRC-derived cell line and CRC/NAT tissue samples 84
Figure 17. – Overview of unique and shared MAP source genes in CRC-derived cell line and CRC/NAT tissue samples 86
Figure 18. – Novel TSAs identified in CRC derive primarily from non-coding regions, while the majority of TAAs derive from exons 89
Figure 19. – Correlation of MAPs and TSAs 90
Figure 20. – RNA expression profiles of putative TSAs and TAAs 96
Figure 21. – RNA expression of MCS in cancer and NAT 98

Figure 22. – Validation of TSAs and TAAs 103
Figure 23. – Predicted immunogenicity of TSAs and TAAs 104

List of acronyms and abbreviations

aeTSA: aberrantly expressed tumor-specific antigen

AGC: automatic gain control

AML: acute myeloid leukemia

ATCC: American Type Culture Collection

BCR: B cell receptor

CEA: carcinoembryonic antigen

CIMP: CpG island methylator phenotype

CIN: chromosomal instability

COAD: colon adenocarcinoma

CRC: colorectal cancer

CTA: cancer testis antigen

CTLA-4: cytotoxic T-lymphocyte-associated protein 4

DC: dendritic cell

DCC: Deleted in colorectal carcinoma

DEG: differentially expressed gene

DP: double positive

DRIP: defective ribosomal product

EGF: epidermal growth factor

ER: endoplasmic reticulum

ERE: endogenous retroviral element

FAP: familial adenomatous polyposis

FBS: fetal bovine serum

FDR: false discovery rate

GO : gene ontology

GTEx: Genotype Tissue Expression project

HDI: human development index

HLA: human leukocyte antigen

ICI : immune checkpoint inhibition

IEDB : Immune Epitope Database

IFN-γ: interferon gamma

INDEL: insertion/deletion

KPHM: kmers-per-hundred-million

LC-MS/MS: liquid chromatography tandem mass spectrometry

lncRNA: long noncoding RNA

LOH: loss of heterozygosity

MAP: MHC I-associated peptide

MAPK: mitogen-activated protein kinase

MCS: MAP-coding sequence

MHC: major histocompatibility complex

MMR: mismatch repair (DNA mismatch repair pathway)

MS: mass spectrometry

MS/MS: tandem mass spectrometry

MSI: microsatellite instability

MSS: microsatellite stable

mTEC: medullary thymic epithelial cell

mTSA: mutated tumor-specific antigen

NAT: normal adjacent tissue

NK: natural killer

PBS: phosphate buffered saline

PCA: principal component analysis

PCR: polymerase chain reaction

PD-1: programmed death 1

PD-L: programmed death ligand

PSM: peptide spectrum match

PTM: post-translational modification

RNA-seq: RNA sequencing

RPHM: reads-per-hundred-million

SNP: single nucleotide polymorphism

SNV: single nucleotide variant

SPS-MS3: synchronous precursor selection MS3

ssGSEA: single-sample Gene Set Enrichment Analysis

TAA: tumor associated antigen

TCGA: The Cancer Genome Atlas

TCR: T cell receptor

TEC: thymic epithelial cell

TIL: tumor-infiltrating lymphocyte

TME: tumor microenvironment

TMT: tandem mass tag

TPM: transcripts per million

TSA: tumor-specific antigen

UTR: untranslated region

Acknowledgements

I would like to thank my supervisor Pierre Thibault for his guidance and support throughout my degree. I would also like to thank the members of the Thibault and Perreault labs, who provided invaluable contributions to this work (scientific or otherwise). Finally, I would like to thank my parents, for always reminding me that I could do anything.

Chapter 1 – Introduction

1.1 Colorectal cancer

Cancer is defined as uncontrolled cell growth that can develop in nearly any tissue or organ in the body (1). Such growth can disrupt the regular functioning of these tissues or organs and lead to grave health consequences. In addition, cancer cells have the ability to migrate, or metastasize, spreading throughout the body and causing further damage. Colorectal cancers (CRC) are tumors, primarily adenocarcinomas, which develop in the colon or rectum. They represent a large and globally increasing disease burden. While CRC incidence and mortality are decreasing in certain countries with evolving treatments and screening methods, many factors need to be addressed. Different molecular subtypes have varying responses to treatments and much remains to be understood about the intricacies of this illness.

1.1.1 Colorectal cancer epidemiology

Colorectal cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death worldwide, with over 1.9 million cases and 935 000 deaths estimated in 2020 alone (2). This represents approximately 10% of all cancer deaths. The incidence of CRC is expected to increase as global socioeconomic changes occur, with a predicted 2.2 million cases and 1.1 million deaths occurring annually by 2030 (2, 3). In Canada, the predicted incidence in 2020 was 26 900 cases and 9700 deaths (4). Annual cases of CRC in Canada began declining in males and females in the year 2000, however there remains a higher incidence and mortality in males, a trend that is reproduced globally (2).

The incidence of CRC is correlated with high human development indices (HDI), as many risk factors of CRC are associated with a “Western” lifestyle (2). Such risk factors include alcohol consumption, smoking, decreased physical activity, and eating red or processed meats. “Obesity” is often listed as a risk factor; however, the stigma and bias directed towards people who are considered “obese” has a profound effect on patients’ mental health, well-being, and access to proper medical care (5, 6). This bias in turn could lead to delay in diagnoses, resulting in disease

progression and worse clinical outcomes. In contrast, healthier lifestyle choices decrease the risk of CRC, as demonstrated by the decrease in incidence in North America in recent years (2, 7).

In addition to lifestyle changes and disease prevention, CRC mortality has also decreased due to screening efforts. As CRC has a slow progression from precancerous lesions to more advanced stages of disease (8), and early diagnoses are positively correlated with improved prognoses (7), widespread screening efforts are paramount. Both invasive (colonoscopy, sigmoidoscopy) and non-invasive (blood test, stool test) measures exist, each with advantages and with guidelines varying by country. For example, colonoscopies are the primary screening method used in the United States and are considered the "gold standard of screening" and were shown to reduce mortality in CRC overall; however, they did not reduce mortality for cancers of the proximal colon. In addition, this method is quite expensive and is thus not widely available (9). Other non-invasive and more affordable treatments, such as the fecal occult blood test, are simpler and widely available, however they are lower in sensitivity. Advances in artificial intelligence are also proving advantageous for CRC detection, diagnosis, and treatment, such as deep learning models used to analyze screening images, which are frequently shown to improve diagnostic accuracy (10).

Despite the variety of tests available, roughly 55% of Canadian adults in 2012 and less than 70% of American adults in 2018 were up to date with CRC screening (11, 12). The United States has established screening targets in response, with the Office of Disease Prevention and Health Promotion setting the goal of 74.4% of eligible adults being screened by 2030 (https://health.gov/healthypeople/objectives-and-data/, objective C-07). A retrospective study of the age of diagnosis of CRC in the United States demonstrated an increase in diagnoses in individuals under 50 from 2004 to 2015. Accordingly, American recommendations adapted to suggest screening begin at 45 years of age (13, 14).

While the overall incidence and mortality rates of CRC are declining in the United States, these advances are not seen equally across all populations. The incidence rates of CRC were approximately 20% higher in Black individuals compared to non-Hispanic white individuals in 2012 to 2016, while mortality was nearly 40% higher (7). While approximately half of these disparities

are attributable to differences in risk factor prevalence, Black individuals are also less likely to receive follow-up care (15). Additionally, Alaska Natives have the highest CRC incidence and mortality in the United States, which is at least partially due to inadequate availability of screening services (7). In addition to differences in risk factor prevalence and screening and follow-up access, systemic racism in healthcare has been well documented in both Canada and the United States and cannot be ignored in discussions of incidence and mortality among marginalized groups (16, 17). To increase the impact of both cancer screening and treatments it will also be necessary to address racism and other forms of discrimination enacted by healthcare providers and institutions.

1.1.2 Colorectal cancer diagnoses and treatments

Colorectal cancer typically presents with initial symptoms such as rectal bleeding, pain, or alterations in bowel habits (18). Upon diagnosis, tumors are staged with a combination of CT, MRI, or endoscopic ultrasound. They may also be analyzed for various biomarkers or molecular phenotypes which could influence the course of treatment. To treat colorectal cancer, patients typically undergo surgical resection and chemotherapy or radiation, both of which depend on the type and location of the tumor. In more recent years, immune checkpoint inhibition (ICI) has become an attractive treatment option, in which small molecule inhibitors mitigate the cancer's inhibitory effects on the anti-tumor immune response.

1.1.3 Colorectal cancer tumorigenesis

The proposed adenoma-carcinoma sequence (Figure 1) of CRC describes the transition from a normal epithelium to a benign adenoma, followed by the progression to *in situ* carcinoma and finally an invasive and metastatic tumor (19, 20). An adenoma is a type of polyp, or abnormal tissue growth, and can be identified in cancer screenings. While not all adenomas will develop into cancer, approximately 90% of sporadic CRC cases begin with adenomas, and as such colon polyps are removed if possible (21). Cancer progression can also be measured through TNM staging, which evaluates the invasion of the primary tumor both locally and in neighboring lymph nodes, as well as whether any distant metastases are present. In the case of CRC, an *in situ* carcinoma progresses to stage I once it has invaded the submucosa, and left untreated it will

continue to invade the muscular propria (stage II), the pericolorectal tissues (stage III), and finally it could invade the visceral peritoneum or other structures or organs (stage IV)(22). Stage I and II tumors are typically treated with surgery alone, whereas later stages are often treated with chemotherapy (23).

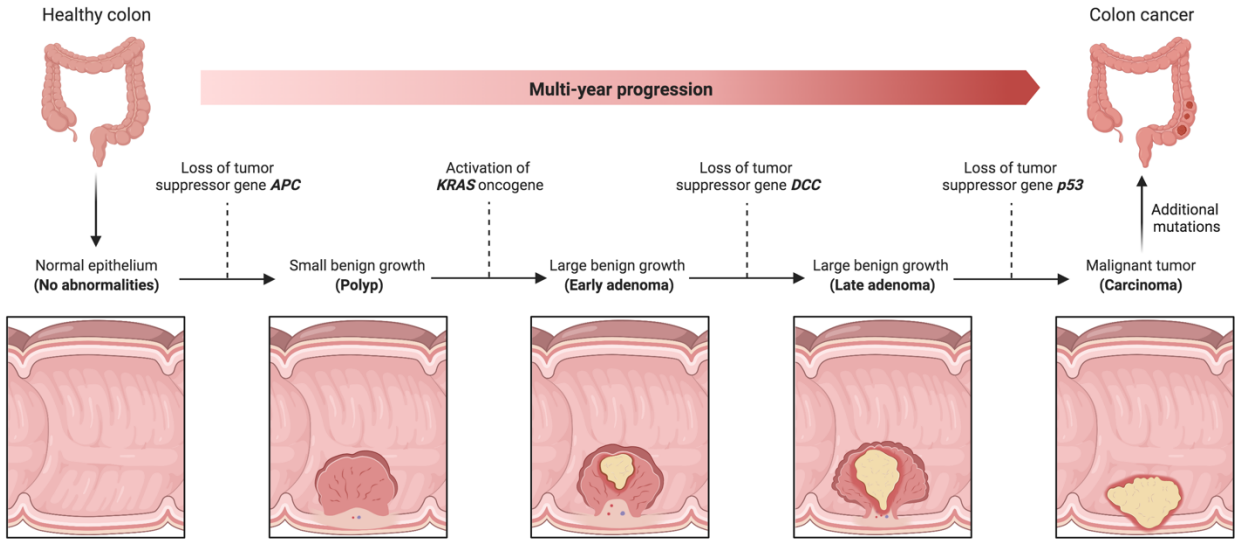

Figure 1. – Adenoma-carcinoma sequence of colorectal cancer.

Depiction of the multi-year progression from healthy colon to colon cancer, through the successive loss of tumor suppressor genes and activation of oncogenes. Through these stages, normal colon epithelium develops first into a polyp, which then progresses to various stages of adenomas, finally resulting in a colon carcinoma. Adapted from "The Multi-Hit Model of Colorectal Cancer", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

Colonic epithelial stem cells employ multiple strategies to maintain their genomic integrity (19). First, these cells rarely replicate, and when they do, they allocate DNA strands in an asymmetric fashion such that the older strand is allocated to the progenitor stem cell, as this strand likely contains fewer errors. The new strand is allocated to a transit-amplifying cell, which will amplify and then be sloughed off after 5-7 days. In addition to their replication strategies, colonic epithelial stem cells are located deep in the crypts away from environmental toxins and exposure to mutagens, and undergo cell death if damaged, rather than attempting DNA repair.

Finally, these cells also express Mdr1 (multidrug resistance 1) at the cell surface, which is able to pump out any mutagens that are able to enter the cell (and conversely is a source of chemo-resistance).

1.1.4 Molecular subtypes of colorectal cancer

1.1.4.1 Chromosomal instability

In spite of these protective mechanisms, several sources of genomic instability contribute to colorectal cancer development. The first, known as the chromosomal instability phenotype (CIN), follows a predictable series of KRAS activation and the inactivation of three or more tumor suppressor genes, which are often APC, p53, and the loss of heterozygosity (LOH) of the long arm of chromosome 18. APC, or the adenomatous polyposis coli gene, is the most frequent initial mutation in CRC, with 34-70% of sporadic CRC cases possessing a mutation in this gene (24). As part of the Wnt signaling pathway, this gene is responsible for blocking cell cycle progression, specifically G1/S transition, and this pathway keeps stem cells in an undifferentiated state. The Wnt pathway is often dysregulated in CRC, through APC or β-catenin mutations or promoter hypermethylation of the APC gene. This results in the accumulation of β-catenin and the retention of stem cells at the surface of crypts (rather than their elimination), and this accumulation of undifferentiated cells results in polyp formation (19). The subsequent addition of further mutations can thus lead to carcinoma development. Additionally, TP53 mutations prevent the G1 cell cycle arrest and DNA repair normally executed by this gene. TP53 is one of the most frequently mutated genes in human cancers, with approximately half of CRC tumors carrying a mutation in this gene (25, 26). The chromosome 18 LOH primarily involves the deletion of the Deleted in colorectal carcinoma (DCC) gene, resulting in abnormal cell survival. This LOH is observed in approximately 70% of CRC.

1.1.4.2 Microsatellite instability

Second, the microsatellite instability (MSI) phenotype of CRC arises from the hypermethylation of the MLH1 promoter in 80% of sporadic cases, or the point mutation of a gene involved in mismatch repair (MMR). DNA mismatch repair is responsible for repairing insertion/deletions and mis-incorporated bases during DNA replication. Microsatellites are small

repeating sequences of DNA of approximately 1 to 6 base pairs, which are present in both coding and non-coding regions and make up approximately 3% of the human genome. Due to their repetitive nature, these sequences are frequently subject to slippage, which under normal conditions is repaired by MMR machinery. Deficiencies in this machinery thus result in elongation or shortening of these microsatellite sequences, which is a source of genomic instability and is termed microsatellite instability. In CRC, MSI is associated with improved prognosis, especially in stage II tumors, and is associated with an improved response to ICI compared to the microsatellite stable (MSS) tumors (27). MSI is present in 95% of the hereditary Lynch syndrome and 15-20% of sporadic CRC and is diagnosed by polymerase chain reaction (PCR) to amplify microsatellite loci, immunohistochemical staining of CRC tissues to detect expression of MMR proteins, or through more recently developed bioinformatic programs that determine microsatellite status from Next Generation Sequencing data (28, 29). MSI tumors arise more frequently in the proximal colon and are characterized by lymphocytic infiltration. Due to the defects in MMR that are normally responsible for repairing DNA damage, MSI tumors are more chemo-sensitive than their MSS counterparts.

1.1.4.3 CpG island methylator phenotype

Finally, a high proportion of hypermethylated genes, known as the CpG island methylator phenotype (CIMP), characterizes a subset of CRC tumors. This involves the covalent attachment of a methyl group to a cytosine in either a repetitive CG sequence or a CpG-rich area of the promoter region of a gene. These regions are normally unmethylated, but their methylation leads to gene silencing; this is thus a method of silencing tumor suppressor genes. CRCs with this phenotype often have the V600E hotspot mutation in BRAF, a member of the mitogen-activated protein kinase (MAPK) pathway. This pathway is involved in cell proliferation, angiogenesis, motility, and metastasis, and is frequently dysregulated in cancer. Driver mutations in KRAS, are very common and occur in roughly 40% of CRC tumors, and ERK has been shown to be overexpressed in CRC (30, 31). The PI3K/AKT pathway can also be activated by mutations in PIK3CA (which is mutated in over a quarter of CRC), PTEN (negative regulator), or TGFBRII, which is mutated in up to 90% of MSI CRC. Similarly to the MAPK pathway, this signaling sequence is involved in angiogenesis, metabolism, growth, and proliferation (32). In addition, this pathway is

initiated by insulin or other growth factors binding to the insulin receptor substrate. As such, insulin dysregulation due to illnesses such as diabetes or hyperinsulinemia predispose CRC due to the overabundance of insulin that activates the PI3K/AKT pathway (33).

1.1.4.4 Familial CRC

CRCs can be divided into inherited familial or sporadic disease. The most prevalent inherited familial diseases are Lynch syndrome (hereditary non-polyposis colorectal cancer) or familial adenomatous polyposis (FAP). Patients with Lynch syndrome carry a germline mutation in an MMR gene such as MLH1, MSH2, MSH6, or PMS2, or an EPCAM deletion, resulting in an 80% lifetime risk of CRC (as well as an increased risk of other cancers such as endometrial cancer) (19, 34, 35). Approximately 3-5% of CRC cases are caused by this syndrome. Germline mutations in the APC gene cause the autosomal dominant FAP disorder, which causes hundreds or thousands of polyps to develop in the patient's colon beginning on average at the age of 16 (19, 36). Without colectomy (bowel resection) intervention, CRC will develop in all patients with FAP (37).

1.2 Immune system

The immune system serves as the body's defense against illness, through a complex web of biological processes that distinguish self from non-self. It is comprised of two branches: the innate and adaptive immune systems. Innate immunity is fast acting and non-specific but does not have a memory component. In contrast, the adaptive immune system takes longer to act but carries the advantage of specificity and immunological memory. These two branches work together to protect the host from a variety of onslaughts including bacteria, viruses, parasites, physical objects, and even tumors.

1.2.1 Innate immune system

The innate immune response is enacted by a series of distinct but collaborating cell types. Most of these cells circulate in the blood until they are recruited to sites of infection. Monocytes (in the blood) and macrophages (in the tissues) are responsible for phagocytosis of pathogens, initiating inflammation, and recruiting granulocytes to the area of infection or insult. These granulocytes, the most important of which are neutrophils, also enact phagocytosis of extracellular pathogens. Dendritic cells are responsible for ingesting antigens and then migrating into the lymphoid tissues to play a role in the adaptive immune response. Natural killer (NK) cells are considered both innate and adaptive cells and arise from a lymphoid progenitor and have a strong cytotoxic function. These cells are thought to make up approximately 5-20% of the circulating lymphocytes in the body (38). Phagocytosis by innate immune cells typically involves the ingestion of the pathogen into a phagosome, which is then acidified and merges with lysosomes to form a phagolysosome, in which the contents of the lysosome are also released to kill the pathogen (39).

1.2.2 Adaptive immune system

Aside from NK cells, the adaptive immune system consists primary of T and B lymphocytes. While B lymphocytes are responsible for an antibody response and attack extracellular components, T cells evaluate intracellular components. Both cell types recognize antigens, either through antibody/immunoglobulin or B cell receptor (BCR) recognition or T cell receptor (TCR) recognition of antigens presented by the major histocompatibility complex (MHC). Each TCR consists of both alpha and beta polypeptide chains which have a variable region, constant region, and a

transmembrane region with a small cytoplasmic tail, and the two chains are connected through a disulfide bond (39).

MHC class I and class II have two corresponding classes of T cells in humans. These two types of MHC molecules have different structures and functions but are both highly polymorphic and play major roles in the activation of T cells. MHC class I is expressed on the surface of all nucleated cells and is responsible for presenting peptides to CD8+ T cells. It consists of three α domains and one β-2-microglobulin domain (Figure 2) (40). One of its α chains spans the cell membrane. In contrast, MHC class II is only expressed on professional antigen presenting cells, namely B and T cells, macrophages, and dendritic cells. MHC class II activates CD4+ T cells and has one α and one β chain, both of which span the membrane. The two outermost domains of these molecules form a peptide-binding groove, through which they interact with short amino acid sequences and present them to the appropriate T cells for activation. MHC II binds longer peptides of approximately 13-17 amino acids in length and interacts with and activates CD4+ T cells which are responsible for activating other effector cells in response to extracellular pathogens. On the other hand, MHC class I binds short peptides of 8-11 amino acids in length and activate CD8+ T cells which have a cytotoxic response against primarily intracellular pathogens. The CD4+ and CD8+ T cells are so named for the presence of the CD4 and CD8 coreceptors at their cell surface, which are required for MHC recognition. In either case, the MHC complex is unstable without a bound peptide and thus needs to be presenting an antigen to be present at the cell surface (40). The MHC I complex interacts with peptides through the anchor residues near their amino and carboxy termini. These peptides can be released from the MHC complex through acidic denaturation and can thus be isolated and analyzed.

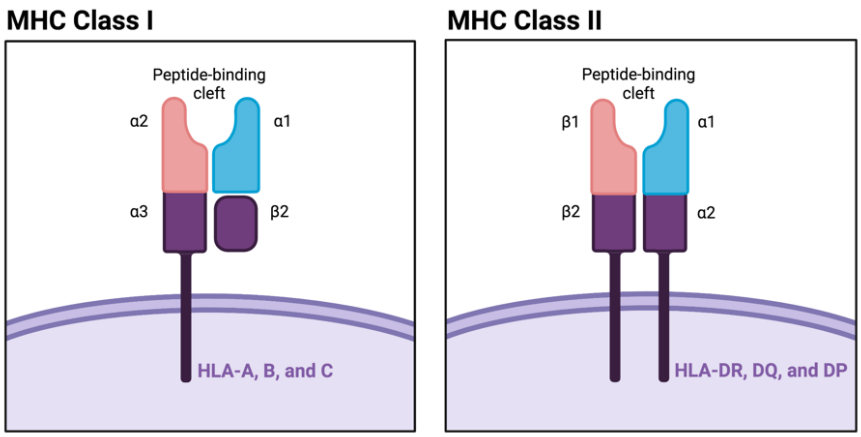

Figure 2. – MHC structures.

Structures of MHC class I (left) and MHC class II (right). The genes associated with these complexes are listed below. Adapted from "MHC Class I" and "MHC Class II", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

Given that the adaptive immune response needs to recognize and eliminate infections from a large array of pathogens which can consist of many different peptide sequences, TCRs need to be capable of recognizing a wide range of antigens. This is accomplished through somatic DNA recombination that occurs during T cell development. Briefly, the TCR is composed of variable (V), diversity (D), joining (J), and constant genes. While there are two distinct types of TCRs, αβ and γδ, αβ TCRs make up the majority and their role in the adaptive immune response is much better understood. During αβ T cell development, a Vα segment combines with a Jα segment, and this VJ region is then transcribed and spliced together to Cα, resulting in mRNA that is then translated to yield the TCRα chain protein (Figure 3) (39). In the β chain, the variable region is encoded by VDJ segments, which are generated and then spliced together with the Cβ gene. Since there are many possible V, D, and J segments, as well as various junctions to be made between them, this results in an impressive diversity of putative TCRs. For example, the V, D, and J components of the β chain alone have 52, 2, and 13 possible segments, respectively, which does not include the ability of the D segments to be read in multiple frames, or the possible junctional diversity that could occur. When combined with the diversity of the α chain, there is a presumed TCR diversity in the order of 10^{18} (39).

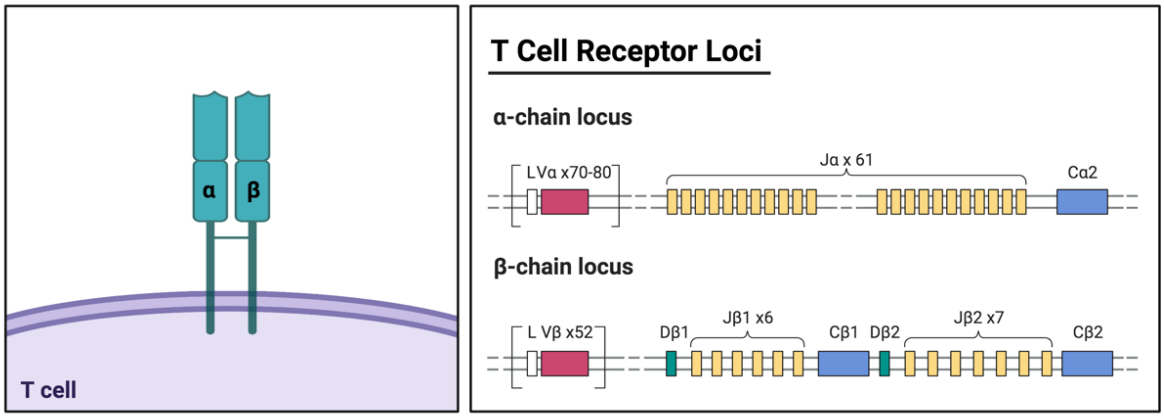

Figure 3. – T cell receptor loci

T cell receptor structure (left) and T cell receptor loci for the alpha and beta chains. The numbers of different segments for each loci are indicated above the segments. Both the alpha and beta chains undergo somatic DNA recombination, and the various combinations of the numerous variable (V), joining (J), and diversity (D) segments results in large TCR diversity. Adapted from "T Cell Receptor Loci", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

MHC complexes need to be able to recognize a wide variety of antigens. As such, different alleles of the genes that make up the MHC complex, which mainly occur in the peptide binding groove, result in various peptide binding repertoires. MHC is both polygenic (involves multiple genes) and polymorphic (has many possible variants). MHC class I in particular is encoded by three human leukocyte antigen (HLA) alleles, HLA-A, B, and C, which are codominant. Thus, each individual is able to express up to six different HLA alleles, which are highly polymorphic, and each have varying peptide-binding specificities. Because of this, an individual's ability to present a given peptide antigen is highly influenced by their HLA alleles.

1.2.3 MHC I peptide presentation

Protein homeostasis in the cell involves tightly regulating the processes of protein synthesis, transport, and degradation to maintain proteins in the correct concentration and localization in the cell. To achieve a balance in concentration, proteins often need to be degraded, which occurs primarily through ubiquitin-mediated proteasomal degradation. The proteasome is expressed in all cell types, and protein degradation can additionally occur by specialized proteasomes such as the immunoproteasome, expressed constitutively in APCs and T cells but also in other cell types

upon cytokine signaling such as interferon-γ (IFN-γ), or the thymoproteasome expressed in cortical thymic epithelial cells (41). Additionally, protein products that are defective due to improper splicing, folding, or frameshifts, termed defective ribosomal products (DRiPs), are also degraded (42). Following proteasomal degradation, peptides are trafficked to the endoplasmic reticulum (ER) via the TAP protein and trimmed by aminopeptidases. The corresponding peptides are then loaded onto partially folded MHC, which is bound to a series of chaperone proteins (Figure 4). Following peptide binding, MHC class I completes its folding and is transported to the cell membrane. It is through this pathway that MHC class I presents peptides from intracellular proteins. Less frequently MHC class I presents exogenous peptides in a process known as "cross-presentation". This process, which occurs most commonly in DCs, involves endocytosis of exogenous antigens that can then be loaded onto MHC I through two possible mechanisms (43): In the vacuolar pathway, internalized antigens are degraded by proteases and loaded onto MHC I entirely within the endosome; in the more dominant endosome-to-cytosol pathway, endocytosed antigens are trafficked to the cytosol where they undergo proteasomal degradation and peptides are then loaded onto MHC I in the ER. Cross-presentation thus allows DCs to present pathogen-derived peptides from which they were not infected, effectively allowing DCs to activate an immune response against viruses that infect other cell types (39, 44). This process is also key in the activation of an adaptive immune response against tumors (43).

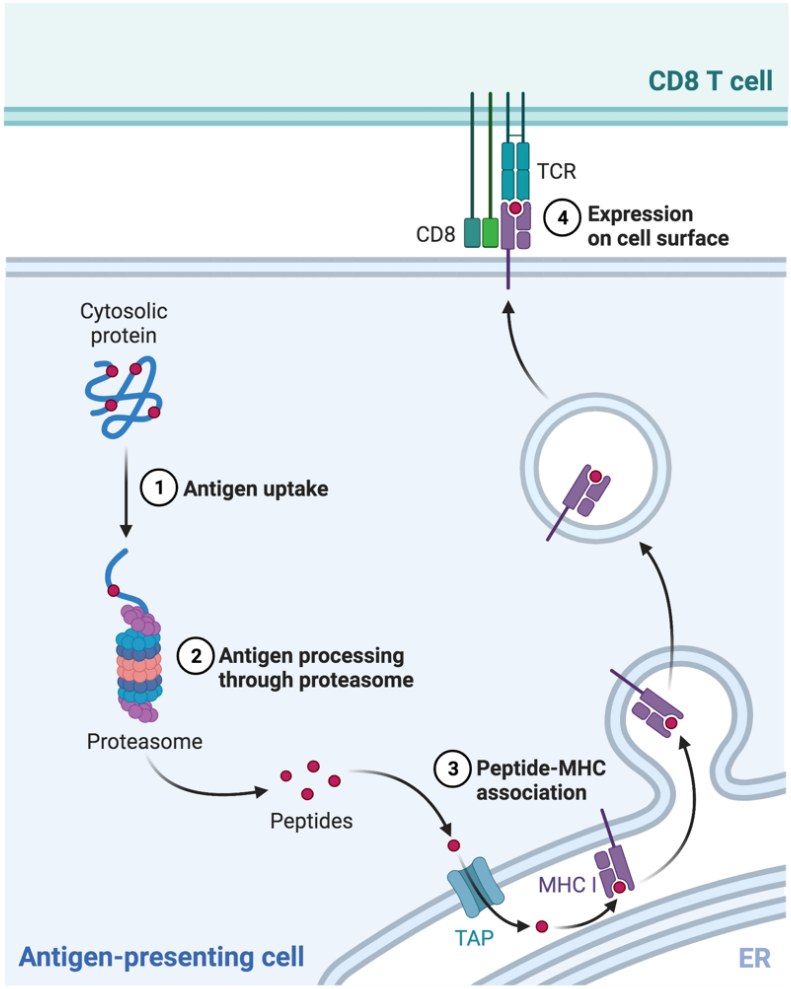

Figure 4. – MHC Class I peptide presentation pathway

In order for antigens to be presented by MHC I at the surface of an antigen presenting cell, their source proteins must first be taken up by the proteasome (1) and undergo proteasomal degradation (2). The resulting peptides are then loaded onto MHC I complexes in the endoplasmic reticulum (ER) (3), and the peptide-MHC I complexes are then trafficked to the cell surface (4), where they can then be recognized by T cell receptors. Adapted from "MHC class I and II Pathways", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

1.2.4 T cell receptor rearrangement

Hematopoietic stem cells are the precursors of both the myeloid cell lineage (dendritic cells, macrophages, monocytes, etc) as well as the lymphoid lineage (NK cells, B cells, and T cells). After originating from the bone marrow, progenitor cells migrate to the thymus for maturation, and undergo a rigorous series of selection prior to their evolution into mature T cells (39). When these progenitor cells arrive in the thymus, they have not yet undergone TCR genomic rearrangement and have few surface receptors. Following interactions with thymic epithelial cells or stromal compartment, these progenitor cells begin to proliferate and express certain surface molecules, but not yet CD4 or CD8 – such T cells are termed double negative thymocytes (i.e., CD4- and CD8-). As they develop in the 'double negative' stage, T cells begin expressing other receptors such as CD44 and CD25. As well, they begin the rearrangement of the β chain. The rearranged β chain then pairs with a surrogate α chain, termed the pre-T-cell receptor α chain (or pTα) (45). This pre-TCR pairs with CD3 to conduct signal intracellularly and allow the cell to proliferate and express both CD4 and CD8, thus termed double positive (DP) thymocytes (46). It is in this double positive state that the α chain rearranges to finalize the TCR.

1.2.5 T cell selection

In order to select only T cells that are capable of recognizing self-peptide presented by self-MHC and initiating an effective adaptive immune response, DP T cells undergo a process known as positive selection. Essentially, thymocytes will undergo cell death unless their TCR is able to recognize self-peptide:self-MHC complexes (47). As DP T cells are sampling these complexes, the strength of the signaling that they receive through CD4 or CD8 will determine to which of these two lineages they commit. The 10-30% of T cells that are able to recognize self-peptide:self-MHC will then progress to the single positive stage, in which they are either CD4+ or CD8+ (39). While T cells need to be able to recognize self-peptide presented by self-MHC, they should not initiate an immune response in response to self-peptides; this is often what is dysregulated in cases of autoimmunity. To prevent this, T cells in the thymus that are activated by self-antigens will be deleted. In order to test T cell response against a majority of proteins in the body, including those that are tissue-specific, this collection of proteins is expressed by certain medullary thymic stromal cells, controlled by the autoimmune regulator gene (47). Thus, T cells

that react too strongly to a self-peptide will undergo apoptosis. After maturation, T cells will exit the thymus and begin circulating in the blood and lymph.

1.2.6 T cell activation

While circulating in the body, T cells will continuously sample peptide:MHC complexes presented at the surface of cells. Once they are activated by the antigen:MHC complex to which they are specific, they will proliferate and differentiate into effector T cells. In the case of CD4+ T cells, this differentiation could result in different helper T cell classes, depending on the pathogen. For CD8+ T cells, this differentiation results in cytotoxic effector T cells, which proceed to kill infected cells presenting the particular antigen. The activation of a T cell does not uniquely consist of antigen recognition; it also requires signaling through costimulatory molecules (for example, CD28) and cytokine signaling that is important for differentiation into different effector cells (48). Aside from cytotoxic or helper effector types, T cells may also differentiate into memory T cells, which will be capable of effecting a more rapid response upon a potential second infection.

1.2.7 Immune checkpoints

Following an effector immune response, the immune system requires an “off-switch”, or some way to halt the T cell response and inflammation once the threat has been removed. There are several existing mechanisms to mitigate and “turn off” the immune response. Two such mechanisms, termed “immune checkpoints”, are molecules that bind various receptors on T cells to inhibit their effector functions. For example, in addition to TCR:MHC binding, T cells require additional signals for their activation. The surface receptor CD28 is typically bound by CD80 and CD86 costimulatory molecules to advance T cell activation and proliferation. Cytotoxic lymphocyte associated protein 4 (CTLA-4) is a competitive binder to CD28. Thus, the binding of CTLA-4 to CD28 is able to inhibit the proliferation of T cells in the lymph (49). CTLA-4 is also constitutively expressed on regulatory T cells, or Tregs, which mitigate the activity of effector T cells. By binding CD28, CTLA-4 prevents T cell costimulation and downstream PI3K/Akt signaling (50). These pathways guide the T cell towards aerobic glycolysis and are necessary for T cell differentiation and effector function (e.g. cytokine production) (51, 52). CTLA-4 also removes CD80 and CD86 from the surface of APCs through a process known as trans-endocytosis (53).

Additionally, programmed death 1 (PD-1) expressed by a variety of immune cells, including activated T and B cells, NK cells, macrophages, and DCs (54). PD-1 is bound by its ligands programmed death ligand 1 or 2 (PD-L1 or PD-L2) to inhibit T cell proliferation and reduce T cell survival. Upon PD-L1 engagement of PD-1, immunoreceptor tyrosine-based switch motif (ITSM) is phopshorylated and the subsequent dephosphorylation of CD28 by the recruited Src homology region 2 domain containing phosphatase-2 (SHP2) inhibits TCR activation (50). PD-1 expression is a mark of T cell exhaustion and is characteristic of chronic infections or cancer. CTLA-4 is thus able to inhibit the early stages of an immune response, typically preventing T cell activation in the lymph nodes. In contrast, PD-1's inhibition of T cell proliferation and survival occurs later in peripheral tissues. Together, these two immune checkpoints, along with Tregs, maintain homeostasis following a normal immune effector response.

1.2.8 Immune system and cancer

Over 20 years ago, Hanahan and Weinberg published the seminal "Hallmarks of Cancer", describing the features through which normal cells transform and become malignant (55). The original six hallmarks included genomic instability, epigenetic modifications, cancer cell proliferation, enhancement of cancer anti-apoptotic pathways, stimulation of angiogenesis, and cancer dissemination. In 2011, they published a follow-up to their original work, in which they discussed important advances in the field and proposed two additional emerging hallmarks and two enabling characteristics: these were deregulating cellular energetics, avoiding immune destruction, genomic instability and mutation, and tumor-promoting inflammation (56). The addition of 'avoiding immune destruction' and 'tumor-promoting inflammation' as characteristics of malignancy highlights the important roles of the immune system in cancer, both in its prevention and its promotion.

1.2.9 Tumor microenvironment

In the early stages of tumor initiation, individual or small groups of cells obtain successive mutations which result in their malignant transformation. However, immune cells are able to eliminate these cells through various anti-tumorigenic mechanisms. Eventually, certain cancer cells are able to progress through the incorporation of immune evasion strategies and the

recruitment of immunosuppressive cells (57). This tug-of-war between the effector and tolerogenic responses of the immune system decides the fate of the tumor and is highly influenced by the tumor microenvironment (TME), the highly heterogeneous region composing and surrounding a tumor. This region, which is of course composed of tumor cells but also stroma, blood vessels, and immune cells, is able to alter the course of the immune response (Figure 5). Many innate immune cells initially execute an anti-tumor response upon cancer development. NK cells and granulocytes are able to kill cancer cells through secretion of perforin or granzymes. NK cells secrete IFN-γ to activate T cell effector functions, as well as recognizing and responding to MHC allele loss that frequently occurs in cancer (58). Macrophages are initially also able to directly kill cancer cells, and dendritic cells play major roles in both the innate and adaptive immune responses to cancer. Unfortunately, an insidious feature of the TME is its ability to reconfigure and redirect the innate immune response for its own benefit, through the release of inhibitory cytokines. This pro-tumorigenic redirection involves the transition from anti-tumorigenic to tumor-associated macrophages, which advance processes such as angiogenesis and tumor-associated inflammation (59). DCs can also be redirected to tolerize the immune response to the tumor, and granulocytes can be coopted to have a pro-metastatic role (58, 60). Tumor cells are also able to downregulate their surface expression of MHC I, reducing their overall antigen presentation, in addition to the previously described immune checkpoints employed by the tumor.

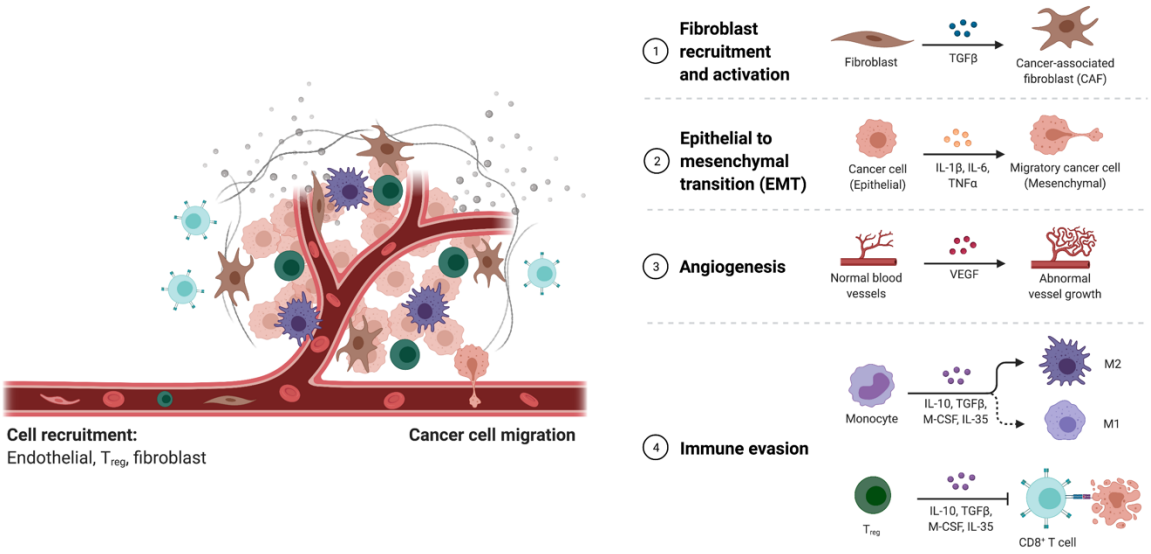

Figure 5. – Tumor microenvironment.

Overview of the various components of the tumor microenvironment and the cancer-associated changes that may occur. 1) Fibroblasts recuited to the tumor microenvironment develop pro-tumorigenic roles in response to TGF-B signaling; 2) Cancer cells are able to migrate and metastasize following cytokine signaling in the tumor microenvironment; 3) Vascular endothelial growth factor (VEGF) causes abnormal blood vessel growth, ultimately allowing further tumor development; 4) Monocytes are redirected to cancer-associated macrophages following cytokine signaling, and Treg recruitment contributes to further immune evasion by inhibiting CD8+ T cell function. Adapted from "The Tumor Microenvironment: Overview of Cancer-Associated Changes", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

1.2.10 Adaptive immune response to cancer

The adaptive immune system also plays a significant role in the anti-tumorigenic response. Mutations and dysregulation of cancer cells lead to the presentation of mutated or aberrantly expressed MHC I-associated peptides (MAPs) at the cell surface. These tumor-associated or tumor-specific antigens (TAAs and TSAs, respectively) may be recognized as non-self by the immune system, and thus activate a cytotoxic T cell response against these antigens. For over 50 years, these elusive tumor-associated and -specific antigens, which would allow an immunological attack on tumor cells, have been the subject of much study. Notably, Phil Gold and Samuel O. Freedman discovered carcinoembryonic antigen (CEA) in 1965, through a series of absorption and tolerance experiments using rabbits exposed to human colon carcinomas with normal tissue from

the same individual (61). Antisera from rabbits immunized with CRC tumor extract that was absorbed with excess normal colon tissue extract formed an antibody-antigen precipitate when exposed to CRC tumor extract. Thus, it was demonstrated that some component of the rabbit antisera was uniquely responding to a component of the CRC tumor. Further investigation revealed that CEA was found in several human cancers, and was expressed in embryos and then downregulated in later development, hence its name (62). This family of tumor-associated molecules are now the most used tumor marker to date and are often used in CRC screening and as a prognostic factor (63, 64). From this ground-breaking study, it was revealed that tumor-associated markers exist and that more likely remain to be discovered.

1.2.11 Immune checkpoint inhibition

Many tumor cells express CTLA-4 or PD-L, which is one of many mechanisms of immune evasion possessed by cancer. In light of this, such molecules were naturally targeted as cancer immunotherapy. The first immune checkpoint inhibitor, ipilimumab, was an anti-CTLA-4 monoclonal antibody initially approved for treating advanced melanoma (65-67). Since then, ICI has been tested and approved in several other cancers, including CRC (68). Despite the initial success of ICI in advanced melanoma, it was soon revealed that not all tumors respond equally, even within an individual cancer type. “Immune cold” or “immune desert” TMEs lack immune infiltration into the tumor bed which prevents effector T cells from eliminating the cancer; such tumors typically do not benefit from ICI therapy (66). In CRC, there is a sharp distinction in response to ICI between MSS and MSI subtypes. A 2015 phase II study using a PD-1 inhibitor in metastatic CRC demonstrated that patients with MSI tumors had both increased progression-free survival and increased overall survival compared to their MSS counterparts (69). In addition, MSI tumors are characterized by increased immune infiltration, particularly of tumor-infiltrating lymphocytes, and better prognosis. When taken together, these data suggest that MSI tumors are presenting more tumor antigens at the cell surface, which is causing the increased recruitment of TILs and the improved anti-tumorigenic immune response.

1.3 Analyzing MAPs by mass spectrometry

The human genome consists of approximately 25 000 genes, 20 000 of which are supposedly protein-coding. Given that protein variation can be introduced through alternative splicing events, polymorphisms, and post-translational modifications, the human proteome is many times more complex than the genome. This complexity is greater still when considering the dysregulation that occurs in cancer which potentially produces novel protein products (70). Thus, the human cancer proteome could consist of hundreds of thousands of proteins; a portion of these would generate MAPs and highly sought-after TSAs. The use of mass spectrometry in this regard has led to countless advances in the field. After isolating MAPs from cells through acid elution or immunoprecipitation, they are separated by reverse phase chromatography and ionized by electrospray ionization prior to mass spectrometry analysis (LC-MS). The peptides, suspended in solution, are dispersed as charged droplets. The solvent then evaporates, leading to smaller and increasingly charged droplets, from which the ions are eventually ejected in the gas phase (71, 72). These newly ionized peptides are then introduced into a mass analyzer which measures the mass to charge ratio (m/z) of these ions (Figure 6). The collection of m/z ratios of ions isolated at a given time is known as an MS1 spectrum (73). From this initial spectrum, the mass spectrometer can isolate a specific isolation window for further ion fragmentation through collision with inert gases. This fragmentation typically breaks the peptide apart at amino acid junctions, and the resulting m/z differences between peaks in the MS/MS spectrum can be used to sequence the peptide. These peptides can be sequenced by correlating the corresponding fragment ions with 'theoretical spectra' contained in a user-defined database (74). The use of MS/MS sequencing ensures unambiguous identification of MAPs presented at the cell surface by MHC class I instead of using *in silico* prediction based on RNASeq and HLA selection, a process riddle with an overwhelming number of false positives especially when predicting neoantigens (75). And so, rather than attempting to predict *in silico* the peptides that are presented at the cell surface by MHC class I, a process which is riddled with unknowns, mass spectrometry allows the direct identification and sequencing of these peptides.

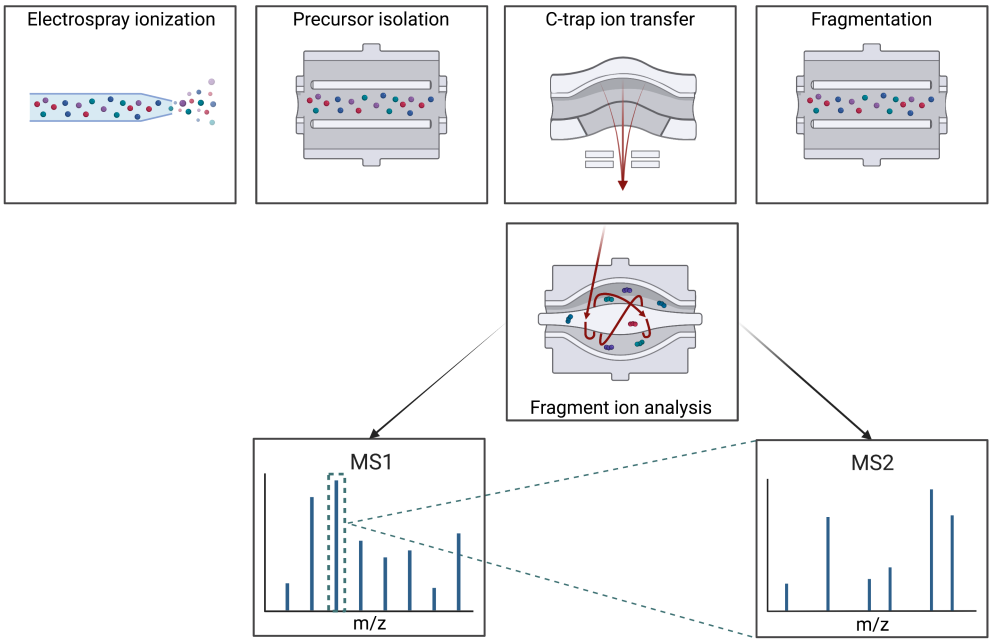

Figure 6. – Schematic of mass spectrometry analysis.

Depiction of mass spectrometry-based analysis of peptides. Peptides are first ionized by electrospray ionization, and precursor ions are then selected by a quadrupole. Precursors are transferred to the mass analyzer via the C-trap, after which they are analyzed to generate an MS1 spectrum. The precursors may also be fragmented in a second quadrupole, and sent back to the C-trap and then subsequently the mass analyzer, in order to generated MS2 spectra composed of the m/z ratios of a given fragment. Created with BioRender.com

1.3.1 Peptide quantification

Of course, the advantages of mass spectrometry do not end there. This powerful instrument also carries the ability to quantify peptides through various methods. The simplest, known as label-free quantification, quantifies the area-under-the-curve of a peptide across the MS1 spectra captured throughout an injection, and allows the relative quantification of a given peptide across conditions (73). However, this approach is not very precise, and peptides are not necessarily detected and identified across all conditions. Metabolic labeling techniques, such as Stable Isotope Labeling by Amino acids in Cell culture (SILAC) involve the *in vitro* labeling of proteins with light, medium, and heavy stable isotopes (ie ^{13}C, ^{15}N) of amino acids. The resulting peptides can be quantified in the MS1 spectrum to determine their relative intensities across the three samples. Unfortunately, the inherent complexity of the corresponding MS1 spectrum

means that the same peptide can be sequenced multiple times which limits proteome coverage. Alternatively, quantification can be performed at the MS/MS stage using isobaric peptide labeling which allows multiplex quantitative measurements of up to 18 conditions in a single injection (76). Tandem mass tag (TMT) is the most popular reagent used in isobaric peptide labeling, which involves the covalent addition of the mass tag on the free amino group of lysine and the N terminus of peptides. These tags consist of a reporter group and a mass normalization group, which have varying distributions of heavy isotopes that give all the tags the same mass but the reporter ions resulting from ion fragmentation will vary in m/z (73). Thus, a peptide from differentially labeled samples will appear as a single peak in the MS1 spectrum, but upon fragmentation the reporter ions will be distinct peaks in the MS2. This allows the relative quantification of a peptide between conditions, or even the absolute quantification if a known concentration of peptide is used. While TMT or other isobaric labeling approaches allow sample multiplexing, which reduces technical variability and increases throughput, the most commonly occurring problem with this approach is known as ratio distortion. This can result in several peptides being quantified together, and this interference or ratio distortion clouds the quantification of these peptides. Fortunately, several methods have been elucidated to address this. The most commonly used method is MS3, in which ions undergo another round of fragmentation to remove co-isolating peptides (73); the drawback of this approach is a reduction in sensitivity and comprehensiveness. Another exciting feature of mass spectrometry is the ability to perform targeted analysis of a sample, in which the peptides of interest are known and can be selected for fragmentation based on their m/z, charge, or other properties. This technique, known as parallel reaction monitoring or PRM, permits the detection of low abundance peptides that may not be isolated otherwise (77).

1.3.2 Databases for mass spectrometry analyses

In the Journal of the American Society of Mass Spectrometry in 2015, John R. Yates III published a comprehensive review summarizing the technological and theoretical advances of the past several decades that have revolutionized mass spectrometry (74). In particular, given the wealth of data that can be generated in a single mass spectrometry (MS) injection, developments in computational methods and technology have been especially integral to the development of

MS analyses. For example, incorporating computational calculations and algorithms to interpret spectra allowed MS instruments to incorporate faster scanning speed and acquire more spectra in a given run. In the past, MS analyses relied on spectral libraries or de novo sequencing to identify proteins. Both of these methods carry limitations: spectral libraries are only capable of sequencing peptides that have been previously identified and for which the spectra is included in the library (78), and de novo sequencing often struggles to distinguish b and y ions, resulting in incorrect peptide sequencing (79). Thus, when the human genome was sequenced in the early 2000s, this had a profound impact on MS analyses. Knowing the entire DNA sequence of the human genome allowed these nucleotides to be translated into amino acid sequences. This knowledge allowed researchers to use databases containing protein sequences that would be used to generate theoretical spectra. A peptide could thus be sequenced by matching an experimental spectrum to a theoretical spectrum for a protein sequence contained in the reference database. Initially, these databases were constructed using sequences from large, publicly available genome and protein databases such as Ensembl, RefSeq, or UniProtKB (80). A downfall of these databases is their inability to account for individual mutations such as single nucleotide polymorphisms (SNPs) or other unique variations that would not be characterized in these repositories. To address this, research groups began integrating genomic sequencing with proteomic analyses, which would come to be known as proteogenomics. In this approach, the personalized genomic data from an individual sample can be translated into amino acid sequences to construct the reference database, which will then be used for MS analyses. In fact, the algorithms used to correlate experimental and theoretical spectra were developed in the early 1990s alongside efforts to sequence MHC I- and II-associated peptides (81-83). Later, decoy sequences (inversed protein sequences from the reference database) were incorporated in order for software to calculate the false discovery rate (FDR) of peptide spectrum matches (PSMs). Today, there are several commercially available MS analysis software available, including Peaks (84, 85), MaxQuant (86), and SEQUEST (83).

1.4 Identification of tumor antigens

The identification of TAs from cell lines or tissues is a "numbers game"; It is known that these sequences exist, however, their rarity and the difficulty of MAP identification make them elusive. Years of work have led to the development of optimized methods to identify these antigens. For example, the isolation of MAPs from cells or tissues is a potential source of major material loss if not done carefully. A 2018 study in a B cell lymphoblastoid and an acute myeloid leukemia (AML) cell line demonstrated that immunoprecipitation with an anti-MHC class I antibody allowed the isolation of approximately six times more MAPs than the other commonly used approach of mild acid elution (87). MAPs present a challenge for MS-based identification in that they do not require tryptic digestion, which results in basic C-terminal amino acids, as one might use in a whole cell extract to digest proteins into peptides. This and the common amino acid composition of MAPs means that they are frequently lowly charged, making them more challenging to identify in MS analyses. To address this, TMT labeling was recently shown to improve MAP identification by enhancing the formation of multiply charged ions and increasing peptide hydrophobicity (88). A final challenge presented by identification of tumor antigens is that their identification does not guarantee their immunogenicity or their recognition by T cells. Peptide immunogenicity may be predicted *in silico* or evaluated experimentally, in studies involving *in vitro* T cell assays or humanized mice. Importantly, promising results in immunogenicity experiments do not guarantee T cell reactivity towards these antigens in the patient; it is also required that the patient possess CD8+ T cells specific to these peptides. Fortunately, T cell reactivity can also be evaluated using *in vitro* reactivity assays to determine if the blood from cancer patients contains T cells specific to the antigens of interest, or through measuring cytokine secretion in ELISA or ELISpot assays (89).

1.4.1 TAAs

TAAs are antigens that are more highly expressed on tumors, but are still presented by normal cells. Such antigens typically arise due to genetic or epigenetic changes that result in the increased expression and presentation of these sequences in cancer (Figure 7) (90). Given the cancer-associated processes that cause their overexpression, such sequences thus have the

capacity to be shared among tumors, however they are primarily exonic sequences and their expression on normal tissues makes them difficult to employ in the clinic without inducing either immune tolerance or adverse autoimmune responses. A subset of these antigens, however, known as cancer-testis antigens (CTAs), do show promise. These peptides, such as CEA or melanoma-associated antigens (MAGE) are absent from normal tissues with the exception of the testis, and show aberrant expression in some cancers (91). While certain TAAs and especially CTAs can be shared and immunogenic, research has expanded its purview to investigate TSAs, which also have the potential to be shared but could be comparatively absent from all normal tissues.

1.4.2 Mutated TSAs

In addition to technical challenges posed by the search for TSAs, researchers are also presented with theoretical challenges, such as to which type of TSA to devote their resources as well as how such sequences can be identified. Initially the majority of research in this area was devoted to the identification of mutated TSAs, or mTSAs; such peptides contain cancer-specific mutations, ensuring their tumor-specificity (Figure 7). Moreover, tumors can have enormous mutational burdens, carrying thousands of single-nucleotide variants (SNVs), in addition to insertion-deletions (INDELs) or gene fusion events. In cancers with large mutational burdens, such as melanoma, this suggests that they should be presenting many neoantigens. Indeed, personalized neoantigen vaccines were shown to induce effective anti-tumor T cell responses in melanoma patients immunized with a series of mutated peptides (92). In CRC, the abundance of neoantigens predicted to be immunoreactive correlated with patient survival (93). While many neoantigens would be specific to an individual tumor, the possibility of a neoantigen generated from a driver mutation seemed promising, as such a mutation could be shared among many cancers. For example, KRAS is mutated in approximately 40% of CRC, and so a neoantigen derived from this mutated protein could theoretically be applicable to many patients. In fact, infusion of tumor-infiltrating lymphocytes (TILs), which were primarily CD8+ T cells, specific to the G12D KRAS mutation resulted in regression of seven lung metastases in the CRC patient from which the TILs were derived (94). And yet, the majority of SNVs in a given tumor do not generate antigens presented at the cell surface. For example, out of 159 predicted neoantigens with evidence at the proteome level in sixteen hepatocellular carcinomas, no neoantigens were detected by MS (95).

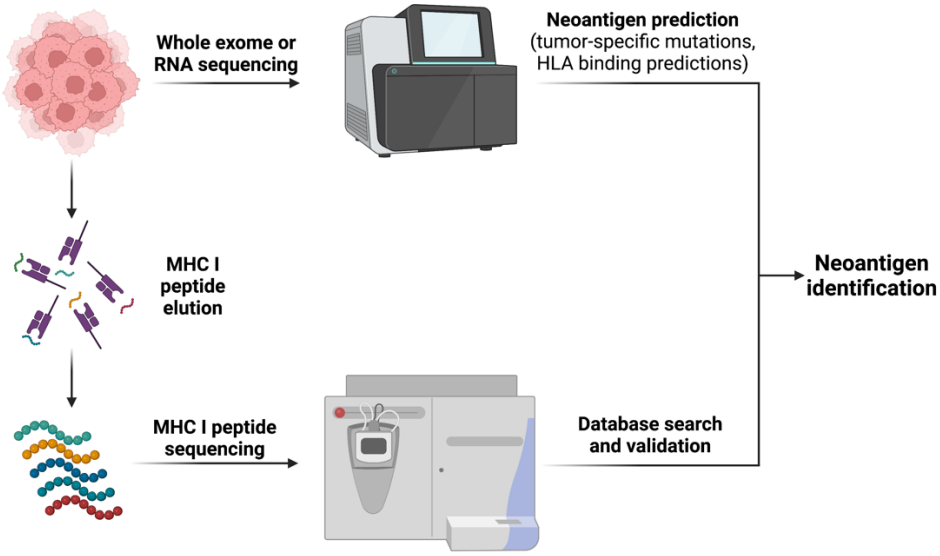

Figure 7. – Standard mTSA/neoantigen identification

Mutated tumor-specific antigens, or neoantigens, are often predicted using Next Generation Sequencing to identify cancer-specific mutations, which then undergo HLA binding predictions to identify which mutations are promising as neoantigen candidates. In tandem, MHC I peptides are eluted from the samples and analyzed by mass spectrometry to identify which predicted neoantigens are effectively being presented at the cell surface. Created with BioRender.com

1.4.3 Aberrantly expressed TSAs

Given that the neoantigen, or mTSA, yield is so low, particularly in cancers with lower mutational burdens, the search for TSAs has turned elsewhere. Cancer cells are ridden with mutations and aberrations that make them fundamentally different in a way that is recognizable to immune cells, and yet SNVs do not seem to be the answer. INDEL mutations, the second most common type of mutation, would presumably result in frameshift events that could generate novel MAPs in cancer. In fact, a frameshift peptide derived from TGFBRII was shown to induce proliferation of CD4+ T cells from CRC patients (96), while TILs specific for another frameshift peptide from the same gene were able to lyse a CRC-derived cell line (97). Although these results suggest a certain immunogenicity of these frameshift peptides, neither of these studies demonstrated that the peptides of interest were presented at the cell surface by MHC. Fortunately, the cancer-specific aberrantly expressed sequences do not end there. SNV and INDEL studies, as well as those that study TAAs, have focused exclusively on coding regions of the genome. However, exons make up only 2% of the genome, whereas up to 75% of the genome can

be transcribed and potentially translated (98). MAPs derived from introns, untranslated regions (UTRs), long non-coding RNAs (lncRNAs), and intergenic regions have all been identified by MS (99). The translation of these normally-silenced sequences and others arises through processes such as epigenetic alterations, slicing aberrations, expression of endogenous retroviral elements (EREs), or the aberrant translation of transcripts (100). The profound implication of this is that there is a wealth of genomic regions remaining to possibly generate TSAs. In addition to being more abundant than mTSAs, aberrantly expressed TSAs (aeTSAs) also have more potential to be shared across tumors, since they do not rely on cancer-specific mutations, many of which are patient-specific.

As the proteogenomic search for TSAs expands to all regions of the genome, so too do databases expand to include the entirely of the potential protein sequences expressed in all genomic regions as well as in all possible reading frames. However, the expansion of database size is not without cost; the increase in theoretical spectra in the database will increase required computation time, and can reduce identification and PSM quality (80). When confronted with this problem, researchers developed various strategies to maintain the ability to identify peptides from all genomic regions while decreasing database size. One such approach involved constructing a database that consisted of a canonical cancer proteome generated through *in silico* translation of RNA-seq data, as well as cancer-specific sequences, generated by removing any sequence present in thymic epithelial cells (TECs), and it was only these cancer-specific sequences that were translated into all reading frames (101). This approach, which involves examining the entire genome, including non-coding regions, for TSAs, has previously led to the identification of aeTSAs in acute lymphoblastic leukemia and lung cancer (101), ovarian cancer (102), and acute myeloid leukemia (103).

1.4.4 TSAs in CRC

As with other cancers, CRC immunopeptidomic studies have focused exclusively on mTSAs or TAAs. One study in MSS CRC organoids demonstrated that out of 304 genes predicted to generate neoantigens, only three such antigens were detectable by MS (104). A more recent study identified only a single mTSA from an MSI tumor with nearly 4000 non-synonymous

mutations (105). A third group compared a large set of paired tumor and normal adjacent tissue (NAT) and identified a series of TAAs of interest, however as with these other studies, they focused exclusively on coding regions of the genome. Additionally, only one study explicitly investigated both MSS and MSI cancers. In summary, no studies have identified aberrantly expressed TSAs in CRC to date.

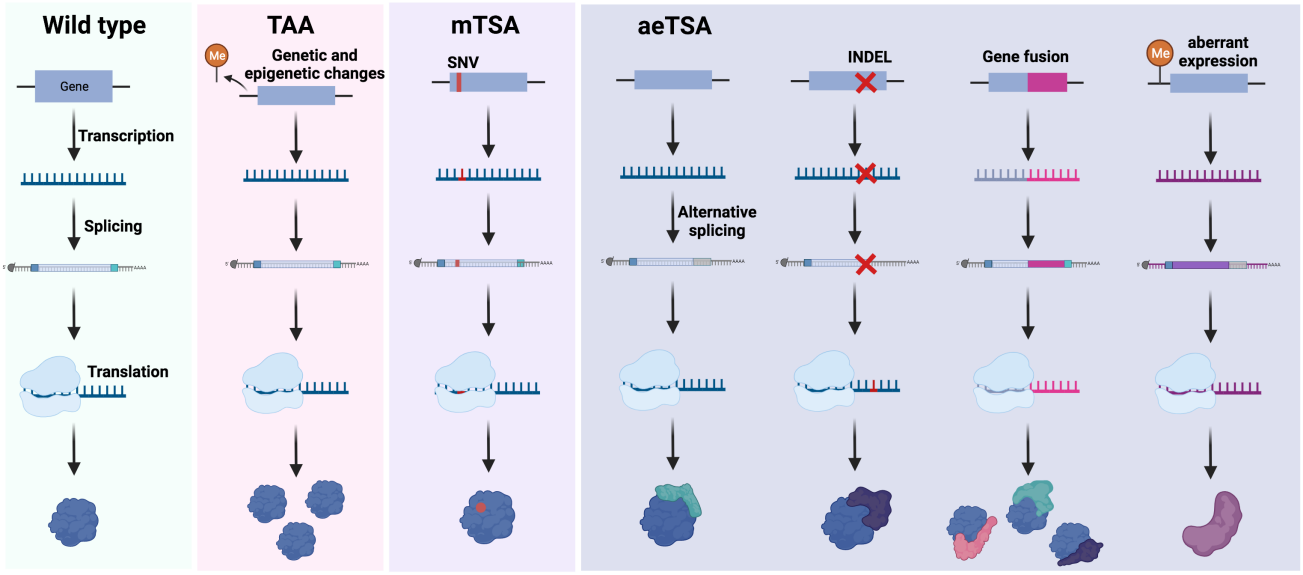

Figure 8. – Types of tumor antigens

Representation of the various mechanisms through which different classes of tumor antigens can be generated. Small blue balls represent the protein generated under normal conditions, which can undergo cancer-induced mutations (red) or other modifications, resulting in translation of non-coding sequences (green protein), out-of-frame exons (dark purple protein), or newly expressed sequences (light purple protein). Created with BioRender.com

1.5 Research Objectives

With advancements in the efficacy and sensitivity of mass spectrometry analyses in recent years, it is now simpler than ever to sequence thousands of MHC I-associated peptides presented at the cell surface, making the immunopeptidome available for study on a silver platter. However, there remain vast regions of the genome often unexplored by traditional analyses and database construction techniques. In addition, the difference in prognosis between MSS and MSI tumors imposes a sense of urgency on the identification of therapies that can bridge the gap in treatment efficacy between these two molecular subtypes. We hypothesize that our approach will uncover TSAs in CRC, the majority of which will be from non-coding regions and will be identified in MSI tumors. Our research objectives were thus as follows:

1. To elucidate the immunopeptidomes of a series of CRC-derived cell lines and matched CRC tumor and normal adjacent tissues using personalized databases for MS-based identification
2. To identify TSAs in both MSS and MSI samples using a stringent identification pipeline
3. To validate the TSAs in terms of intertumoral distribution, immunogenicity, and tumor specificity

1.6 Thesis Overview

The body of this thesis summarizes the work I completed throughout my Master's degree, culminating in a publication currently under review for Molecular and Cellular Proteomics. In this paper, we utilized a novel proteogenomic approach to analyze a series of CRC-derived cell lines and matched tumor and NAT. We applied a stringent set of filters to a large dataset of MAPs identified by mass spectrometry to identify 19 novel TSAs, the majority of which are aberrantly expressed and derive from non-coding sequences. In addition, we were able to identify TSAs in both MSS and MSI CRC tissues. The final chapter of this thesis provides a conclusion and perspectives on the preceding work, centering its potential in future cancer immunotherapeutic strategies.

Chapter 2: Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability

Authors: Jenna Cleyle[1,2], Marie-Pierre Hardy[1], Robin Minati[1,2], Mathieu Courcelles[1], Chantal Durette[1], Joel Lanoix[1], Jean-Philippe Laverdure[1], Krystel Vincent[1], Claude Perreault[1,3], Pierre Thibault[1, 4.]

[1] Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, QC, Canada.

[2] Molecular Biology Program, Université de Montréal, Montreal, QC, Canada.

[3] Department of Medicine, Université de Montréal, Montreal, QC, Canada.

[4] Department of Chemistry, Université de Montréal, Montreal, QC, Canada.

Under revision:

Molecular and Cellular Proteomics, 2022, Volume 20, Special Issue: Immunopeptidomics

Author contributions:

Jenna Cleyle: Methodology, Investigation, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization, Funding acquisition; Marie-Pierre Hardy: Methodology, Investigation, Formal analysis, Writing – Review & Editing; Robin Minati: Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization; Mathieu Courcelles: Software, Formal analysis, Writing – Review & Editing, Visualization; Chantal Durette: Investigation, Writing – Review & Editing; Joel Lanoix: Investigation, Writing – Review & Editing; Jean-Philippe Laverdure: Software, Formal analysis, Writing – Review & Editing; Krystel Vincent: Methodology, Writing – Review & Editing; Claude Perreault: Conceptualization, Writing – Review & Editing, Funding acquisition; Pierre Thibault: Conceptualization, Methodology, Writing – Review & Editing, Project administration, Funding acquisition.

2.1 Abstract

Colorectal cancer is the second leading cause of cancer death worldwide, and the incidence of this disease is expected to increase as global socioeconomic changes occur. Immune checkpoint inhibition therapy is effective in treating a minority of colorectal cancer tumors; however, microsatellite stable tumors do not respond well to this treatment. Emerging cancer immunotherapeutic strategies aim to activate a cytotoxic T cell response against tumor-specific antigens, presented exclusively at the cell surface of cancer cells. These antigens are rare and are most effectively identified with a mass spectrometry-based approach, which allows the direct sampling and sequencing of these peptides. While the few tumor-specific antigens identified to date are derived from coding regions of the genome, recent findings indicate that a large proportion of tumor-specific antigens originate from allegedly noncoding regions. Here, we employed a novel proteogenomic approach to identify tumor antigens in a collection of colorectal cancer-derived cell lines and biopsy samples consisting of matched tumor and normal adjacent tissue. The generation of personalized cancer databases paired with mass spectrometry analyses permitted the identification of more than 30 000 unique MHC I-associated peptides. We identified 19 tumor-specific antigens in both microsatellite stable and unstable tumors, over two-thirds of which were derived from non-coding regions. Many of these peptides were derived from source genes known to be involved in colorectal cancer progression, suggesting that antigens from these genes could have therapeutic potential in a wide range of tumors. These findings could benefit the development of T cell-based vaccines, in which T cells are primed against these antigens to target and eradicate tumors. Such a vaccine could be used in tandem with existing immune checkpoint inhibition therapies, to bridge the gap in treatment efficacy across subtypes of colorectal cancer with varying prognoses. Data are available via ProteomeXchange with identifier PXD028309.

2.2 Introduction

CRC is the third most commonly diagnosed cancer and the second leading cause of cancer death worldwide, with over 1.8 million cases and 881 000 deaths estimated in 2018 alone (106). The incidence of CRC is expected to increase as global socioeconomic changes occur, with a predicted 2.2 million cases and 1.1 million deaths occurring annually by 2030 (3, 106). This significant disease burden highlights the necessity of developing new and effective treatments.

The positive correlation between the abundance of TILs and increased overall survival in both colon and rectal cancer suggests that T cells can recognize biologically relevant tumor antigens in these tumors (107, 108). The potential immunogenicity of these antigens made ICI a promising treatment for cancer patients; however, early clinical trials evaluating their efficacy in CRC have yielded mixed results. Colorectal tumors characterized by deficiencies in mismatch repair proteins resulting in the accumulation of repetitive DNA sequences (microsatellites), known as MSI, have shown relative success in phase II clinical trials with anti-PD1 treatment (69). In contrast, such treatments have had very little efficacy in clinical trials against MSS tumors that do not possess a high mutational burden, which make up approximately 80% of CRC cases (69, 109).

Given the significance of the immune response in CRC and the limited success of ICI alone, a promising research avenue in recent years has been neoantigen-based vaccines or T cell receptor-based therapy, which could be administered with ICI and would ideally bridge the gap in treatment efficacy across MSI and MSS tumors. In line with this, TAAs, which are overexpressed in cancer cells compared to normal cells, have been previously identified in CRC (110, 111). While several TAAs have been tested in vaccine and phase I trials against CRC, most were met with “limited success”, likely due to the negative selection of TAA-responsive T cells in the thymus (112). In a study by Parkhurst et al., the treatment of metastatic CRC with genetically engineered anti- CEA T cells resulted in tumor regression in one individual but “serious inflammatory colitis” in all patients, demonstrating that an adverse autoimmune response is another possible consequence of targeting TAAs (113).

The mixed responses to TAA-based therapy suggest that targeting TSAs would be more effective. These antigens may be generated through genetic, epigenetic, and post-translational variations, including but not limited to single-nucleotide variants, aberrantly expressed transcripts, or novel splicing events, and are exclusively presented by tumor cells (100). The high prevalence of single nucleotide variants, splice variants, and INDEL mutations in CRC suggests that there is a higher probability of unique antigen presentation by the MHC molecules of tumors compared to other cancers with lower mutational loads. These antigens, or MAPs, would make it possible to invoke a tumor-specific immune response (114). TSAs have recently been identified in CRC and have demonstrated some success in phase I and II vaccine trials. A 2015 vaccine trial using frameshift antigens originating from MSI-high tumors demonstrated significant and specific immune responses among all patients (115). However, as this study used antigens derived from frameshift mutations associated with MSI, these findings do not apply to the majority of CRC patients. Other studies identifying TSAs in CRC to date have focused exclusively on mTSAs derived from coding regions of the genome (115, 116). An investigation of MSS CRC organoids revealed that only 0.5% of non-silent mutations generated mTSAs; this was a significantly lower proportion than what was anticipated by HLA-binding prediction software (104). It was recently demonstrated that the majority of actionable TSAs arise from non-coding regions of the genome and from aberrantly expressed transcripts, rather than somatic mutations (101-103). While mTSAs are tumor-specific unless derived from common driver mutations, these aeTSAs are particularly noteworthy because they may be shared by multiple tumors. In addition, previous studies did not employ MS techniques to quantify the expression of those TSAs on tumor cells, which is information that could influence the therapeutic potential of targeting a given TSA (115, 116).

In the present study, we use an MS-based approach that leverages personalized databases to directly identify TSAs presented by CRC-derived cell lines and tumor biopsies and allows the identification of TSAs from non-coding regions. By using this approach, we identify 19 TSAs across our samples, as well as a variety of TAAs. Further, we identify TSAs in both MSS and MSI tumors, suggesting that MSS tumors present immunologically relevant antigens that could be exploited to bridge the gap in treatment efficacy of ICI in various subtypes of CRC.

2.3 Experimental Procedures

2.3.1 Samples

2.3.1.1 Cell lines

Four human colorectal cancer cell lines [COLO 205 (ATCC® CCL-222™), HCT 116 (ATCC® CCL-247™), RKO (ATCC® CRL-2577™), SW620 [SW-620] (ATCC® CCL-227™)] and one human normal fetal small intestine cell line [HIEC¬6 (ATCC® CRL¬3266™)] were obtained from the American Type Culture Collection (ATCC). COLO205, HCT116, and SW620 were grown in RPMI-1640 (Gibco) supplemented with 10% Fetal bovine serum (FBS), RKO was grown in Eagle's Minimum Essential Medium (EMEM) (ATCC) supplemented with 10% FBS, and HIEC-6 was grown in OptiMEM 1 Reduced Serum Medium (Gibco) supplemented with 20 mM HEPES (Gibco), 10 mM GlutaMAX (Gibco), 10ng/mL epidermal growth factor (EGF) (Gibco), and FBS to a final concentration of 4%. All cells were maintained at 37°C with 5% CO_2.

For collection, cells were rinsed with warm phosphate-buffered saline (PBS) before being trypsinized with TrypLE™ Express Enzyme (1X) (Gibco) for 5-15 minutes at 37°C with 5% CO_2. Harvested material was then spun at 1000rpm for 5 minutes, rinsed once with warm PBS, then resuspended in ice-cold PBS. After cell count, replicates of 2 x 10^8 CRC cells were pelleted and frozen at -80°C until further use. MHC class I surface density of the CRC cell lines was determined by Qifikit (Agilent) using the W6/32 anti-HLA class I antibody (BioXCell), according to the manufacturer's instructions.

2.3.1.2 Primary tissues

Six pairs of primary human samples consisting of matched colon adenocarcinoma tumor and NAT were purchased from Tissue Solutions. Tissue samples were taken from patients receiving surgery as the first line of treatment and were flash-frozen in liquid nitrogen. More information about primary tissue samples can be found in Table 2.

2.3.2 RNA extraction and sequencing

2.3.2.1 RNA extraction

For RNA extraction of cell lines, 1-2 million cells were collected and washed once with ice-cold PBS. The cells were then resuspended in Trizol (Invitrogen). Total RNA was isolated using the RNeasy Mini kit (Qiagen) or the AllPrep DNA/RNA/miRNA Universal kit (Qiagen) as recommended by the manufacturer, for cell lines and tissues, respectively.

2.3.2.2 RNA sequencing

500 ng of total RNA was used for library preparation. RNA quality control was assessed with the Bioanalyzer RNA 6000 Nano assay on the 2100 Bioanalyzer system (Agilent Technologies) and all samples had an RNA integrity number (RIN) above 6.8 for NAT and above 8 for cancer samples. Libraries were prepared with the KAPA mRNAseq Hyperprep kit (Roche). Ligation was made with Illumina dual-index UMI (IDT). After being validated on a BioAnalyzer DNA1000 chip and quantified by QuBit and qPCR, libraries were pooled to equimolar concentration and sequenced with the Illumina Nextseq500 using the Nextseq High Output 150 (2x75bp) cycles kit. A mean of 129 and 95 million paired-end PF reads were generated for the cell lines and tissue samples, respectively. Library preparation and sequencing were performed at the Genomic Platform of the Institute for Research in Immunology and Cancer (IRIC).

2.3.2.3 Bioinformatic analyses

Sequences were trimmed using Trimmomatic version 0.35 (117) and aligned to the reference human genome version GRCh38 (gene annotation from Gencode version 33, based on Ensembl 99) using STAR version 2.7.1a (118). Gene expressions were obtained both as read count directly from STAR as well as computed using RSEM (119) to obtain normalized gene and transcript-level expression, in transcript-per-million (TPM) values, for these stranded RNA libraries.

2.3.3 Transcriptomics

2.3.3.1 HLA genotyping

HLA genotyping of cell lines and tissues was performed using OptiType, an online HLA genotyping tool that uses RNA-Seq data to predict a sample's HLA alleles (https://github.com/FRED-2/OptiType) (120). HLA alleles of cell lines were confirmed with what is documented in the literature, and if these differed from Optitype predictions, we preferentially selected those in the literature.

2.3.3.2 Microsatellite instability detection

MSI status of the primary tumor samples was evaluated using the MSIsensor-pro1.0a program using paired tumor and NAT (https://github.com/xjtu-omics/msisensor-pro) (121).

2.3.3.3 Differential expression analysis

DESeq2 version 1.22.2 (122) was used to normalize gene read counts. Principal component analyses (PCA) were generated using normalized log read counts for the first two most significant components. The PCA was generated in an unsupervised manner. The 500 genes were those presenting the biggest standard deviation based on their expression levels across all samples. DESeq2 was only used to normalize the read counts, not to perform a differential expression analysis. For differential expression analysis of the cell lines, fold changes were computed between the mean expression of the four CRC cell lines compared to the normal cell line (HIEC-6). Significant differentially expressed genes (DEGs), those with padj < 0.05 and $|\log2 \text{ fold change}| > 1$, were considered for gene ontology (GO) terms using the Metascape tool (123). For paired differential expression analysis of the tissues, TPM normalized values were used to compare tumor/NAT pairs. As only a single replicate of the tissues was sequenced, rather than filtering by adjusted p-value, we selected only genes that were significantly differentially expressed in all six subjects for GO term analysis with $|\log2 \text{ fold change}| > 1$. When examining differentially expressed genes between MSS and MSI tissues, the same fold change thresholds were applied. For GO term analysis of MSI DEGs, genes were selected that were exclusively differentially expressed in both MSI tissues (i.e. not considered DEGs in any MSS tissues). For GO

term analysis of MSS DEGs, genes were considered if they were differentially expressed in three or more MSS tissues.

2.3.3.4 Transcriptome analysis of tissue samples

The proportion of various biotypes in the transcriptome of tissue samples was determined as previously described (124). Briefly, following quantification and alignment of Ensembl annotated transcripts by Kallisto (119), transcripts and repetitive elements were annotated using a Kallisto index containing Ensembl annotated transcripts supplemented with genetic repeat identifications from the UCSC Table Browser GRCh38 repeat masker database (125). Transcript expression was quantified in TPM.

2.3.3.5 Mutation profiles and genetic variant annotation

Genetic variant calling was performed for both cell line and primary biopsies using SNPEff (https://pcingola.github.io/SnpEff/#snpeff) (126).

2.3.4 Database generation

Global cancer databases were constructed as previously described (101). In brief, RNA-sequencing (RNA-seq) reads were trimmed using Trimmomatic version 0.35 (117) and aligned to the reference human genome version GRCh38 (gene annotation from Gencode version 33, based on Ensembl 99) using STAR version 2.7.1a (118). Kallisto (https://pachterlab.github.io/kallisto) was used to quantify transcript expression in TPM (119). Sample-specific exomes were constructed by integrating single nucleotide variants (quality>20) identified with Freebayes (https://github.com/ekg/freebayes) into PyGeno (127). Annotated open reading frames with TPM > 0 were then translated in silico and added to the canonical proteome in fasta format. We selected medullary thymic epithelial cells (mTECs) and TECs as a positive control because they 1) express a large collection of self-peptides and 2) establish central tolerance in the thymus (negative selection of T cells). mTECs (n=6) and TECs (n=2) were thus used to generate the cancer-specific proteome for cell lines (GEO accessions GSE127825, GSE127826). The respective NAT for each primary tumor sample was used in place of mTECs for this portion of database construction, as it approximates 'normal' expression for that subject. RNA-seq reads were cut into 33-nucleotide sequences known as k-mers and only k-mers present <2 in mTECs or matched NAT for

cell lines and tissues, respectively, were kept. Overlapping k-mers were assembled into contigs, which were then 3-frame translated in silico. Of note, short peptide sequences generated through the k-mer approach were then concatenated into longer sequences of approximately ten thousand amino acids. To reduce the number of small separate sequences in the cancer-specific, these peptides were concatenated using the 'JJ' sequence as a separator, which is recognized internally by the PeaksX+ software to split sequences upon occurrence of this sequence. Then, the canonical proteome and the cancer-specific proteome were concatenated to create the global cancer databases. Cell line databases consisted of 3.38×10^6 sequences on average.

2.3.5 Isolation of MAPs

CRC cell line pellet samples (2×10^8 cells per replicate, four replicates per cell line) were resuspended with PBS up to 2 mL and then solubilized by adding 2 mL of ice-cold 2X lysis buffer (1% w/v CHAPS). Tumor and normal adjacent tissue samples (average 568mg) were cut into small pieces (cubes, ~3 mm in size) and 5 ml of ice-cold PBS containing protein inhibitor cocktail (Sigma, cat#P8340-5ml) was added. Tissues were first homogenized twice for 20 seconds using an Ultra Turrax T25 homogenizer (IKA-Labortechnik) set at a speed of 20 000 rpm and then 20 seconds using an Ultra Turrax T8 homogenizer (IKA-Labortechnik) set at speed 25 000 rpm. Then, 550 µl of ice-cold 10X lysis buffer (5% w/v CHAPS) was added to each sample. After 60-minute incubation with tumbling at 4°C, tissue samples and CRC cell line samples were spun at 10 000g for 30 minutes at 4°C. Supernatants were transferred into tubes containing 1 mg of W6/32 antibody covalently-cross-linked protein A magnetic beads and MAPs were immunoprecipitated as previously described (128). MAP extracts were then dried using a Speed-Vac and kept frozen before MS analyses.

2.3.6 TMT labeling

MAP extracts were resuspended in 200mM HEPES buffer pH 8.1. 50 µg of TMT reagent (Thermo Fisher Scientific) in anhydrous acetonitrile was added to samples as follows: CRC cell line replicates were labeled with TMT6plex (lot #UG287166) channels TMT6-126 to 129; Tissue samples were labeled with TMT10plex (lot # UH285228) -126 (NAT) and -127N (tumor). Samples were gently vortexed and reacted at room temperature for 1.5 hours. Samples were then

quenched with 50% hydroxylamine for 30 minutes at room temperature, then were diluted with 4%FA/H2O. CRC cell line replicates and individual NAT-tumor pairs were combined. Samples were then desalted on homemade C18 membrane (Empore) columns and stored at -20°C until injection. Labeling efficiency was calculated using PeaksX+ search results (see 'MAP identification' section below), by taking the proportion of TMT PSMs over the total number of PSMs.

2.3.7 Mass spectrometry analyses

Dried peptide extracts were resuspended in 4% FA and loaded on a homemade C18 analytical column (20 cm x 150 μm i.d. packed with C18 Jupiter Phenomenex) with a 106-minute gradient from 0% to 30% ACN (0.2% FA) and a 600 nL/min flow rate on an EASY-nLC II system. Samples were analyzed with an Orbitrap Exploris 480 spectrometer (Thermo Fisher Scientific) in positive ion mode with Nanoflex source at 2.8kV. Each full MS spectrum, acquired with a 240 000 resolution was followed by 20 MS/MS spectra, where the most abundant multiply charged ions were selected for MS/MS sequencing with a resolution of 30 000, an automatic gain control target of 100%, an injection time of 700ms, and collisional energy of 40%. LC-MS instrument was controlled using Xcalibur version 4.4 (Thermo Fisher Scientific, Inc).

2.3.8 MAP identification

Database searches were conducted using the PeaksX+ software, version 10.6 (Bioinformatics Solutions Inc.) (84). Error tolerances for precursor mass and fragment ions were set to 10.0ppm and 0.01 Da, respectively. A non-specific digest mode was used. TMT10plex was set as a fixed post-translational modification (PTM), and variable modifications included phosphorylation (STY), Oxidation (M), Deamidation (NQ), and TMT10plex STY. Peaks searches were then loaded into MAPDP (129), which was used to apply the following filters: selecting peptides of 8-11 amino acids in length, with rank eluted ligand threshold ≤ 2% based on NetMHCpan-4.1b predictions, using a 5% false discovery rate (FDR). FDR was calculated using the decoy hits imported from Peaks, which employ the decoy-fusion strategy (85).

2.3.8.1 MAP source gene analysis

GO term analysis was performed for CRC-derived cell lines and primary tissues with the Metascape tool (123). A list of source genes was generated for each sample by taking all of the source genes associated with the MHC I immunopeptidome of that set of samples and removing duplicates (i.e., although a source gene may generate more than one unique peptide, it would only be counted once in the source gene analyses). For tissues, only source genes shared by four or more tissues were included in this analysis.

2.3.9 Quantification of MAP coding sequences in RNA-Seq data

MAP coding sequences (MCSs) were quantified in RNA-seq data as previously described (103). Briefly, MCSs were reverse translated into all possible nucleotide sequences with an in-house python script (deposited on Zenodo at DOI: 3739257). The nucleotide sequences were then mapped onto the genome with GSNAP (130) to determine all possible genomic locations able to code for a given MAP. MCSs were also mapped onto the transcriptome to account for MAPs overlapping splice sites, and portions of the transcriptome corresponding to these MAPs were then also mapped onto the reference genome with GSNAP. For MAPs of interest, we performed genomic alignment of all reads containing the MCS. GSNAP output was filtered to keep only perfect matches between the sequence and reference, resulting in a file containing all possible genomic regions able to code for a given MAP. We summed the number of reads containing the MCSs at their respective genomic locations in each desired RNA-Seq sample (such as CRC and NAT, Genotype Tissue Expression project (GTEx), or the Cancer Genome Atlas (TCGA) samples), aligned on the reference genome with STAR. Lastly, all read counts for a given MAP were summed and normalized on the total number of reads sequenced in each sample of interest to obtain a reads-per-hundred-million (RPHM) count.

2.3.10 Determination of MAP source transcripts

To investigate what proportion of tissue sample MAPs were derived from certain transcript biotypes, the most abundant putative source transcript based on kmer-per-hundred-milion (KPHM) quantification was determined. For peptides from the cancer-specific (kmer) database, the MCSs were reverse translated into all possible nucleotide sequences and all

possible genomic regions able to code for a given MAP were identified (see 'Quantification of MAP coding sequences in RNA-Seq data' above). Finally, Kallisto was used to determine the most expressed transcript at that location, which was then assigned as the most probable transcript for the given peptide. Peptides that had more than one putative source transcript were excluded from the analysis.

2.3.11 Identification of TSA candidates

TSA candidates were identified through a stringent TSA identification pipeline. First, MAPs underwent peptide classification in which the peptide sequence accessions were retrieved from the protein database and used to extract the nucleotide sequences of each peptide. RNA-Seq data from each cancer and normal samples were transformed into 24-nucleotide-long k-mer databases with Jellyfish 2.2.3 (using the –C option) and used to query each TSA candidate coding sequence's 24-nucleotide-long k-mer set. The number of reads fully overlapping a given peptide-coding sequence was estimated using the k-mer set's minimum occurrence ($kmin$), as in general, one k-mer always originates from a single RNA-Seq read. We then transformed this kmin value into several k-mers detected per 10^8 reads sequenced (kphm) using the following formula: kphm = (kmin × 10^8)/$rtot$, with $rtot$ representing the total number of reads sequenced in a given RNA-Seq experiment. Peptides were kept only if their RNA coding sequences were expressed at least 10-fold higher in cancer than in normal (pooled mTEC samples for cell lines, matched NAT for tissues), and expressed < 2 KPHM in normal. Subsequent filtering removed any peptides with indistinguishable isoleucine/leucine variants; a peptide with an IL variant was kept only if the most expressed variant met the above-mentioned criteria. The MCSs of the remaining peptides were quantified in RNA-seq data as described above and were kept only if their expression was < 8.55 RPHM in mTECs and other normal tissues (GTEx). Genomic localization for each peptide was assigned by mapping reads containing each MCS to the reference genome (GRCh38.99) using BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat). Peptides were excluded if the genomic localization was unclear or if they mapped to a hypervariable region (HLA, Ig, or T cell receptor (TCR) genes). Finally, the MS/MS spectra of the remaining candidates were manually validated. Peptides were classified as mTSAs if their amino acid sequence was different from the reference, and if the mutation was not a known germline polymorphism. Peptides were classified as aeTSAs

if they were overexpressed ≥10-fold in tumor compared to normal and ≤0.2 KPHM in mTECs (and NAT in the case of tissues) and as TAAs if they were overexpressed ≥10-fold in cancer but the expression in mTECs and/or NAT was > 0.2 KPHM. Ultimately, the transcript of origin of each TSA/TAA was attributed by selecting the most highly expressed peptide-overlapping transcript from the kallisto quantification file (see Database Generation section).

2.3.11.1 Intertumoral sharing

To examine the intertumoral distribution of TSA and TAA sequences in other CRC tumors, the log(RPHM+1) expression of the peptide coding sequences in 151 colon adenocarcinoma (COAD) samples from TCGA was determined (see 'Quantification of MAP coding sequences in RNA-Seq data').

2.3.11.2 Immunogenicity prediction

The predicted immunogenicity of MAPs of interest was determined with the R package Repitope v3.0.1 (https://github.com/masato-ogishi/Repitope) (131).

2.3.12 TSA validation and relative quantification with synthetic peptides

2.3.12.1 Validation of TSA peptide candidates

Synthetic peptides of TSA and select TAA sequences were obtained from Genscript. Synthetic peptides were solubilized in DMSO to a concentration of 1nmol/μL and all synthetic peptides were combined in a stock solution at a concentration of 10picomol/μL. The stock solution was desalted in aliquots of 150picomol on homemade C18 membrane (Empore) columns and dried using a Speed-Vac. Dried peptide extracts were labeled with a TMT10plex channel as described (see 'TMT labeling' section), desalted, and dried down in Speed-Vac. Labeled synthetic peptides were resuspended in 4% FA and 1picomol of each synthetic peptide was loaded on a homemade C18 analytical column (20 cm x 150 μm i.d. packed with C18 Jupiter Phenomenex) with a 76-minute gradient from 0% to 30% ACN (0.2% FA) and a 600 nL/min flow rate on an EASY-nLC II system. Samples were analyzed with an Orbitrap Exploris 480 spectrometer (Thermo Fisher Scientific) in positive ion mode with Nanoflex source at 2.8kV. Each full MS spectrum was acquired with a 120 000 resolution, and an inclusion list was used to select ions for fragmentation with

40% collision energy and an isolation window of 1 m/z. MS/MS were acquired with a resolution of 30 000. MS/MS correlations were computed as previously described (102). Briefly, expected peptide fragments were computed with pyteomics v4.0.1 (https://bitbucket.org/levitsky/pyteomics) and reproducibly detected peptide fragments were identified. Root scaled intensities of these fragments were correlated between endogenous and synthetic peptide scan pairs and SciPy v1.2.1 (https://www.scipy.org/) was used to compute Pearson correlation coefficient, p-value, and confidence intervals. Mirror plots of the scan pair with the lowest p-value were generated for each peptide using spectrum_utils v0.2.1(https://github.com/bittremieux/spectrum_utils).

2.3.12.2 Relative quantification

To relatively quantify MAPs of interest in primary tissue samples, synthetic peptides at concentrations of 10, 100, or 1000 fmol labeled with TMT 10plex-129N, 130N, and 131N, respectively, were spiked into remaining purified MAPs from NAT and CRC tissue samples labeled with TMT10plex-126 and 127N, respectively. Note that the channel TMT 10plex-127C was left empty to assess contamination. Samples were analyzed with an Orbitrap Fusion Tribrid spectrometer (Thermo Fisher Scientific) in positive ion mode with Nanoflex source at 2.4kV. For synchronous precursor selection MS3 (SPS-MS3), full MS scans were acquired with a range of 300-1000 m/z, Orbitrap resolution of 120 000, automatic gain control (AGC) of 5.0e5, and a maximum injection time of 50ms, using an inclusion list for the peptides of interest. We used a 3s top speed approach for MS2 in the ion trap, with an isolation window of 0.4m/z, collision induced dissociation of 35%, a 'normal' ion trap scan rate mode, 2.0e4 AGC target, and 50ms maximum injection time. This was followed by the selection of eight synchronous precursor ions for MS3 acquisition, which was done with a scan range of 110-500m/z, Orbitrap resolution of 50 000, AGC of 1.0e5, maximum injection time of 300ms, an isolation window of 2.0m/z, and 65% HCD collision energy. LC-MS instrument was controlled using Xcalibur version 4.4 (Thermo Fisher Scientific, Inc). Error tolerances for precursor mass and fragment ions were set to 10.0ppm and 0.5 Da, respectively. A non-specific digest mode was used. TMT10plex was set as a fixed PTM, and variable modifications included phosphorylation (STY), Oxidation (M), Deamidation (NQ), and TMT10plex STY. For quantification, PSMs were filtered to exclude those with contamination in

the TMT10plex-127C channel, and to select those within the 70th intensity percentile. MS2 precursor profiles and intensity profiles of all relevant channels were manually inspected to select peptides for quantification. Intensity ratios for each peptide were calculated using the average TMT10plex-127N and TMT10plex-126 intensities of good quality PSMs.

2.3.13 Data analysis and visualization

Figure 1 was generated with BioRender.com. The majority of other figures were created with Python v3.7.6, R v3.6.3, or Origin (Pro)2019b. R packages include:

Repitope v3.0.1 (https://github.com/masato-ogishi/Repitope) (131),

UpsetR v1.4.0 (https://github.com/hms-dbmi/UpSetR) (132),

GSVA v1.38.2 (https://github.com/rcastelo/GSVA) (133),

ESTIMATE v1.0.13 (https://bioinformatics.mdanderson.org/estimate/) (134).

2.3.14 Experimental Design and Statistical Rationale

To effectively elucidate the MHC I immunopeptidome of colorectal cancer, four CRC cell lines and six samples from human subjects consisting of both matched tumor and NAT were selected. NAT was used as an approximation of healthy tissue, as it is the most effective control for each respective tumor. Since no matched samples were available for cell lines, a pool of eight TEC samples was used in the creation of global cancer databases, to obtain a wide berth of approximate normal RNA expression. All instances of p-values are determined using two-sample t-test, except in the determination of significance for immunogenicity scores, in which case the Mann-Whitney test was used as the data did not have a normal distribution, as determined by the Shapiro test. For t-tests, we performed f-tests to determine whether the dataset had significant variation; if yes, then we used the t-test assuming variation, and otherwise the t-test assuming no variation was used. For CRC-derived cell lines, four technical replicates of 2×10^8 cells were prepared, which were TMT labeled and multiplexed prior to injection. Due to limited tissue material, half of the purified MAPs from primary samples were injected to obtain global immunopeptidomic data, and the remaining sample was used for targeted analysis with synthetic peptides to confirm the sequences and abundance of putative TSAs and select TAAs. To select

high quality PSMs for quantification, those of low intensity or with contamination in an empty TMT channel were excluded. Further, only peptides with favorable MS2 precursor and intensity profiles were quantified.

2.4 Results

2.4.1 Immunopeptidomic analyses using a proteogenomic approach

To determine the composition of the immunopeptidome of colorectal cancer, we analyzed a collection of samples comprised of four colorectal cancer-derived cell lines and six sets of primary adenocarcinoma samples, which consist of matched tumor and normal adjacent tissue (Tables 1 and 2). Paired-end RNA-seq allowed the creation of a global cancer database, consisting of a canonical cancer proteome as well as a cancer-specific proteome for each sample, by generating cancer-specific kmers which, once combined into contigs, are translated into three reading frames to encompass non-canonical sequences from any genomic origin (Figure 8 – green box). mTECs present peripheral antigens in the thymus and mediate the negative selection of auto-reactive T-cells (135). Thus, in the case of CRC-derived cell lines, cancer-specific kmers were obtained following the subtraction of mTEC-derived kmers, which approximated the expression of these sequences in healthy tissues. For the primary tissue samples, the cancer-specific kmers were generated following subtraction of the sequences from matched NAT. This approach enabled the determination of sequences expressed in tumor and not observed in healthy colon tissue of the same individual. In addition to database construction, RNA-seq data were also used for transcriptomic analysis, including GO term analysis, investigation of immune infiltration, mutation profiling, and determination of transcript abundance (Figure 8 – purple box).

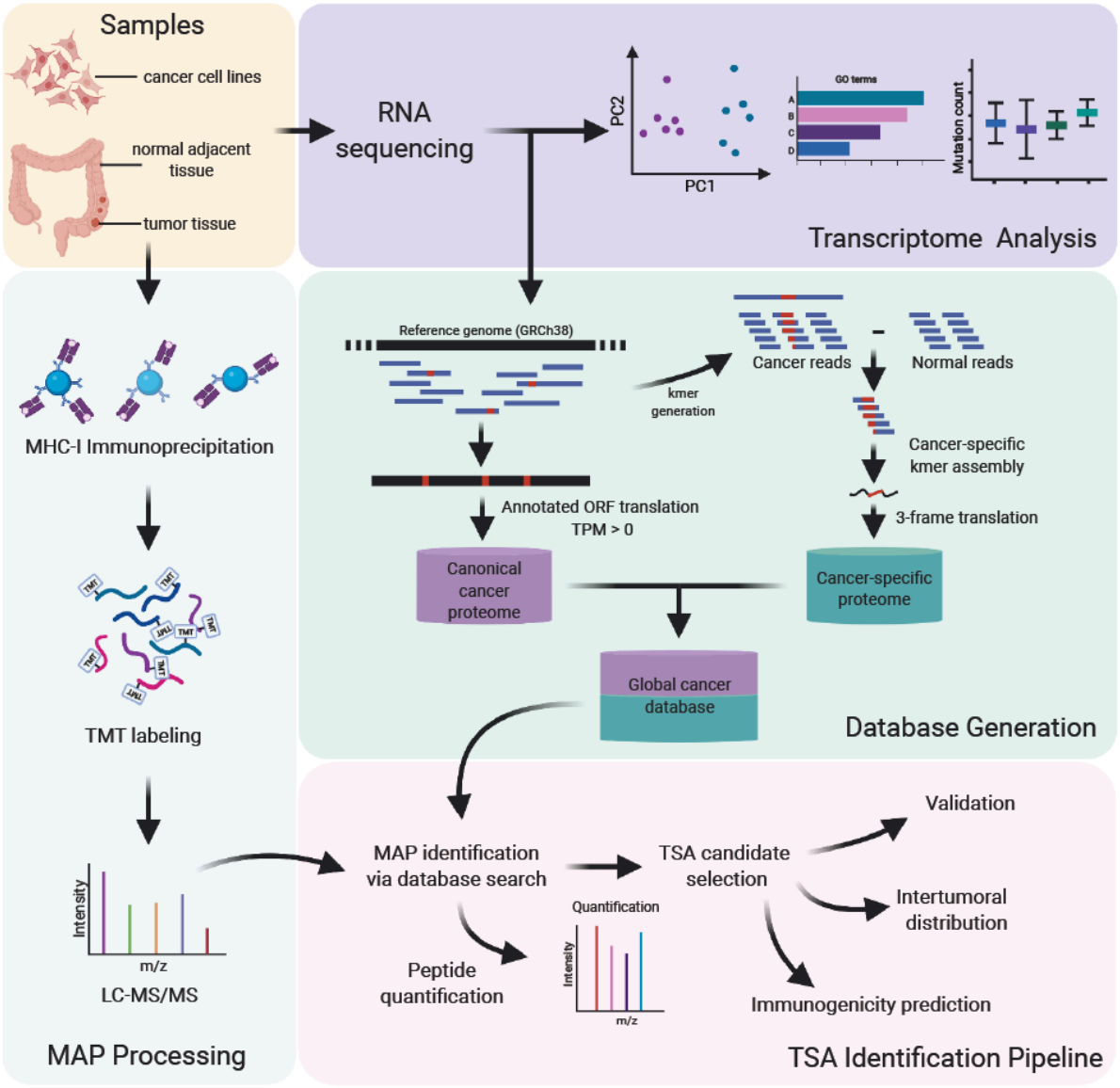

Figure 9. – Proteogenomic workflow for the discovery of tumor-specific antigens (TSAs) in both colorectal cancer (CRC)-derived cell lines and primary tumor samples

Illustration of proteogenomic strategies used to identify TSAs. Samples generated from CRC- and normal intestine-derived cell lines and matching primary tumor/normal adjacent tissue (NAT) biopsies obtained from six individuals were all processed for both RNA sequencing and major histocompatibility complex class I (MHC-I) immunoprecipitation (IP). RNA sequencing data were used for both the transcriptomic characterization of the samples and the generation of customized global cancer proteome databases. For each sample, the MHC-I associated peptides (MAPs) isolated via IP were identified via LC-MS/MS using the respective database. After validating both the identification and the tumor specificity of our TSA candidates, their therapeutic potentials were evaluated though the prediction of both their immunogenicity and inter-tumoral distribution. Created with BioRender.com.

We used immunoprecipitation to isolate MHC I-peptide complexes, and we labeled the eluted MAPs with TMT isobaric labeling reagent, as TMT labeling was recently shown to enhance the detection of MAPs by increasing their charge state and hydrophobicity (Figure 8 – blue box) (88). We then sequenced and analyzed MAPs by liquid chromatography tandem mass spectrometry (LC-MS/MS) and identified using the personalized cancer databases generated through RNA-Seq. Identified MAPs then underwent a rigorous series of classifications and validations to identify putative TSAs and TAAs. Tumor antigens identified in CRC tissues were then validated and quantified with synthetic peptides to determine to what extent they were overexpressed on tumor compared to matched NAT, and we also investigated their predicted immunogenicity and intertumoral distribution to evaluate their clinical potential (Figure 8 – pink box). The TSA and TAA selection process was composed of a stringent set of filters based on expression in cancer and normal tissues (Figure 9).

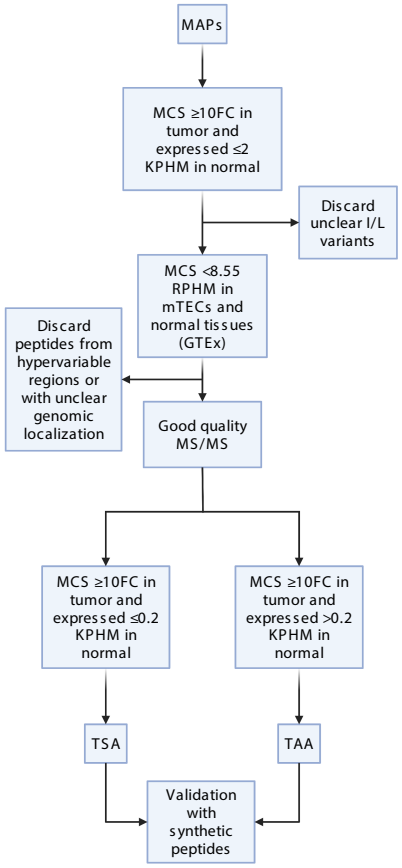

Figure 10. – TA identification flowchart

Flowchart depicting workflow and filters used to identify TSA and TAA candidates, as described in methods. MHC I-associated peptides were selected based on the overexpression of their RNA coding sequences in tumor compared to normal, and their low expression in mTECs and normal tissues. Peptides with from hypervariable regions or with unclear I/L variants or genomic localizations were discarded. Putative TSAs and TAAs were distinguished based on their expression in normal tissues.

We used four CRC-derived cell lines with different HLA alleles and characteristics as summarized in Table 1. HCT116 and RKO are derived from primary tumors and are characterized by MSI, whereas Colo205 and SW620 are derived from metastases of ascites and lymph node, respectively, and are both MSS. Among the four cell lines are mutations in several key genes, such as BRAF, RAS, SMAD4, TP53, and PI3CA. These cell lines have a varying MHC I surface expression ranging from 1.44×10^5 to 5.07×10^5 MHC I molecules/cell, as determined by Qifikit, and a diversity of HLA alleles which were identified using OptiType, an HLA genotyping tool that uses RNA-Seq data to predict a sample's HLA alleles, in combination with the HLA alleles for these cell lines documented in the literature (Table 1) (120).

Table 1. – Description of CRC-derived cell lines

Cell line	Tissue	Morphology	Disease	Biomarkers	MHC I molecule/cell	HLA genotyping	Mutations of interest
Colo205	Colon; derived from metastatic site: ascites	Epithelial	Dukes' type D, colorectal adenocarcinoma	MSS, CIMP	$1.44 \times 10^5 \pm 0.00282 \times 10^5$	HLA-A*01:01 HLA-A*02:01 HLA-B*07:02 HLA-B*08:01 HLA-C*07:01 HLA-C*07:02	BRAF (V600E), SMAD4, TP53
HCT116	Colon	Epithelial	Colorectal carcinoma	MSI, CIMP	$5.07 \times 10^5 \pm 0.30 \times 10^5$	HLA-A*01:01 HLA-A*02:01 HLA-B*18:01 HLA-B*45:01 HLA-C*05:01 HLA-C*07:01	RAS (G13D), PI3CA, CDKN2A, CTNNB1 (B-catenin)
RKO	Colon	Epithelial	Carcinoma	MSI, CIMP	$2.82 \times 10^5 \pm 0.11 \times 10^5$	HLA-A*03:01 HLA-B*18:01 HLA-C*07:01	BRAF (V600E), PI3CA
SW620	Colon; derived from metastatic site: lymph node	Epithelial	Dukes' type C, colorectal adenocarcinoma	MSS, CIN	$1.69 \times 10^5 \pm 0.0017 \times 10^5$	HLA-A*02:01 HLA-A*24:02 HLA-B*07:02 HLA-B*15:18 HLA-C*07:02 HLA-C*07:04	APC, RAS (G12V), SMAD4, TP53

All of the primary tumor samples are derived from stage II adenocarcinomas, which vary only slightly in tumor grade and TNM (tumor-node-metastases) classification (Table 2). The CRC tissue samples had a tumor content of 95-100% and an average mass of 0.6625 g. The tumors all originated from the sigmoid colon, with the exception of S1 (cecum) and S5 (ascending colon). NAT were collected at least 6cm away from the tumor margins. Similar to the cell lines, the tissue samples also possess a variety of HLA alleles. A visualization of the number of HLA alleles unique to or shared by cell line and tissue samples is available in Figure 10. There is an average of 1.3 and 3.2 unique alleles per cell line and tissue, respectively.

Table 2. – Description of primary tumor and matched NAT

Sample ID	Sex	Age	Ethnic background	Matrix	Diagnosis	Histological diagnosis	Stage	Tumor content %	Mutations of interest	HLA
S1_N				colon	NAT					HLA-A*24:02
S1_T	F	73	Caucasian	cecum	cancer	adenocarcinoma	IIC	100	KRAS G12D	HLA-B*07:02 HLA-B*35:01 HLA-C*04:01 HLA-C*07:02
S2_N				colon	NAT					HLA-A*02:01 HLA-A*03:02
S2_T	M	60	Caucasian	Sigmoid	Cancer	Adenocarcinoma	IIA	95		HLA-B*27:05 HLA-B*58:01 HLA-C*02:02 HLA-C*07:01
S3_N				colon	NAT					HLA-A*01:01 HLA-A*32:01
S3_T	F	63	Caucasian	sigmoid	cancer	adenocarcinoma	IIA	100	KRAS Q61H	HLA-B*38:01 HLA-B*50:01 HLA-C*06:02 HLA-C*12:03
S4_N				colon	NAT					HLA-A*01:01 HLA-A*11:01
S4_T	F	85	Caucasian	sigmoid	cancer	adenocarcinoma	IIA	100	KRAS G12D	HLA-B*15:01 HLA-B*57:01 HLA-C*03:03 HLA-C*06:02
S5_N				colon	NAT					HLA-A*03:01 HLA-A*30:01
S5_T	F	43	Caucasian	ascending colon	cancer	adenocarcinoma	IIA	95		HLA-B*13:02 HLA-B*52:01 HLA-C*06:02 HLA-C*12:02
S6_N				colon	NAT					HLA-A*03:01 HLA-A*23:01
S6_T	F	48	Caucasian	sigmoid	cancer	adenocarcinoma	IIA	95		HLA-B*07:02 HLA-B*18:01 HLA-C*07:01 HLA-C*07:02

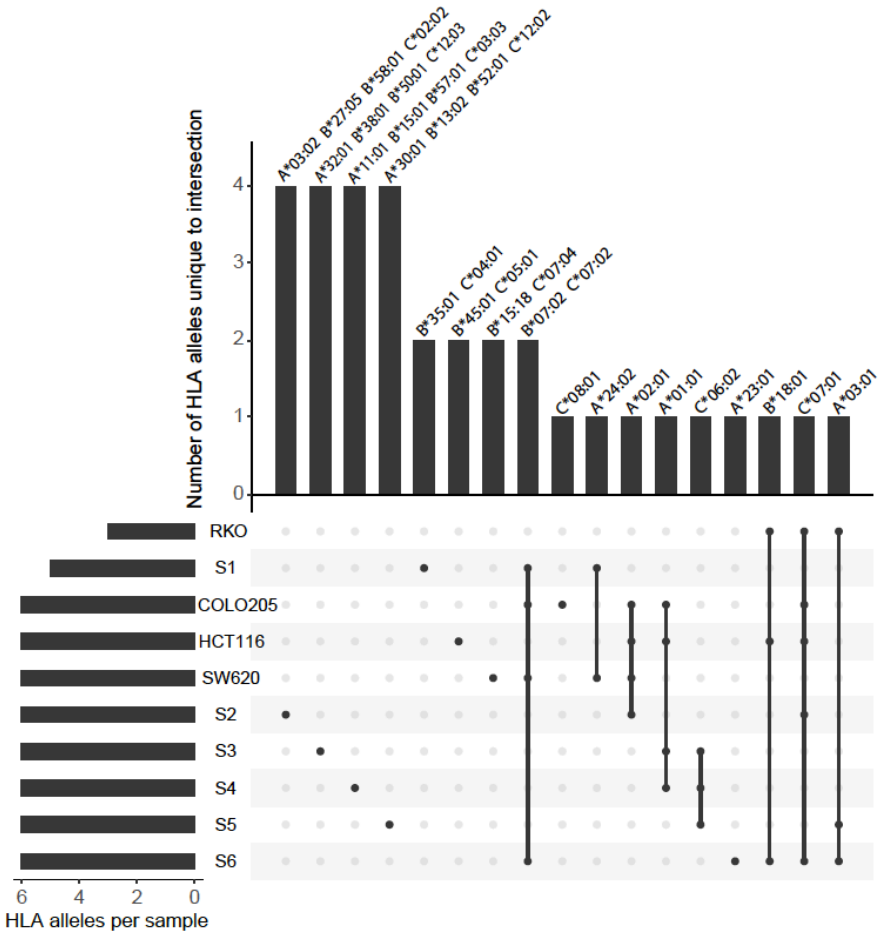

Figure 11. – Upset plot of HLA alleles

UpsetR plot displaying the number of HLA alleles unique to a given intersection of samples, specifically MAPs that are unique to a given sample or that are uniquely shared by two samples.

2.4.2 Transcriptomic analyses reveal heterogeneity between MSI and MSS samples

Because the outcome for CRC patients within a given disease stage differs greatly based on the molecular characteristics of the tumor (136, 137), RNA sequencing data were used to characterize the molecular heterogenicity of the samples. After first examining the mutational status of key biomarkers (such as KRAS, NRAS, or BRAF) which are commonly used to guide therapeutic decisions and prognoses in the clinics (138, 139) (Table 1), the microsatellite statuses of cell lines and primary samples were respectively determined from the literature (140, 141) and using the MSIsensor package (142). While MSI is found in a limited subset of CRC tumors (i.e. 15% of sporadic CRC and 90% of nonpolyposis colorectal cancer) (143), in this study, 50% of the tumorigenic cell lines and 33% of the primary biopsies present this phenotype (Table 3). Although several elements in the literature suggest that MSI and MSS tumors are immunologically different (69, 100, 144, 145), this study will provide the first comparison of MSI and MSS colorectal tumors at the immunopeptidomic level.

Table 3. – MSISensor-pro results for CRC primary tissues

Sample	Number of total sites	Sites with enough coverage	Sites with enough coverage (%)	MSI sites (somatic)	MSI sites (somatic) (%)	Class(MSI > 3.5%)
S1	1011195	111243	11	226	0.20	MSS
S2	1011195	69004	7	197	0.29	MSS
S3	1011195	56779	6	104	0.18	MSS
S4	1011195	45848	5	177	0.39	MSS
S5	1011195	78340	8	8267	10.55	MSI
S6	1011195	60715	6	3085	4.08	MSI

Principal component analysis of the top 500 varying genes between normal and tumor biopsy samples (Figure 11A) or cell lines (Figure 12A) confirms their distinct transcriptomic profile. Accordingly, pathway and process enrichment analysis of biopsy samples revealed a transcriptomic profile enriched in terms associated with their tumorigenic status (Figure 11B). As expected for CRC, the most significantly up- and down-regulated terms are respectively linked to cell proliferation (Figure 11B upper panel) and muscle phenotype and contractility (Figure 11B lower panel). While the enrichment of terms related to proliferation and cell cycle is a general hallmark of cancer (55, 56), the downregulation of muscle-related pathways is inherent to CRC and results from the functional dichotomy between poorly differentiated tumor areas and highly contractile NAT. In contrast, inter-tumoral transcriptomic differences were mostly explained by the MSI/MSS status of the tumor samples of our datasets (Figure 11A and Figure 12A). While MSI samples tend to cluster tightly together, MSS tumors appear more dispersed and therefore transcriptionally more heterogeneous. Functionally, when analyzed separately, the MSS and MSI CRC samples are enriched in very different gene sets. When compared to their corresponding NAT, MSI tumors are characterized by a significant up-regulation of various immune-related terms (Figure 12B) whereas MSS tumors are more associated with an increased expression of genes related to both Wnt and PI3K-Akt signaling (Figure 12C). Although the link of these two signaling pathways with CRC is well established in the literature (146), no reference could be found to support that their contribution in CRC may differ between MSS and MSI tumors.

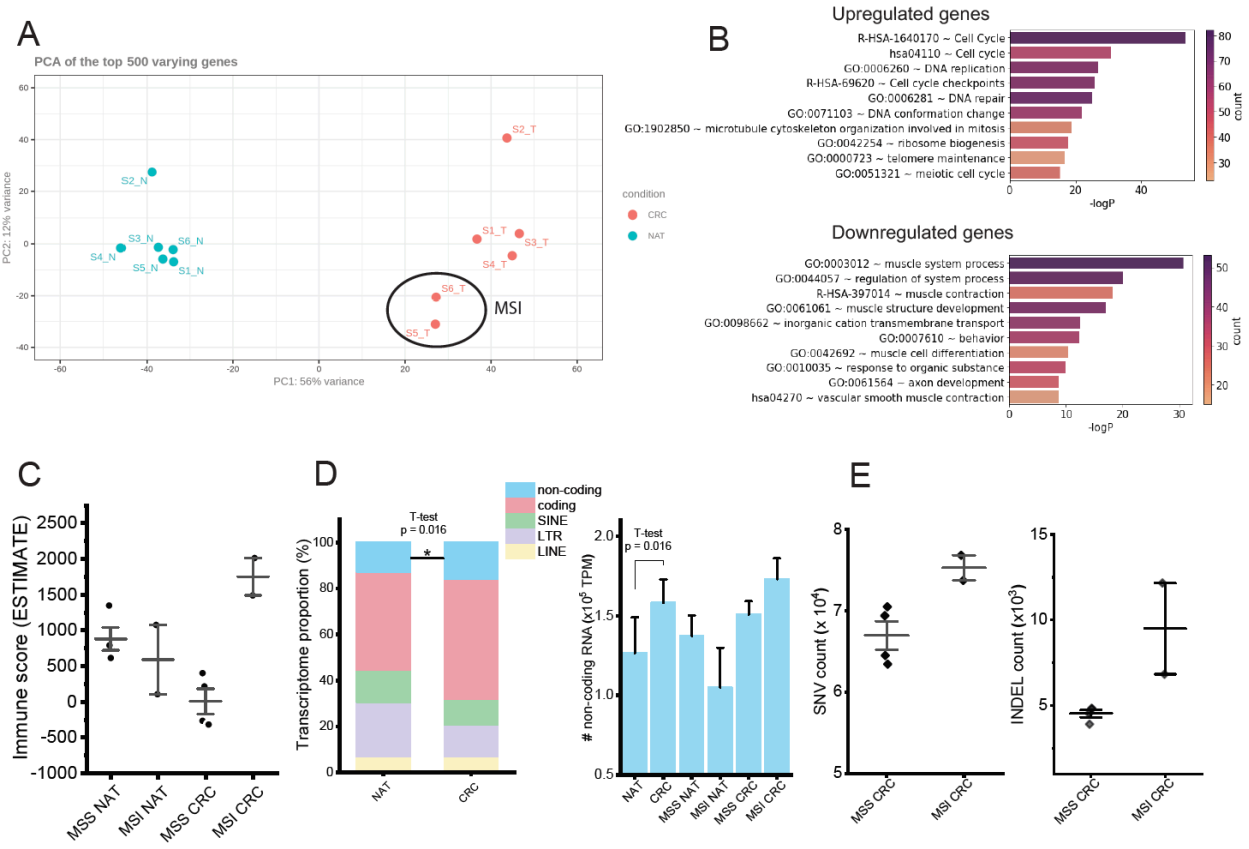

Figure 12. – Transcriptomic profile of primary tumor/normal adjacent tissue CRC biopsies

A) Principal component analysis (PCA) of the top 500 varying genes of each tumor/NAT sample following paired-end RNA seq and gene readcount normalization with DESeq2. MSI tissues (as determined by MSISensor) are encircled. B) GO term analysis of genes up/downregulated in CRC tissues compared to their adjacent NAT. Genes submitted to GO term analysis were those with |log2FC| >1 and that were found to be differentially regulated in all samples, using TPM normalized values. C) Bar graph showing the mean ESTIMATE immune score of MSS NAT, MSI NAT, MSS CRC, and MSI CRC, with standard deviation shown. D) Stacked bar graph showing the mean proportion of the transcriptome attributable to five distinct transcript biotypes in NAT vs CRC samples, with the differences in the proportion of non-coding transcripts being statistically significant between NAT and CRC (non-coding: p = 0.016; coding: p = 0.078; SINE: p = 0.15; LTR: p = 0.056; LINE: p = 0.95). E) Scatterplots displaying the SNV counts and INDEL counts of MSS and MSI CRC tissues determined by SNPEff genomic annotation, with mean and standard error bars.

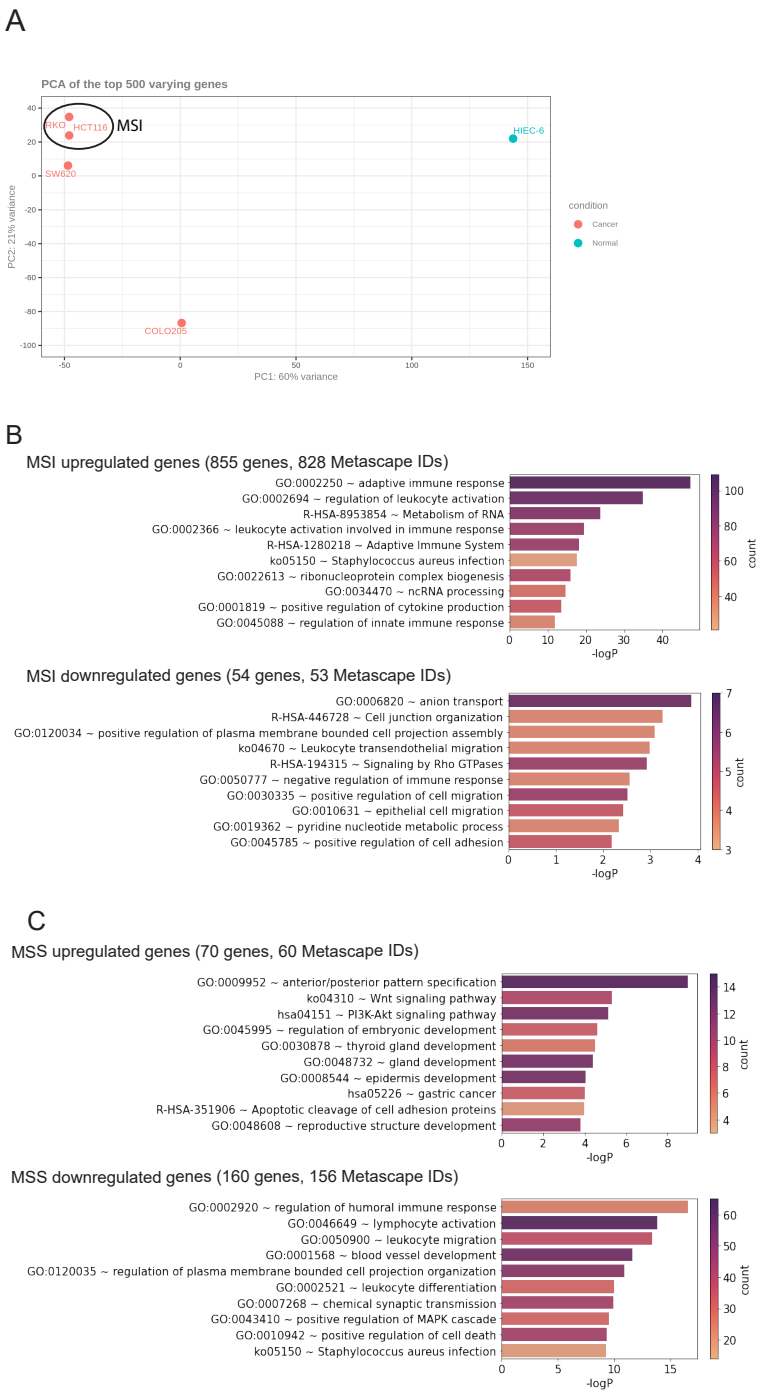

Figure 13. – Transcriptomic profile of CRC-derived cell lines and GO term analysis of MSI and MSS primary tissue samples

A) Principal component analysis (PCA) of the top 500 varying genes of CRC-derived cell line and one normal intestinal cell line (HIEC-6) following paired-end RNA seq and gene read count normalization with DESeq2. Known MSI cell lines are encircled. B) GO term analysis of genes up/downregulated in MSI tumors compared to their adjacent NAT. Genes used for GO term analysis were those with |log2FC| >1 when compared to their respective NAT using TPM

normalized values and that were found to be uniquely differentially expressed in both of the MSI tumor samples (i.e. genes that were only up/downregulated in both of the MSI tumors but not any of the MSS tumors). C) GO term analysis of genes up/downregulated in MSS tumors compared to their NAT. Genes used for GO term analysis were those with |log2FC| >1 when compared to their respective NAT using TPM normalized values and that were found to be uniquely differentially expressed in three or more of the MSI tumor samples (i.e. genes that were up/downregulated in at least three MSS tumors but neither of the MSI tumors).

Next, we estimated the degree of immune infiltration of each sample via two independent approaches using the immune infiltration score from the ESTIMATE package, which has been shown to effectively predict tumor purity when compared with histological analyses (134) (Figure 11C), and with an enrichment score for known TIL markers (147) based on a single-sample Gene Set Enrichment Analysis (ssGSEA) (148) (Figure 13A). While all NAT samples presented similar levels of immune infiltration, MSI and MSS tumors were respectively characterized by increased and decreased immune infiltration scores (Figure 11C and Figure 13A). Consistent with what has been previously reported in the literature (144, 149-152), such differences suggest that MSI tumors may be more immunogenic than their MSS homologs.

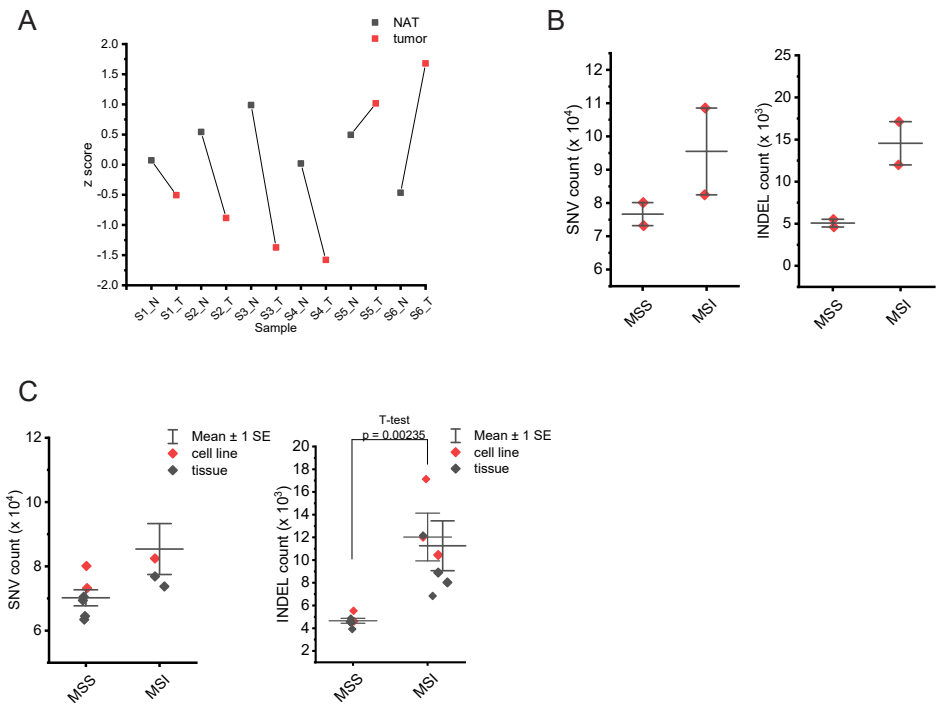

Figure 14. – ssGSEA analysis of immune infiltration in CRC tissues and mutation profile of all samples

A) ssGSEA analysis of immune infiltration in tumor and matched NAT using expression markers for tumor-infiltrating lymphocytes described in Danaher et al. 2017, and the Gene Set Variation Analysis (GSVA) R program (https://github.com/rcastelo/GSVA) to estimate immune cell enrichment in these tissues. B) Scatterplots displaying the SNV counts and INDEL counts of MSS and MSI CRC-derived cell lines determined by SNPEff genomic annotation, with mean and standard error bars. C) Scatterplots displaying the SNV counts and INDEL counts of all MSS and MSI samples (cell lines and tissues), determined by SNPEff genomic annotation, with mean and standard error bars. (SNV: p = 0.062; INDEL: p = 0.0024).

Because TSAs can arise from a wide range of cancer-specific events/dysregulations (100, 114) and the immunopeptidome contribution of each antigenic source varies significantly across malignancies (100), RNA-seq data were also used to inform which TSA classes might be enriched in our samples. By examining the genomic origin of the transcripts, we observed that both the proportion and the absolute abundance of non-coding polyadenylated RNAs are significantly increased in tumors compared to NATs (Figure 11D). While on average the absolute abundance increase is limited to 25%, our data suggest that the tumor-specific gain of non-coding transcripts could be higher in MSI tumors than in MSS. Although this comparison remains limited due to the low number of MSI samples (n=2), one could expect to identify a higher number of aeTSA deriving

from non-coding transcripts in MSI samples compared to MSS. As well, both the SNV burden and the INDEL burden are notably increased in MSI samples compared to MSS (an average difference of 8326 and 4965 between MSI and MSS mean SNV and INDEL burdens, respectively), an observation that is also noted for cell lines (Figure 11E and Figure 13B). Considering both cell line and tissue samples together resulted in a statistically significant difference in the number of INDEL mutations between MSS and MSI samples ($p = 0.0024$) (Figure 13C). Because both MSI and INDEL accumulation result from defects in the MMR pathway (153), one can hypothesize that the number of INDEL-derived TSAs (most likely frameshift-derived antigens) identified in a sample will be proportional to its MSI level.

2.4.3 Immunopeptidomic analyses highlight the diversity of CRC antigens

To elucidate the MHC I immunopeptidomes of CRC-derived cell lines and tissues, we immunoprecipitated MAPs from four replicates of 2×10^8 cells for each cell line and from each tissue sample. We then derivatized each replicate with a separate TMT6plex channel (channels 126, 127, 128, 129) for cell lines or with TMT10plex-126 and -127N for primary NAT and tissue samples, respectively. The four replicates of each cell line, and half of the respective NAT and tumor MAPs from each subject, were multiplexed and analyzed by LC-MS/MS. The median labeling efficiencies were 72.4% or 87.8% for cell lines and tissue samples, respectively. We ascribe the lower efficiency of labeling in cell lines to meager MAP yields. We identified 5281 and 27 583 unique MAPs in the cell line and tissue datasets, respectively, with a mean of 1433 unique MAPs per cell line and 5855 per tissue (Figure 14A, upper panel, and 14B; Supplementary files 1 and 2). While the identification varied between each line, the number of MAPs identified was strongly correlated with the abundance of MHC I molecules per cell (Figure 14A, lower panel; Pearson's $r = 0.96$).

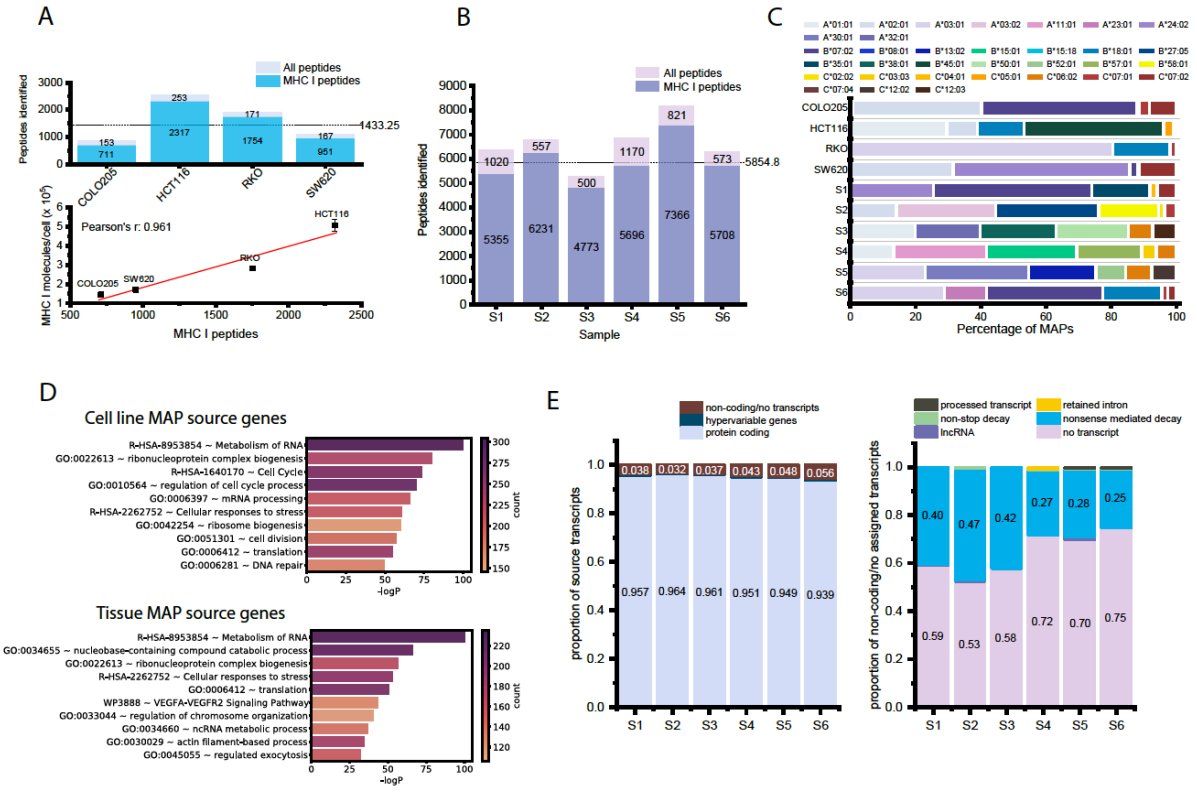

Figure 15. – Immunopeptidomics of CRC-derived cell lines and tissues

A) Top panel: Stacked bar chart displaying the number of unique peptides identified in CRC cell lines, and a horizontal line indicating the average number of MAPs per cell line. Bottom panel: Scatterplot indicating the correlation between the number of unique MAPs identified in each cell line and the presentation of MHC I at the cell surface (Pearson's r = 0.96). B) Stacked bar chart displaying the number of unique peptides identified in primary tissue samples, and a horizontal line indicating the average number of MAPs per tissue sample. 'All peptides' in A) and B) indicates the number of peptides identified with a 5% FDR, while 'MHC I peptides' indicates the number of peptides identified with the corresponding peptide score, 8-11 amino acids in length, and a rank eluted ligand threshold $\leq$ 2% using netpanMHC4.1b predictions. C) Bar chart indicating the proportion of unique MAPs predicted to bind to a given HLA allele in each sample, using NetMHCpan-4.1b predicted affinity. D) GO term analysis of MAP source genes for CRC-derived cell lines and primary tissues. For tissues, only source genes shared by four or more tissues were included in this analysis. E) Left panel: Stacked bar chart displaying the proportion of MAPs in each tissue sample derived from protein-coding, hypervariable gene (immunoglobulin or TCR), or non-coding transcripts, or those from unannotated transcripts. Right panel: stacked bar chart displaying the proportion of non-coding MAPs derived from processed transcripts, retained introns, nonstop decay products, nonsense mediated decay products, lncRNA, or those that have no annotated transcript.

When taking the cell line and tissue samples together, we identified a total of 30 485 unique MAPs. Within the MAP repertoire of each sample, 32-68% of the peptides are sample-specific, and very few shared MAPs were observed when comparing only cell line or primary samples (Figure 15A and B). This large proportion of unique MAPs can be attributed to the diversity of HLA alleles among our samples, which is a major factor influencing which peptides can be presented at the cell surface (Figure 14C; Figure 10). On average, the number of MAPs shared by any two cell lines or any two tissue samples is 59 or 640 MAPs, respectively. There are noteworthy outliers – tissue samples S1 and S6 shared 2079 MAPs (1673 of which are unique to these samples (Figure 15C)), thus their MHC I immunopeptidomes have approximately 23% similarity, as measured by the Jaccard index. (Figure 15D). The next closest similarity in MAP repertoires between two tissues is 1328 MAPs shared by the two MSI tissues (S5 and S6), which is only an 11% similarity between their immunopeptidomes. Despite the decreased MAP identification in cell lines, these trends are reproduced. For example, HCT116 and RKO share the most MAPs, though these peptides represent only 4% similarity, and this is likely a feature of their larger peptide repertoires (Figure 15C). In contrast, COLO205 and SW620 share 152 MAPs, approximately 10% similarity.

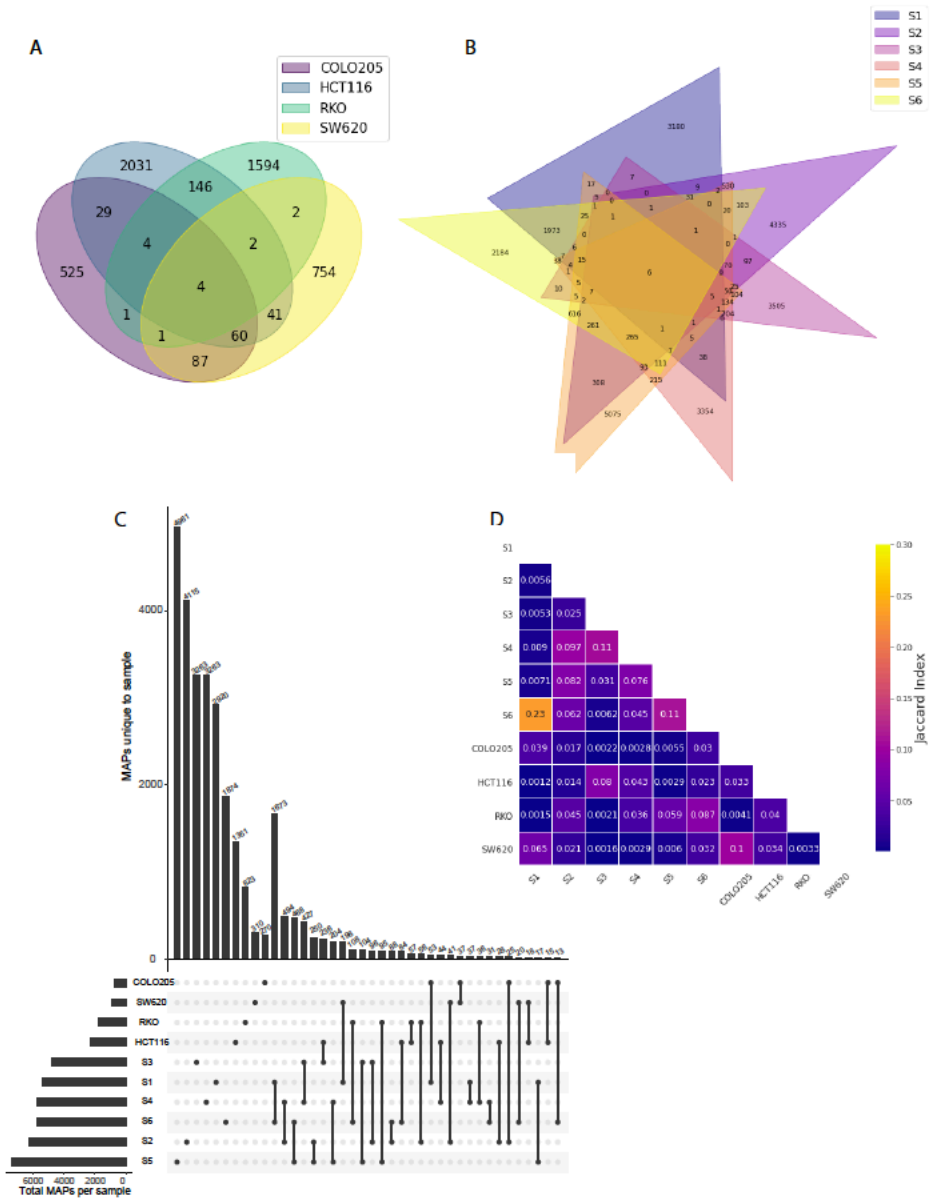

Figure 16. – Overview of unique and shared MAPs in CRC-derived cell line and CRC/NAT tissue samples

A) Venn diagram displaying the overlap of MAPs in the MHC I immunopeptidomes of four CRC-derived cell lines. B) Venn diagram displaying the overlap of MAPs in the MHC I immunopeptidomes of six primary tissue samples. C) UpsetR plot displaying the number of MAPs unique to a given intersection of samples, specifically MAPs that are unique to a given sample or that are uniquely shared by two samples. D) Heatmap demonstrating the Jaccard index of MAP similarity between any two cell line or tissue samples.

To contextualize these comparisons, we can again consider the HLA alleles of our samples. Out of the 2079 MAPs shared by S1 and S6, 1595 MAPs are predicted to bind the same allele HLA-B*07:02 in approximately 90% of these cases (Supplementary File 2). In addition, the S1 allele HLA-A*24:02 and the S6 allele HLA-A*23:01 have very similar allele-binding motifs as shown by HLAthena (154). Similarly, 126 out of 152 MAPs shared by COLO205 and SW620 are bound by the allele HLA-A*02:01 for 94 MAPs (Supplementary File 1). Thus, the MHC I immunopeptidomes of our samples is majorly influenced by the HLA repertoire.

At the gene level, we identified peptides derived from over 8000 unique source genes, with an average of 1014 and 3168 source genes per cell line and tissue sample, respectively (Figure 16A, upper panel). This was highly correlated with the number of MAPs identified (Figure 16A, lower panel). Roughly 6-14% of the source genes in a given immunopeptidome were sample-specific (Figure 16B), which could be attributed to sample-specific biological features or it could reflect an imperfect sampling of the immunopeptidome (Figure 16E). We do not expect to identify every MAP presented at the cell surface, and as a majority of source genes in each sample are attributable to only a single MAP (Supplementary files 1 and 2), it is almost certain that additional source genes contribute to the MAP repertoire and their corresponding peptides are simply not detected. When comparing any two tissue samples, they had on average 32% source gene similarity, while comparing any two cell lines resulted in an average of 13% shared gene similarity (Figure 16C-E). Thus, distinct cell lines appear to be less homogenous than tissue samples at the source gene level. This likely reflects differences in sample composition, as the tissue samples have source genes derived from NAT, stroma, infiltrating cells, etc, while cell lines consist of only a single cell type. In addition, lower MHC I presentation of cell lines and the resulting decreased identification of MAPs means fewer source genes were sampled, lowering the likelihood of overlap. Regardless, all samples are more similar at the source gene-level compared to the immunopeptidome level, and sample-specific MAPs are being derived from shared source genes.

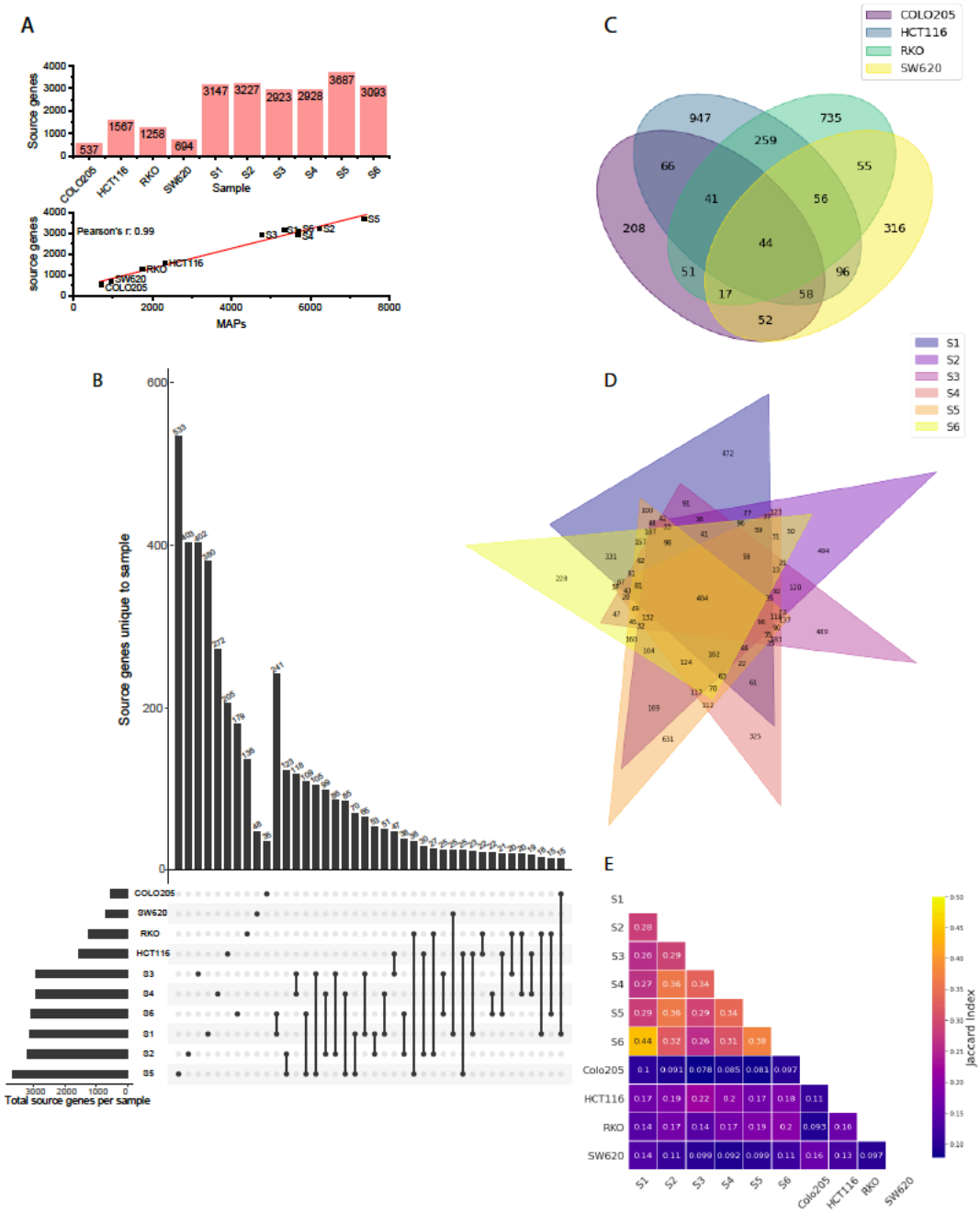

Figure 17. – Overview of unique and shared MAP source genes in CRC-derived cell line and CRC/NAT tissue samples

A) Top panel: Bar chart displaying the number of unique source genes identified per sample. Bottom panel: Scatterplot indicating the correlation between the number of unique MAPs identified in each sample and the corresponding number of unique source genes (Pearson's r = 0.99). Source genes were identified for peptides from coding sequences; any peptide that mapped to more than one source gene was excluded. B) UpsetR plot displaying the number of source genes

unique to a given sample or intersection, specifically source genes that are unique to a given sample or that are uniquely shared by two samples. C) Venn diagram displaying the overlap of source genes in the MHC I immunopeptidomes of four CRC-derived cell lines. D) Venn diagram displaying the overlap of source genes in the MHC I immunopeptidomes of six primary tissue samples. E) Heatmap demonstrating the Jaccard index of source gene similarity between any two cell line or tissue samples.

To obtain an overview of the genomic function of the MHC I immunopeptidome and investigate the overlap of source genes, we performed GO term analysis on all the source genes identified in the cell lines as well as those identified in four or more tissues. Several common features between cell lines and tissues are detectable at the immunopeptidome level, including a significant enrichment of genes involved in RNA metabolism, ribonucleoprotein complex biogenesis, translation, and cellular responses to stress (Figure 14D). Thus, despite the large diversity of HLA alleles between and among our cell lines and tissue samples and the low MAP identification in cell lines, there is significant similarity in terms of what genes are contributing to the MHC I immunopeptidome.

To investigate what proportion of MAPs from our tissue samples were from non-coding transcripts, we first determined, for each peptide, the most abundant putative source transcript (Ensembl Annotation 99). For peptides from the cancer-specific database, we mapped the MCS onto the genome, and determined the most expressed transcript at that location (see 'Quantification of MAP coding sequences in RNA-Seq data' in Methods section). We thus determined that on average, 95.3% of our MAPs from tissue samples were from protein coding transcripts (i.e. UTR or CDR) (Figure 14E, left panel). Approximately 4.2% of peptides are from non-coding regions if we include the 2.8% of peptides deriving from unannotated RNA transcripts, as these peptides are likely coming from intergenic sequences. Approximately one-third of all noncoding MAPs (including those from unannotated transcripts) are derived from nonsense-mediated decay transcript products, while less than 1% of them are coming from lncRNA, nonstop decay products, retained introns, or processed transcripts (transcripts that do not contain open reading frames) (Figure 14E, right panel).

2.4.4 Identification of tumor-specific and tumor-associated antigens in CRC

Following the identification of over 30 000 unique MAPs, we filtered peptide coding sequences to select those overexpressed at least 10-fold in cancer and expressed ≤2KPHM in pooled TEC samples or matched NAT, for cell lines and primary samples, respectively. A recent immunopeptidomic study in AML demonstrated that MCSs with RPHM < 8.55 have less than 5% probability to generate MAPs (103). We thus quantified the expression of the MCSs in RNA-Seq data and kept only those that were expressed below 8.55 RPHM in mTECs and other normal tissues (GTEx). Following manual validation of the remaining peptides, we classified peptides as mTSAs if their amino acid sequence contained a cancer-specific mutation (*i.e.* not an SNP). MAPs for which the sequence was the same as the reference genome and overexpressed at least 10-fold in tumor compared to normal were classified as aberrantly expressed TSAs (aeTSAs) if they had no or residual RNA expression (≤0.2 KPHM) in mTECs (and NAT in the case of tissues) or as TAAs if their expression in mTEC and/or NAT was greater than 0.2 KPHM.

While the TSA yield in CRC-derived cell lines was relatively meager, possibly due in part to low MAP identification, we uncovered an average of three TSAs per primary tissue sample (Figure 17A). Overall, we identified one putative TSA in a CRC-derived cell line and 18 putative TSAs in primary tissues, and the TSA yield from each sample was correlated with the number of MAPs identified (Pearson's r = 0.76) (Supplementary Figure 18). Of these, approximately one-third were derived from coding regions, while the majority of the putative TSAs identified originated from non-coding regions (Figure 17B). Among the TSAs from coding regions, two were from non-canonical reading frames, deriving from exon frameshift sequences, and another two were mutated TSAs identified in MSS tissues S2 and S3 (Figure 17A and 17B). Among the non-coding TSAs, a large proportion originated from intronic or intergenic regions, with a smaller number being derived from 5' UTR, 3' UTR, or lncRNAs (Figure 17B). The sequences of six aeTSAs (four introns, one intergenic, one lncRNA) overlapped ERE sequences (Supplementary file 3). Due to the ubiquitous nature of EREs, TSAs derived from aberrant ERE expression are potentially shared by tumors and have been shown to be immunogenic (155, 156). Of note, none of our putative TSAs were shared between samples, even those with a high proportion of shared MAPs. However, we did identify two unique TSAs in different tissues that were derived from the same transcript

of COL11A1 (one exon frameshift and one 5’ UTR), which was recently shown to play a role in CRC development and prognosis (157). The majority of other TSA source genes have also been shown to be biologically relevant in CRC (Table 4).

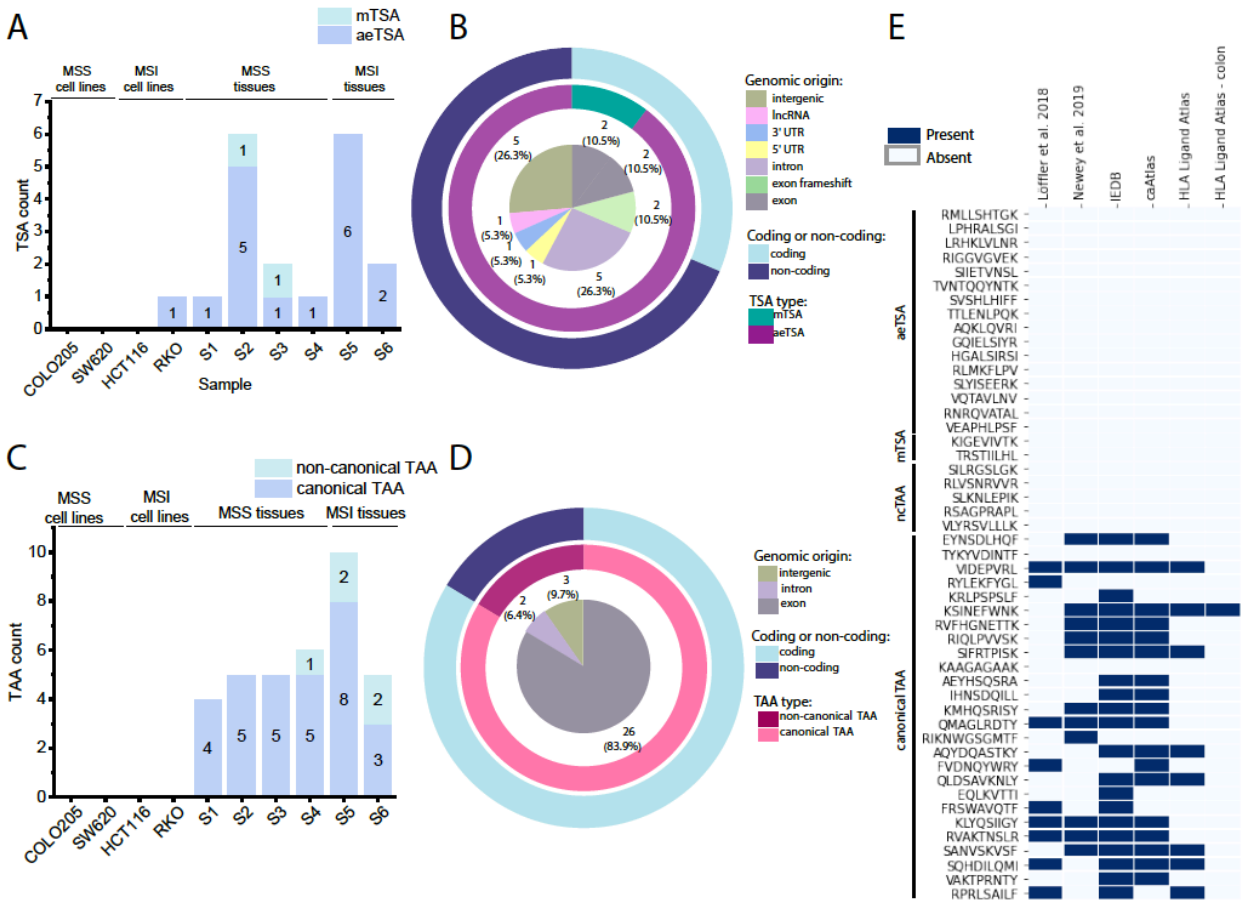

Figure 18. – Novel TSAs identified in CRC derive primarily from non-coding regions, while the majority of TAAs derive from exons

A) Bar chart displaying the number of TSAs identified per sample. B) Stacked pie chart identifying the genomic origin of TSAs in the inner pie, as well as what proportion of TSAs are mutated in the middle pie. The outer pie demonstrates what proportion of TSAs are from coding or non-coding sequences. C) Bar chart displaying the number of TAAs identified per sample. D) Stacked pie chart identifying the genomic origin of TAAs in the inner pie, and what proportion of TAAs are canonical or non-canonical in the middle pie. The outer pie displays what proportion of TAAs are from coding or non-coding sequences. E) Heatmap displaying the presence or absence of putative TSAs and TAAs in two previous publications on CRC immunopeptidomics (Löffler et al. 2018 and Newey et al. 2019), caAtlas, IEDB, and HLA Ligand Atlas (all tissues, and only colon tissue).

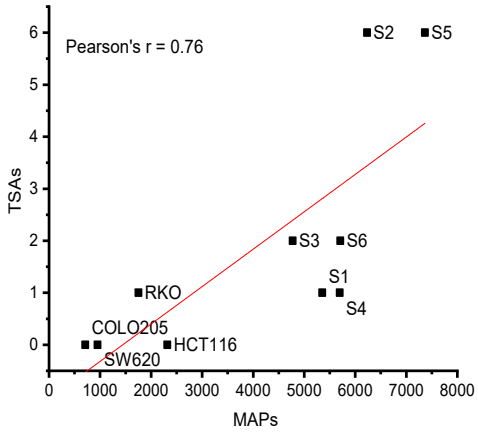

Figure 19. – Correlation of MAPs and TSAs

Scatterplot indicating the correlation between the number of unique MAPs identified in each sample and the number of TSAs identified and validated (Pearson's r = 0.76).

Table 4. – Biological relevance of TSA source genes in CRC

Source gene	Reference	Biological relevance in CRC
COL11A1 - Collagen type XI alpha 1	PMID: 33597969	Upregulated in CRC (mRNA), marker of poor prognosis, role in CRC development
CYP39A1 - ytochrome P450, family 39, subfamily A, polypeptide 1	PMID: 27341022	Expression is increased in CRC with poor prognosis
DPH6 - Diphthamine biosynthesis 6		No known association
GRIN2B - Glutamate ionotropic receptor NMDA type subunit 2B	PMID: 27243824	Identified as non-driver hub gene involved in progression to stage II CRC
HKDC1 - Hexokinase domain-containing protein 1	PMID: 30005951	HKDC1 contributes to increased metabolism, proliferation, and metastasis of CRC cells
HSPD1 - Heat shock protein family D (Hsp60) member 1	PMID: 28261350; PMID: 29246022	Differentially expressed in CRC, potential biomarker for diagnosis; Exosomal HSPD1 identified as putative diagnostic and prognostic biomarker in CRC
IPP (KLHL27) - Intracisternal A particle-promoted polypeptide	Human Protein Atlas (PMID: 28818916)	Favorable prognostic marker in colorectal cancer; unfavorable in renal and liver cancers
LY6G6F-LY6G6D readthrough - Lymphocyte antigen 6 family member G6F and G6D	PMID: 26894861	LY6G6D/F overexpressed in CRC, potential cell surface marker
NKD1 - Naked cuticle homolog 1	PMID: 25446263; PMID: 19956716	Negative feedback regulator of Wnt pathway, intestinal tumor marker in mice; mutations in NKD1 alter Wnt signaling
PATJ - PALS1-associated tight junction protein		No known association
PLK1 - Serine/threonine-protein kinase PLK1 / polo-like kinase 1	PMID: 22648245	Overexpressed in CRC, associated with metastasis and invasion
SUCNR1 – Succinate receptor 1	PMID: 32365557	SUCNR1 activation induces Wnt ligand expression and activates WNT signaling and EMT in a CRC-derived cell line
TRPC6 - Transient receptor potential cation channel subfamily C member 6	PMID: 26422106	mRNA expression of TRPC6 lower in CRC than in normal tissue, may contribute to tumorigenesis

While our primary objective was to identify putative TSAs in CRC, we also identified an average of 5.2 TAAs in our CRC tissue samples, though none were identified in our CRC-derived cell lines (Figure 17C). In contrast to the primarily non-coding putative TSAs, the majority of the TAAs we identified were from canonical, exon-coding sequences, with only a small number being derived from introns or intergenic sequences (Figure 17D). Two non-canonical TAAs overlapped ERE sequences (Supplementary file 3). Of note, four separate TAAs were identified in more than one sample. These shared TAAs were all derived from canonical exons, with source transcripts originating from ASPM, MKI67, MMP12, and HI-5, all of which have documented associations with cancer (Table 5).

Table 5. – Biological relevance of TAA source genes in CRC

Source gene	Reference	Biological relevance in CRC
ASPM -Abnormal spindle microcephaly associated	**PMID: 31966766; Human Protein Atlas (PMID: 28818916)**	**Overexpressed in CRC; suggested to be unfavorable prognostic marker (involved in mitosis, cell cycle, tumorigenesis); known to be unfavorable prognostic marker in liver, lung, endometrial, pancreatic cancers**
BUB1 - Mitotic spindle checkpoint kinase	PMID: 23747338; PMID: 11782350	Mutations in BUB1 linked to early onset CRC; inactivation may drive metastasis and progression in CRC
CDCA8 - Cell division cycle associated 8	PMID: 25260804	overexpressed in CRC, associated with cancer progression
CENPE - Centromere-associated protein E		No known association
CENPF – Centromere protein F	**PMID: 30550624**	**phosphorylation changes associated w CRC malignancy; unfavorable prognostic marker in other cancers (liver, renal, etc; human protein atlas)**
DIAPH3 - Diaphanous related formin 3	Human Protein Atlas (PMID: 28818916)	DIAPH3 is prognostic, high expression is favorable in colorectal cancer
FANCA - Fanconi anemia group A protein	**PMID: 27165003; PMID: 21286667**	**Fanconi anemia predisposes certain cancers; genes in FA pathway participate in CRC pathogenesis (involved in HR repair)**
HI-5 - H1.5 linker histone, cluster member	PMID: 16959974	Frequently mutated in CRC
IDO2 - Indoleamine 2,3-dioxygenase 2	PMID: 18418598	Upregulated expression in CRC
MACC1 - Metastasis-associated in colon cancer 1	PMID: 27424982; PMID: 25003996	Promotes growth and metastasis of colorectal cancer; associated with carcinogenesis through B-catenin signaling and EMT transition
MCM10 - Minichromosome maintenance 10 replication initiation factor	PMID: 32597491	Decreased mRNA expression in colon and rectal adenocarcinoma samples compared to normal tissues
MGAM2 – Maltase glucoamylase 2	PMID: 30996822	Expressed in GI cancers (TCGA data)
MKI67 – Marker of proliferation Ki-67	**PMID: 26281861; PMID: 27855388;** **PMID: 30727976; PMID: 33658388**	**Favorable prognostic marker in CRC, IHC staining (2016); favorable prognostic marker in stage III and IV CRC, IHC staining (2016); poor prognostic marker in CRC based on database meta-analysis (2019); Ki-67 expression important for tumorigenesis**
MMP12 - Matrix metallopeptidase 12	**PMID: 27431388**	**Overexpressed in CRC compared to control, negative prognostic marker in CRC**
NOS2 – Nitric oxide synthase 2	Human Protein Atlas (PMID: 28818916)	Cancer enhanced (colorectal cancer); RNA data
SPC25 (kinetochore protein)	**PMID: 32351050; Human Protein Atlas (PMID: 28818916)**	**Highly expressed in CRC (among other cancers); unfavorable prognostic marker in liver cancer, endometrial cancer, and lung cancer**
ZNF215 – Zinc finger protein 215	Human Protein Atlas (PMID: 28818916)	Cytoplasmic expression in subsets of immune cells, most abundant in gastrointestinal tract and lymphoid tissues (protein data)

Bold = validated

We initially expected to identify an above-average number of both TSAs and TAAs in MSI tissues. This was the case in S5, however the same was not true for the other MSI tissue (Figures 17A and 17C). This could be due to S6 having a lower 'degree' of instability, as reflected in the MSIsensor-pro results (Table 3). Further, the sample that had the highest number of identified TSAs was S2, an MSS tissue. Thus, the yield of TSAs and TAAs per sample seems to be irrespective of MSI status and may be due to other unique biological features of the tumor outside the scope of this study.

To determine if any of our putative TSAs or TAAs have been previously identified, we verified if the peptide sequences were reported in the Immune Epitope Database, caAtlas (158), the HLA Ligand Atlas (159), and two previous publications that sought to identify tumor antigens in CRC from Löffler *et al.* 2018 (111) and Newey *et al*. 2019 (104). Of note, none of the putative aeTSAs, mTSAs, or non-canonical TAAs were previously reported in any of these resources. Of the 26 putative canonical TAAs identified, 24 of them were reported either in the Immune Epitope Database (IEDB), caAtlas, Löffler et al. 2018, Newey et al. 2019, or some combination of the four (Figure 17E). Eight of these were also reported in the HLA Ligand Atlas, with one of them specifically being documented in healthy colon tissue. Interestingly, none of the TAAs previously identified in these earlier publications were reported as tumor antigens, and, conversely, six of the 12 tumor antigens of interest reported in Löffler et al. were also identified in the immunopeptidomes of our work, though they did not pass our TSA or TAA selection criteria, most often due to high expression in NAT (Table 6). We have thus identified novel TSAs in colorectal cancer that derive primarily from non-coding regions, as well as a selection of mainly coding TAAs, some of which have been previously reported as MAPs.

Table 6. – Justifications for exclusion of Löffler et al. 2018 tumor antigens

Peptide	Löffler et al. 2018 classification	Löffler et al. 2018 HLA restriction	Sample	Reason for exclusion
RLASRPLLL	Vaccine candidate	B*07	S5	FC < 10 between cancer and NAT
YRNSYEIEY	Vaccine candidate	C*07	S1, S2, S6	MCS expression > 2 KPHM in NAT
APTPARPVL	Vaccine candidate requiring further validation	B*07	S1	MCS expression > 8.55 RPHM in mTEC
RLAEPSQMLK	Vaccine candidate requiring further validation	A*03	S6	MCS expression > 2 KPHM in NAT
SPKATGVFTTL	Vaccine candidate requiring further validation	B*07	S1, S6	MCS expression > 2 KPHM in NAT
SVLTQPPSV	Vaccine candidate of unclear relevance	A*02	S2	MCS expression > 2 KPHM in NAT

2.4.5 RNA expression of putative tumor-specific and tumor-associated antigens

First, we investigated the expression, in TPM, of the source transcripts in their respective tumor samples compared to the matched NAT, as well as the mean average of that transcript in the CRC/NAT sample (Figure 19A). This analysis naturally does not include peptides derived from intergenic regions. Note that in Figure 19A, both S4 and S5 plots have a canonical TAA point that is not visible, as it overlaps with another canonical TAA source transcript; however, these sequences were still included in downstream analyses. The average log2FC for the source transcripts of our putative TSAs and TAAs in the samples in which they were identified was 3.6 and 3.2, respectively. In some instances, the source transcript of an aeTSA were only slightly more abundant in the tumor than in the NAT, however, this reflects only the overall abundance of the entire transcript, and the peptide coding sequences were in fact more abundant in the cancer (Figure 20). This was also true for aeTSAs, in which the peptide coding region was either entirely absent or lowly expressed in the NAT but was more highly expressed in the cancer tissue.

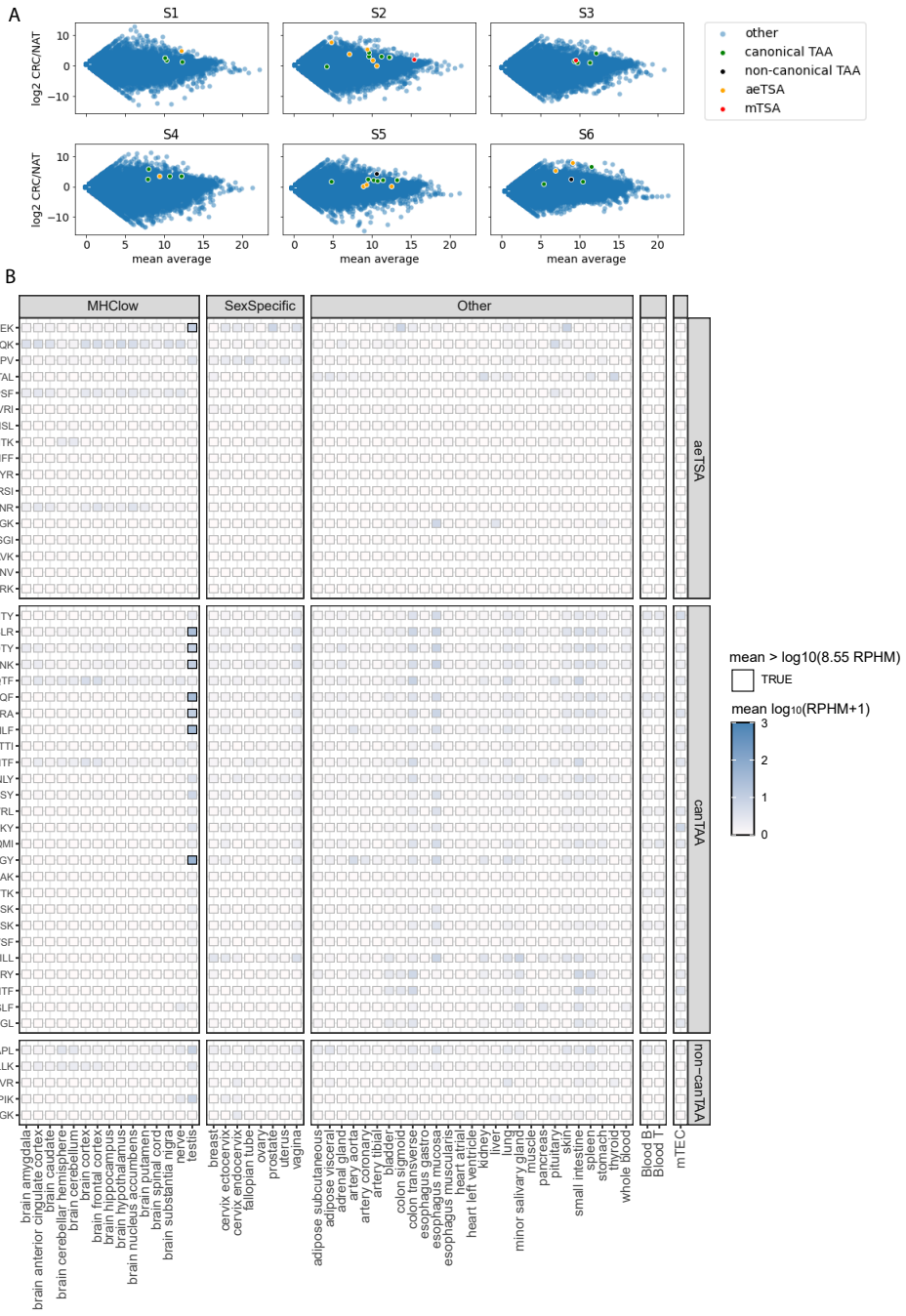

Figure 20. – RNA expression profiles of putative TSAs and TAAs

MA plots displaying the log2FC of transcripts, in TPM, in CRC compared to the matched NAT on the y-axis and the mean average expression in a given tissue sample (mean of CRC and NAT). Highlighted points indicate the source transcripts of putative TAA and TSAs. Both S4 and S5 plots have a canonical TAA point that is not visible, as it overlaps with another canonical TAA source transcript. B) Heatmap of mean RNA expression in log(rphm+1) of aeTSA coding sequences and TAA coding sequences (divided as canonical TAAs (canTAA) and non-canonical TAAs (non-canTAA) in normal tissues from Genotype Tissue Expression (GTEx) Portal and in pooled TEC samples.

MHClow tissues include those from brain, nerve, and testis, which have been shown to lowly express MHC I. A black outline indicates a mean RNA expression >8.55 rphm.

A

B

Figure 21. – RNA expression of MCS in cancer and NAT

Interactive Genome Viewer screenshots of RNA-seq data for two peptides of interest, which act as a proof of concept that although the source transcript of a TSA or TAA may not be highly overexpressed in cancer compared to NAT, the MAP-coding sequence (MCS) can be significantly overexpressed in the tumor. A) MCS for aeTSA sequence SIIETVNSL in S2 RNA-seq data. The source transcript of SIIETVNSL has a log2FC of approximately -0.26 in CRC compared to NAT. B) MCS for aeTSA sequence GQIELSIYR in S5 RNA-seq data. The source transcript of GQIELSIYR has a log2FC of approximately 1.1 in CRC compared to NAT.

To evaluate the specificity of our putative tumor antigens, we determined the mean expression of the peptide-coding sequences in the large dataset of healthy tissues provided by GTEx (Figure 19B). The TSA sequences were not expressed above 8.55 RPHM in any healthy tissues, except RIGGVGVEK, an aeTSA identified in S2, which was expressed above threshold in the testis. This suggests that this TSA could also be classified as a cancer-testis antigen (CTA), a class of aeTSA that is expressed in male germ cells but may also be aberrantly expressed in cancer. Due to the absence of MHC I in testis, these antigens are also promising candidates for cancer immunotherapy (160). This putative TSA is an LY6G6F-LY6G6D exon frameshift. While these genes have not been previously reported as CTAs, another member of the same gene family, LY6K, has been reported as a CTA in lung and esophageal cancers (161). TAA expression was below threshold in healthy tissues, although it tended to be higher in the esophagus and the transverse colon. Seven of these peptides were also expressed above threshold in the testis.

2.4.6 Cancer specificity and predicted immunogenicity of TSAs and TAAs

Following our identification of putative TSAs and TAAs, we validated all of the TSAs and a subset of nine TAAs with synthetic peptides. These TAAs were selected based on favorable initial TMT intensity ratios and precursor ion fractions in cancer vs matched NAT. These candidates all had MS/MS that correlated well with those of the synthetic peptides, with Pearson correlation score $\geq$ 0.6 (Supplementary file 4). We then labeled synthetic peptides with TMT10plex-129N,130N, and 131 at concentrations of 10, 100, and 1000 fmol, respectively, and spiked into remaining purified MAPs from tissue samples that were labeled with TMT126 (NAT) and 127N (CRC). SPS-MS3 was then used to quantify peptides of interest in these samples. Despite the decreased sensitivity of SPS-MS3, we were able to quantify seven TSAs and seven TAAs. We selected good quality PSMs for quantification, and as expected for antigens of this nature, all were more abundant in their respective CRC compared to NAT (Table 7). Determining the ratio of intensity of TMT127N peptides compared to TMT126 peptides revealed that TSAs had a median intensity fold change of 16.96 in CRC compared to NAT, while TAAs had a fold change of 6.93. In addition, the TSA with sequence RYLEKFYGL was also overexpressed in the S1 tumor, despite only

passing our transcriptomic thresholds for S6. Thus, we were able to demonstrate that the TSA identification methodology used in this study successfully identified TSA and TAA sequences that are more highly abundant at the surface of cancer cells than that of NAT.

Table 7. – Relative quantification ratios of validated tumor antigens in CRC

Sequence	Nature of antigen	Sample	Endogenous sample ratio	Mean intensity	SPS-MS3 ratio (127N/126)	Synthetic calibration curve R^2
RMLLSHTGK	aeTSA	RKO	N.D.	N.D.	N.D.	N.D.
LPHRALSGI	aeTSA	S1	-0.364	N.D.	N.D.	N.D.
GTNPTAAVK	aeTSA	S2	2.095	7238.425242	12.174	1.000
LRHKLVLNR	aeTSA	S2	0.307	N.D.	N.D.	N.D.
RIGGVGVEK	aeTSA	S2	1.965	29256.45	6.740	1.000
SIIETVNSL	aeTSA	S2	0.288	N.D.	N.D.	N.D.
TVNTQQYNTK	aeTSA	S2	-0.021	N.D.	N.D.	N.D.
SVSHLHIFF	aeTSA	S3	-1.100	N.D.	N.D.	N.D.
TTLENLPQK	aeTSA	S4	0.134	3140.8875	3.783	0.999
AQKLQVRI	aeTSA	S5	0.793	N.D.	N.D.	N.D.
GQIELSIYR	aeTSA	S5	0.328	N.D.	N.D.	N.D.
HGALSIRSI	aeTSA	S5	0.777	N.D.	N.D.	N.D.
RLMKFLPV	aeTSA	S5	0.171	N.D.	N.D.	N.D.
SLYISEERK	aeTSA	S5	0.046	N.D.	N.D.	N.D.
VQTAVLNV	aeTSA	S5	1.089	N.D.	N.D.	N.D.
VEAPHLPSF	aeTSA	S6	1.059	43782.84192	41.318	1.000
RNRQVATAL	aeTSA	S6	1.090	12174.6625	5.722	1.000
RNRQVATAL	Not assigned	S1	0.890	15514.2375	3.507	1.000
KIGEVIVTK	mTSA	S2	2.506	70659.6	13.637	1.000
TRSTIILHL	mTSA	S3	1.381	34365.32187	48.807	0.997
VLYRSVLLLK	non-canonical TAA	S6	0.997	N.D.	N.D.	N.D.
TYKYVDINTF	canonical TAA	S1	1.969	29834.36875	8.226	0.998
RYLEKFYGL	canonical TAA	S1	2.840	27614.24286	7.661	0.997
RYLEKFYGL	canonical TAA	S6	2.970	106928.2875	16.090	0.999
KSINEFWNK	canonical TAA	S2	2.212	56110.11667	5.238	0.999
RIQLPVVSK	canonical TAA	S4	1.083	7612.378571	2.073	0.999
QMAGLRDTY	canonical TAA	S3	1.140	36090.60294	2.884	0.999
AQYDQASTKY	canonical TAA	S4	1.452	N.D.	N.D.	N.D.
FVDNQYWRY	canonical TAA	S4	0.721	5853.986533	10.954	1.000
SANVSKVSF	canonical TAA	S5	1.114	12780.925	2.321	0.999

N.D.: not detected.

Endogenous sample ratio: 127N/126 ratio in endogenous samples

To examine the intertumoral distribution of these TSAs and TAAs in other CRC tumors, we plotted the log(RPHM+1) expression of the peptide coding sequences in 151 colon adenocarcinoma samples from The Cancer Genome Atlas (TCGA) (Figure 21A). To evaluate the sharing potential of our antigens, for each peptide of interest, we first calculated the average of log-transformed (log(rphm+1)) values of pooled GTEx (n=2442) and mTEC (n=8) samples. Overall, nine TSAs (53%) and nine TAAs (100%) had an expression ≥10-fold above their corresponding averaged GTEx/mTEC value in at least 5% of TCGA COAD tumors. This demonstrates that TAAs are more frequently shared among COAD TCGA tumors than their TSA counterparts. However, this also means that most TSAs are highly shared in these samples.

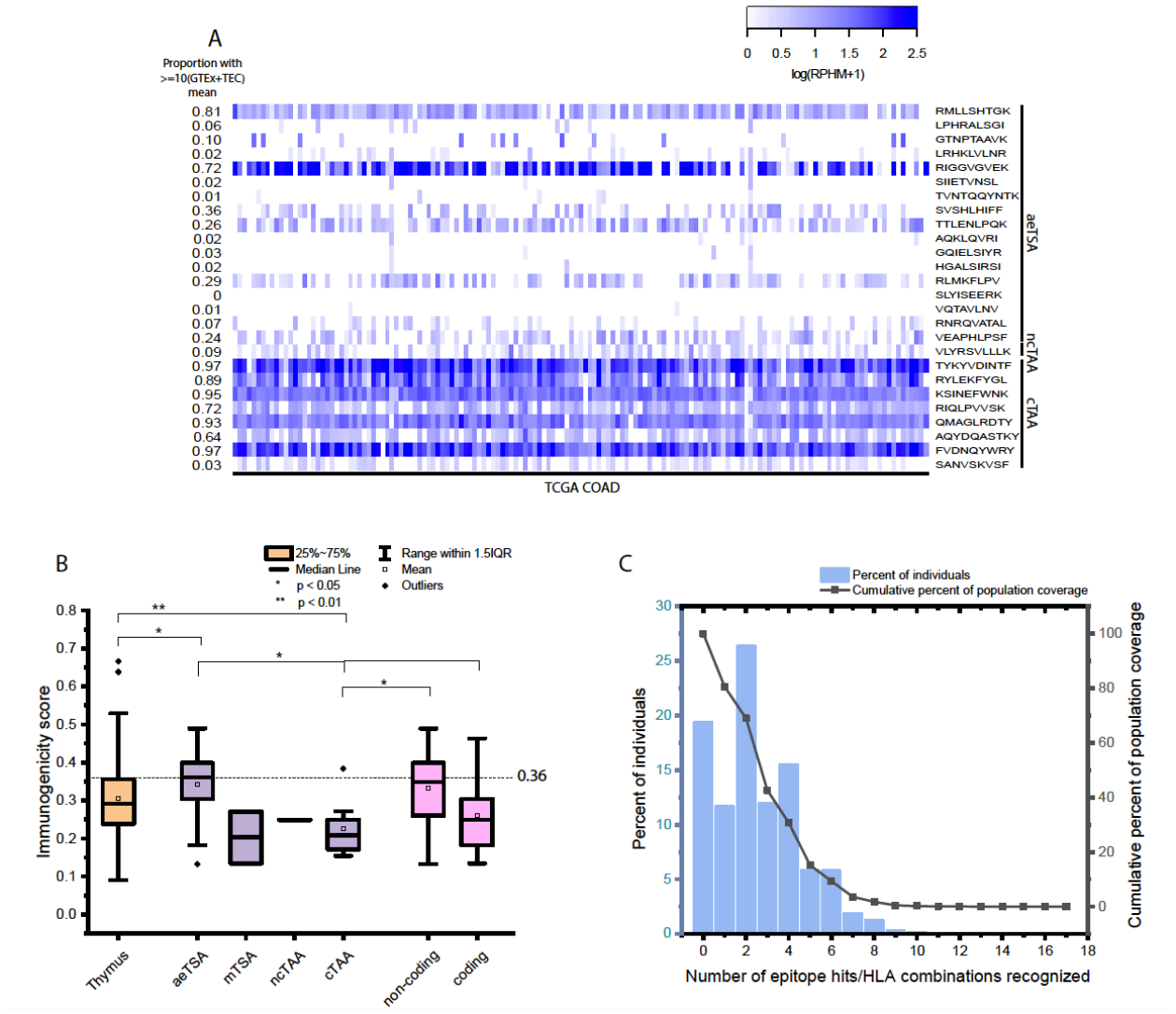

Figure 22. – Validation of TSAs and TAAs

A) Heatmap displaying mean RNA expression in log(rphm+1) of TSAs and TAAs in 151 TCGA COAD samples. The proportion of TCGA COAD samples expressing the TSA and TAA sequences at least 10-fold higher than the log-transformed (log(rphm+1)) mean expression of pooled GTEx and mTEC samples is displayed on the left. B) rEpitope immunogenicity scores of various groupings of validated TSAs and TAAs compared to presumably non-immunogenic thymic peptides reported in Adamopoulou et al. 2013. rEpitope suggested threshold of immunogenicity for MHC I peptides (0.36) is indicated by the dashed line. E) Predicted prevalence of tumor antigen-binding MHC class I alleles in US population (IEDB).

Another important consideration in the identification of tumor antigens is whether these peptides are able to invoke an effective anti-tumor immune response. Repitope predictions of immunogenicity revealed that our aeTSAs are predicted to be significantly more immunogenic than a set of thymic peptides which are presumed non-immunogenic (162) (Figure 21B). In addition, aeTSAs had significantly higher immunogenicity scores compared to canonical TAAs and to coding TAs overall (TSAs and TAAs derived from coding regions). In fact, TAAs from canonical regions were predicted to be significantly less immunogenic than thymic peptides ($p < 0.01$). This could be partially due to the low number of TAAs that we validated. If we consider these predictions with the entire set of 31 TAAs, this is no longer the case (Figure 22). Considering all 31 TAAs revealed that MSI TAs are predicted to be more immunogenic than thymic peptides, while there is also a statistically significant increase in predicted immunogenicity of TAs derived from MSI tissues compared to MSS.

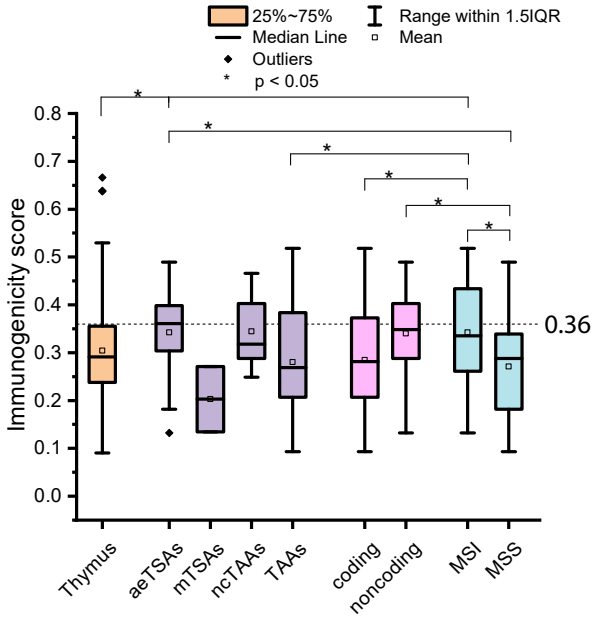

Figure 23. – Predicted immunogenicity of TSAs and TAAs

rEpitope immunogenicity scores of various groupings of validated TSAs and all TAAs compared to presumably non-immunogenic thymic peptides reported in Adamopoulou et al. 2013. rEpitope suggested threshold of immunogenicity for MHC I peptides (0.36) is indicated by the dashed line. This figure differs from figure 21B in that it includes all TAAs reported in this work, not only the nine that were chosen for validation.

Finally, we sought an approximation of the proportion of individuals who possess the alleles that are predicted to bind and present our tumor antigens (Figure 21C). Many of the antigens in our samples are prevalent, and an estimation with the IEDB population coverage tool predicted that 80.64% of the United States population expresses at least one of the alleles associated with the TAs identified in this study.

2.5 Discussion

Mass spectrometry is currently the best method to identify MAPs of interest, as it can directly sample the MHC I immunopeptidome and eliminates the need for error-prone prediction software, which are unable to incorporate the largely misunderstood intricacies of MAP processing and presentation (163). Our approach has previously led to the identification of TSAs in lung cancer (101), ovarian cancer (102), and acute myeloid leukemia (AML) (103), the majority of which are aberrantly expressed and derive from non-coding regions. The addition of TMT labeling allows us to both improve MAP identification and quantify peptide abundance between samples (88). While TSAs and TAAs have been identified in CRC, studies to date have only taken interest in the coding portion of the genome. By elucidating the MHC I immunopeptidomes derived from both canonical and non-canonical sequences in CRC cell lines and tumors, we thus present, to our knowledge, the first successful identification of aberrantly expressed TSAs in CRC. As novelty to our well-established identification workflow, we incorporated matched NAT of the respective CRC primary samples in our analysis, thus allowing for the most accurate possible 'control' samples of normal expression of peptide-coding regions. The aeTSAs identified here derive primarily from non-coding regions, which has also been previously demonstrated in other cancers (101-103).

MSI tumors are characterized by more favorable responses to ICI (particularly PD-1 inhibition) and increased immune infiltration compared to their MSS counterparts (69, 164), which was demonstrated in our samples using bioinformatic tools (Figure 11C, Figure 12B). The increased mutational load (Figure 11E, Figure 13B and C) and increased immunogenicity of MSI in CRC (144) suggested that these tumors would be characterized by a larger TSA or TAA burden. While the MSI tissue samples were sources of many TSAs and TAAs, we were able to identify eight aeTSAs, two mTSAs, and 18 unique TAAs in MSS tissues (seven of which were validated with synthetic peptides). Thus, it could be that the unfavorable response of MSS tumors to ICI is not due to a lack of tumor antigens, but rather to a lack of immune activation against these antigens. Accordingly, when considering all 31 of our identified TAA sequences, we saw a statistically significant decrease in the Repitope immunogenicity scores of TAs derived from MSS tissues compared to their MSI counterparts (Figure 22). While this trend was not observable when only

considering our validated TAA sequences, this could be attributed to the decrease in sample size for both subtypes. Despite a population-level decrease in immunogenicity, there are MSS-derived TAs with immunogenicity scores above the suggested threshold, which could still hold promise for immunotherapy. The 'immune cold' status of MSS tumors could alternatively be resulting from a lack of recruitment to the tumor site (165). Fortunately, there is a wide array of strategies designed to overcome the lack of immune infiltration into cold tumors, which could perhaps be used in combination with ICI or other immunotherapeutic approaches, such as vaccines, making use of TSAs such as those described here.

Among these TSAs, we identified two mTSAs unique to tumors derived from PLK1 and HDSP1, with missense mutations A520T and V345I occurring in 27% and 49% of RNA-Seq reads, respectively. The HDSP1 mutation is predicted to be benign by software such as Polyphen and CADD (166, 167). While the PLK1 mutation is documented in dbSNP (rs1004523813), it was not excluded from our analyses as the mutation is tumor-specific (not present in paired NAT) and it is very rare in the population (minor allele frequency <0.01) (168). Further, it is well-documented in COSMIC (cancer.sanger.ac.uk) and is predicted to be pathogenic by the Functional Analysis through Hidden Markov Models (166, 169). A recent study in immunopeptidomics of CRC organoids previously reported the discovery of three mTSAs (104). Unless derived from driver mutations, mTSAs are rare and thus it is unlikely that they are shared between tumors, and these particular mTSAs are absent from our study. In addition to mutation rarity, differences in peptide processing and presentation at least partly attributable to HLA diversity among samples further lessens the likelihood of identifying shared mTSAs. However, it is worth mentioning that MAPs from the source genes were reported in several of our samples (two, three, and three unique peptides from U2SURP, MED25, and FMO5, respectively). When taken together, these observations suggest that the overlap of mTSAs between specimens is relatively low, and it is thus not surprising to note the absence of previously reported mTSAs in our study, despite the relatively high SNV burden among our samples (Figure 11E, Figure 13C). mTSAs are immunologically relevant and have the capacity to be immunogenic but due to their lack of sharing between individuals, their potential for use in large-scale immunotherapy is limited.

While mTSAs are expected to be rare and unique to a given tumor, it has been shown that aeTSAs can be shared among patients (90). Here, we did not identify any common aeTSAs among our six patients, however, we did identify two unique aeTSAs in different patients that were derived from the same transcript of the COL11A1 gene, which is known to be associated with CRC (Table 4). It should be noted that in this study, we worked with only six primary samples, which were largely diverse in their HLA alleles, thus reducing the likelihood of shared TSAs. The fact that we were still able to identify TSAs from the same transcript is encouraging, as it suggests that this transcript is generating biologically relevant peptides across different tumors. It is possible that these TSAs could be presented by other tumors with similar alleles, or that this transcript could be generating novel TSAs capable of being presented by other HLA alleles not examined here. Additionally, the TSA sequence RNRQVATAL was originally identified as a TSA candidate in S6 only. It was not considered a TSA in S1 originally due to the level of expression in normal tissue (RNA coding sequences not expressed at least 10-fold higher in cancer than in NAT), and yet at the immunopeptidomic level it had a 3.5-fold higher intensity in CRC than in NAT. The fact that this peptide could also be considered a tumor antigen in S1 relates to the fact that mRNA abundance and protein abundance are not highly correlated (170), and our stringent identification pipeline excluded it based on RNA-seq data. This reinforces the need for mass spectrometry to directly sample the immunopeptidome, to relatively quantify the abundance of such peptides at the cell surface, and to validate the immunogenicity of TSAs in large-scale *in vitro* studies.

Outside of our six tissue samples, the decreased sharing of some TSAs among TCGA COAD tumors suggests that certain TSA sequences are not widely shared (Figure 21A). Of the nine TSAs that are expressed ≥10-fold above their corresponding averaged GTEx/mTEC value in at least 5% of TCGA COAD tumors, three are from intergenic sequences, two from exons, two from exon frameshifts, and one each from intronic or 5’ UTR sequences. This small sample size prevents us from drawing any conclusions, however, there may be a therapeutic advantage to distinguishing highly shared TSAs from those that are less abundant across COAD populations. While high tumoral RNA expression of a TSA sequence does not guarantee MHC I presentation of that peptide, it does increase the likelihood that a given TA sequence, or perhaps other sequences from the same transcript, could have dysregulated MHC I presentation in cancer.

In contrast to TSAs, multiple canonical TAAs are shared between different primary CRC samples, with up to three samples presenting the same TAA. Additionally, the same genes can generate multiple relevant TAAs across tumor samples, with three unique TAAs being derived not only from the ASPM gene but from the same transcript (among these, SANVSKVSF was validated) (Table 5; Supplementary file 3). The increased intertumoral sharing of TAA sequences compared to TSA sequences is also reflected in TCGA data, in which canonical TAA coding sequences are expressed more frequently and more abundantly in colon adenocarcinoma samples compared to their TSA counterparts (Figure 21A). This is to be expected due to the very nature of TAAs, which are expressed in normal tissues but overexpressed in cases of malignancy, compared to TSAs which arise only in cases of mutated or aberrantly expressed sequences. As such, TAAs are more challenging to use in immunotherapy approaches as they have been known to induce auto-immune responses, or even T cell tolerance (171). We also demonstrated that the TAAs identified here are predicted to be significantly less immunogenic than the TSAs (Figure 21B). However, TAAs can certainly be advantageous as cancer biomarkers, as is the case with CEA, the first TAA discovered in colorectal cancer in the mid-1960s (62).

We were initially surprised at the lack of CEA-derived TAAs in our tissue samples, despite the presence of several CEA-derived MAPs in our dataset (supplementary files 1-2). A closer examination revealed that CEA-derived MAPs (for example, those derived from CEACAM5 or CEACAM7) were excluded from our analysis following the initial peptide classification, which removes MAPs that are not overexpressed at least 10-fold higher in cancer than in matched NAT, and those that are expressed more than 2 RPHM in NAT. In the interest of comparing our findings with other contemporary studies on CRC immunopeptidomics, we queried our dataset for the MAPs identified as potential vaccine candidates in Löffler *et al.* 2018 (111). Out of the 12 TAAs they selected, six of them were identified in our tissue samples. However, five of these peptides did not pass the initial classifications in our pipeline ($\geq$10FC in cancer compared to NAT and $\leq$2RPHM in NAT), and the other was found to be expressed more than 8.55 RPHM in mTECs (Table 6). We would like to note here that our pipeline was designed to identify TSAs, and thus has a stringent set of criteria meant to exclude peptides present in normal tissues. As no universal thresholds have been established to classify TAAs, differences in thresholds and filtering steps

between studies will naturally result in differential TAA identification. Löffler *et al.* also demonstrated T cell responses to their TAAs (111), suggesting that these antigens do have clinical potential.

While this study is not meant to be a comprehensive view of CRC immunopeptidomics, the primary goal of our work was to provide a proof of concept that aeTSAs can be identified and are more abundantly presented at the cell surface of CRC than of paired NAT. Despite typical limitations of immunopeptidomic studies such as the amount of available material (particularly for tissue biopsies) and instrument sensitivity, we present here the identification of 19 TSAs. An additional drawback of this study was the low MAP identification in CRC cell lines attributable to the low MHC I abundance at the cell surface, which decreased the probability of identifying TSAs. In the future, MHC I presentation of cell lines could be boosted with IFN-γ treatment to increase identification, particularly of lowly abundant peptides (172). This approach would be useful to investigate, for example, the sharing of tumor antigens across samples, despite the lower identification obtained with cell lines. While it should be kept in mind that IFN-γ treatment alters gene expression (173), studies are currently underway to evaluate the impact of IFN-γ, or other drug treatments, on TSA presentation and identification. Future investigations will include evaluations of TSA and TAA immunogenicity with T cell assays. In addition, expanding the sample size with primary tissues sharing common HLA alleles could drastically increase the likelihood of identifying shared TSAs. Here, we examine only primary samples of stage 2 non-metastatic CRC. Differential peptide presentation could be occurring at other tumor stages due to alterations in tumor biology. Expanding the reach of this study could include a large-scale analysis of multiple CRC samples as well as the investigation of TSAs in other stages of CRC.

2.6 Acknowledgments

This study was supported by grants from the Canadian Cancer Society (705604) to C.P. and P.T. J.C. is supported by Graduate Scholarships from CIHR, FRSQ, and the IRIC graduate program. We thank Raphaëlle Lambert, Jennifer Huber, and the IRIC genomics facility for RNA-seq. We also acknowledge the assistance of Gregory Ehx, Patrick Gendron, and Albert Feghaly for essential bioinformatics analyses, and Eric Bonneil for mass spectrometry expertise. The IRIC proteomics facility is a Genomics Technology platform funded in part by the Canadian Government through Genome Canada. We thank the Genotype-Tissue Expression (GTEx) Project for providing RNA-seq data. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We thank the TCGA Research Network for the data generation upon which the results shown here are based: https://www.cancer.gov/tcga.

2.7 Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (174) via the PRIDE partner repository (175) with the dataset identifier PXD028309 and 10.6019/PXD028309. The transcriptomic data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (176) and are accessible through GEO Series accession number GSE195985 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE195985).

Chapter 3 – Conclusion and Perspectives

3.1 Conclusion

For this thesis work, I investigated the immunopeptidomes of a series of CRC samples, employing various techniques including next generation sequencing, MS analyses, and a stringent TSA identification pipeline to identify MAPs of clinical interest. Transcriptomic profiling revealed differentially expressed genes between MSI vs MSS CRC, as well as an increased immune infiltration and increased mutational burden in MSI samples. We identified over 30 000 unique MAPs by mass spectrometry; despite each sample having a large proportion of unique peptides, these peptides frequently derived from common source genes, even across cell line and tissue samples. Our proteogenomic investigation of CRC-derived cell lines and primary tissue samples unveiled 19 novel, primarily non-coding TSAs, as well as a selection of mostly coding TAAs. These TAs were shown to not be expressed in normal tissues and demonstrated intertumoral distribution among CRC samples from the TCGA database. In addition, TMT isobaric peptide labeling provided a convenient approach to compare MAP abundance across tumor and NAT samples and to validate the identification of TAs with synthetic peptides. TMT-mediated MS quantification revealed that a select subset of TAs were all more abundant at the cell surface of CRC tissues compared to NAT, increasing our confidence in their relevance as TAs of interest. Despite the increased immune infiltration observed in MSI samples in our study as well as their improved prognoses in response to ICI therapy, here, MSI tumors did not consistently present more TSAs than MSS tumors. This suggests that the poor response of MSS tumors to ICI is not due to a lack of TSAs, but rather a lack of immune activation against these antigens. In line with this, the tumor antigens derived from non-coding regions tended to be more immunogenic than those from coding regions, and TAs identified in MSI tissues tended to be more immunogenic than those identified in MSS samples. While it should be noted that our sample size was relatively small for MSI tumors (n = 2), these observations suggest that the quantity of TAs may not be as relevant as their quality when it comes to response to immunotherapy (though this requires future investigation). Overall, we have identified, to our knowledge, the first aeTSAs in CRC, which we have thoroughly validated and demonstrated their promise as therapeutic agents.

3.2 Perspectives

3.2.1 Proteogenomics approach

In the late 1950s, the central dogma of biology described the linear progression of DNA being transcribed into RNA and then translated into protein (177). In the decades that followed, much was learned about the ways in which this process could move 'backward' (such as the reverse transcription of RNA into DNA employed by certain classes of viruses) (178), as well as the myriad ways in which DNA, RNA, and proteins may be regulated and modified (179-181). This is of course reflected in the differences in magnitude between protein coding genes (approximately 20 000) and the size of the human proteome (>1 million). A consequence of these highly modifiable processes is the resulting lack of correlation that exists between RNA and protein abundance for a given transcript and corresponding protein. In fact, the correlation between RNA and protein abundance in eukaryotes is only ~40%, meaning that 60% of the variation in protein abundance is attributable to other factors, such as post-transcriptional modifications, translational regulation, or protein degradation (170, 182).

Even so, a proteogenomic approach that employs transcriptomic data along with proteomic analyses is currently the best way to identify relevant tumor antigens. In our case, this involves the construction of databases that contain all RNA sequences present in the transcriptomic analyses, and the removal of sequences that are present in the 'normal' control, in order to identify cancer-specific sequences. While the absence of RNA likely entails the absence of protein (as there is no material to be translated), the presence or abundance of RNA carries little weight with regard to the resulting presence or abundance of a corresponding protein, and this should be kept in mind when pursuing proteogenomic analyses. Thus, it is quite possible that in our work we have eliminated candidates based on RNA expression in 'normal', when the corresponding protein product is absent and thus would be a tumor-specific antigen. However, these possible 'false negatives' are more favorable than any 'false positives' that could result from a less stringent database construction.

Additionally, this further emphasizes the importance of identifying tumor antigens through MS analyses, as the presence of a sequence at the transcriptomic level does not

guarantee its translation into a protein, much less its degradation and presentation at the cell surface by MHC I. This is also why, in our study, we make use of normal adjacent tissues as the best possible control for our primary tumor samples, as they mimic the 'normal' expression of the same individual, and account for any tissue-specific expression.

3.2.2 Immunogenicity and T cell reactivity

While we went to great lengths to validate our TSAs, there are some validations outside of the scope of this study that need to be completed before pursuing these antigens in clinical trials. For example, our predictions of peptide immunogenicity were purely *in silico*, and based on machine learning algorithms that simulate the behavior of these peptides in TCR-MHC:peptide interactions (131). Certain peptide characteristics can be used to predict immunogenicity, such as structure (183) or amino acid composition (184), and these predictions can aid in the selection of promising candidates for further validation. However, these strategies generate population-level predictions of immunogenicity, not individual ones, and do not consider certain peptide features such as post-translational modifications or other *in vivo* immunological factors (185). As such, further studies are needed to fully understand the immunogenicity of a given peptide, and this should ideally be done with a combination of *in vitro* and *in vivo* strategies. For example, HLA binding assays can be used to measure peptide affinity, and *in vitro* T cell assays, such as ELISpot or tetramer assays, can be used to measure cytokine responses, T cell proliferation, or T cell recognition of the peptide of interest. However, researchers should be cautioned against using only a single assay, as they may have conflicting results; a 2018 study demonstrated that MHC I tetramer assays can miss fully functional T cell clones that recognized the peptide of interest with lower affinity (186). In addition, T cell reactivity assays would be necessary to confirm the validity of these antigens, as T cell recognition of TSAs is required for an anti-tumor immune response. While it might be assumed that the presence of TILs in the TME guarantees T cell reactivity, this is not necessarily the case. It has been demonstrated that TILs in human lung and colorectal cancers are not exclusively reactive to tumor antigens, and may be considered "bystanders" in the anti-tumor immune response (187). Another study reported that as few as 10% of TILs in ovarian and colorectal cancers are reactive to the tumor (188), further emphasizing the need to confirm TIL reactivity to tumor antigens. Finally, these immunogenicity evaluations should be

confirmed with *in vivo* studies, such as in HLA transgenic or otherwise humanized mice, which can provide a more complete representation of the innate and adaptive immune systems in response to an antigen without risking the well-being of patients (89, 189). However, these studies are not without drawbacks. For example, even in HLA-transgenic mice, differences in murine and human protein sequences could result in an immune response against an antigen in mice that otherwise would not occur (89). In addition, many humanized mouse models are not able to completely reconstitute all relevant immune cells (e.g., myeloid and red blood cells) (190).

3.2.3 Cancer vaccines

The goal of identifying TSAs and TAAs is of course so that they can be of use to individuals affected by cancer, either to mitigate or cure the disease. In recent years, cancer immunotherapy has become particularly attractive. Of course, these treatments can take many forms. A variety of therapies could employ tumor antigens including vaccination, adoptive TIL transfer, or engineered T cells that specifically target these antigens at the cell surface of tumors (191). In terms of vaccination, many different approaches are possible. In some approaches, whole tumor cells are injected into the patient, such that T cells can be activated against the tumor-associated components to attack the tumor present in the body (192). Other more specific approaches involve injecting purified tumor antigens, or even loading the antigen onto DCs beforehand and injecting them such that they can act as antigen presenting cells to the existing T cells in the patient (193).

Many studies and clinical trials are currently underway evaluating the efficacy of TA-based vaccines. In fact, such studies have resulted in a commercially available prostate cancer vaccine, which consists of peripheral blood mononuclear cells that have been activated against a prostatic acid phosphatase – granulocyte-macrophage colony-stimulating factor fusion protein (194). Prostatic acid phosphatase is a cancer antigen expressed in >95% of prostate cancers (195). However aside from this, vaccines making use of TAAs have had relatively meagre results (196). As previously discussed, the majority of work in this field thus turned to the investigation of neoantigens, or mTSAs, despite their lack of intertumoral sharing. A recent study demonstrated that treatment with a DNMT inhibitor was able to reinduce cancer-testis antigen expression in

mice bearing metastatic CRC tumors, and that combining this inhibitor with an irradiated whole-cell CRC vaccine (known as GVAX) was able to improve survival compared to the GVAX vaccine alone (197). Thus, while several vaccines have been shown to be effective against TAAs or mTSAs, no aeTSA vaccines have been developed to date. Moreover, there are many questions remaining in terms of how to best formulate such a vaccine.

As I have demonstrated in this thesis, aeTSAs are being presented at the cell surface of CRC, are absent on NAT, and are predicted to be immunogenic, including in tumors with high immune infiltration. And yet, the existence of the tumor demonstrates the inability of the immune system to effectively eradicate cancer cells. In fact, it has been proposed that the immune system remains ignorant of many TSAs given their inefficient presentation by tumor cells (198). This inefficient immune activation is three-fold: first, tumor cells have been shown to be poor T cell activators, given that they lowly express CD28, which is a necessary costimulatory signal for T cell activation. It is thus more likely for T cells to be activated against TSAs cross-presented by DCs; second, epithelial cells present little MHC in general (199), and this is compounded by the various mechanisms of MHC downregulation employed by cancer cells, further lowering the likelihood of TSA presentation by tumor cells; and finally, DC cross-presentation has been demonstrated to be biased to present MAPs derived from stable proteins, whereas direct presentation typically presents MAPs derived from DRIPs (42, 200). So, even if some TSAs are being effectively cross-presented by DCs, others derived from DRIPs could be completely ignored (198). For these reasons, the most effective formulation for an aeTSA vaccine would likely be DCs pulsed with aeTSAs, rather than a purified antigen or whole cell formulation.

While the type of immunotherapy needs to be optimized, the contents of a vaccine would also need a sufficient amount of consideration. It is likely that an effective cancer vaccine would also need to contain adjuvants to recruit immune cells to the injection site and activate APCs and prevent the induction of tolerance in response to the antigen(s) (201). This is the case for the previously mentioned prostate cancer vaccine, which contains an antigen fused to granulocyte-macrophage colony-stimulating factor, which acts as an adjuvant to enhance APC efficacy (194). Other possible adjuvants for a cancer vaccine that induce an immune response include pathogen-associated molecular patterns, or PAMPs, which are recognized by pattern recognition receptors

(PRRs) expressed by innate immune cells (201). Such formulations could be improved by incorporating more than one tumor antigen, to ensure effective immune activation against the tumor, particularly given the inevitable intertumoral variability in terms of expression of these antigens. Additionally, cancer vaccines could be paired with existing therapies, such as ICI, or other strategies to improve immune infiltration in the tumor. This is particularly the case for MSS CRC, for which ICI is ineffective and as these tumors are typically immune 'cold'.

Finally, a limitation of our study, and more widely of immunopeptidomic studies, is the constraint of peptide presentation and identification imposed by HLA binding specificity. As the HLA repertoire of an individual plays a critical role in determining the MAPs they are able to present, there is a limitation in the proportion of individuals able to present given TAs. As seen in our study, no TSAs were shared among samples, and this is undoubtedly attributable at least in part to the large diversity of HLA alleles that these samples possessed, limiting their ability to present the same MAPs. There are two separate ideas to be discussed on this matter. The first is that, fortunately, HLA molecules are able to bind many peptides and HLA binding repertoires are not unique for individual alleles, and there are certainly MAPs that can be bound by multiple alleles. In fact, HLA alleles which have largely overlapping binding specificities can be grouped together into "supertypes", meaning that alleles of this supertype are likely capable of binding many of the same peptides (202). This is very fortunate when considering the development of TA-based vaccines, given that over 6000 HLA alleles have been identified (203) and it would be a much larger undertaking if each of these alleles had entirely unique binding specificities.

The second idea to be discussed is the tendency of immunopeptidomic studies to center HLA-A*02 alleles, and particularly HLA-A*02:01 due to the prevalence of this allele in the Caucasian population, and thus the North American population. A limitation of our study is that, while it did not focus exclusively on HLA-A*02 alleles, the primary tissue samples we used were derived exclusively from Caucasian subjects, and thus our study is not exempt from the flaws of such approaches. In particular, any TAs identified in such studies will primarily benefit certain populations i.e., those with HLA-A*02 alleles (204). Further, these studies do not represent the global majority, and cancer is a leading cause of death worldwide (205). While it would of course be difficult to achieve global representation in a single study, it should be argued that focusing

only on the most prominent alleles is a structural form of exclusion that will inevitably contribute to health inequities. To address this, studies seeking to identify TAs should approach the endeavour with this in mind, and incorporate multiple samples and more alleles to avoid serving only a small proportion of the population, as the objective should be to aid as many people as possible.

3.2.4 Remaining questions

While my thesis work ends here, the study of colorectal cancer immunopeptidomics will forge onward. Unfortunately, not all of the important questions could be addressed, but I hope the answers to them will be elucidated in the near future. If I were to continue this work, I would seek to incorporate more samples into our TSA identification pipeline, for two reasons: first, it would likely allow us to obtain more MSI primary tissue samples, whereas in our study only two samples were MSI, unfortunately preventing any statistical analyses and a clearer observation of trends between MSI and MSS samples, especially in terms of immunogenicity of TAs; and second, it would allow us to investigate whether our TAs are shared among other primary samples, particularly if we were able to obtain samples with similar HLA alleles to those in our initial study. Additionally, while in our study we did not identify any shared TSAs, we did identify two TSAs in different primary samples derived from the same transcript. This is especially interesting, as it suggests that this transcript is biologically relevant in CRC and is capable of generating multiple TSAs with different HLA allele binding specificities. It is thus possible that such a transcript could generate more TSAs in different samples. While we approximated the degree of sharing of TA-coding sequences in TCGA data, the expression of these sequences at the RNA level does not guarantee their presentation as MAPs. As such, it would be important to evaluate the degree of TSA sharing across many samples in terms of MAP presentation, rather than only RNA expression. Finally, while the generation of certain TAs may be linked to processes exclusively dysregulated in CRC, it would be interesting to evaluate whether certain TAs may have pan-cancer relevance, i.e., whether a TA identified in our study could be a TA in other cancers, especially since we have already determined that these sequences are absent or very lowly expressed in normal tissues.

Bibliography

1. Torpy, J. M., Lynm, C., and Glass, R. M. (2010) Cancer: The Basics. *JAMA* 304, 1628-1628

2. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 71, 209-249

3. Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017) Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683-691

4. Brenner, D. R., Weir, H. K., Demers, A. A., Ellison, L. F., Louzado, C., Shaw, A., Turner, D., Woods, R. R., Smith, L. M., and Canadian Cancer Statistics Advisory, C. (2020) Projected estimates of cancer in Canada in 2020. *CMAJ* 192, E199-E205

5. Phelan, S. M., Burgess, D. J., Yeazel, M. W., Hellerstedt, W. L., Griffin, J. M., and van Ryn, M. (2015) Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obes Rev* 16, 319-326

6. Tomiyama, A. J., Carr, D., Granberg, E. M., Major, B., Robinson, E., Sutin, A. R., and Brewis, A. (2018) How and why weight stigma drives the obesity 'epidemic' and harms health. *BMC Med* 16, 123

7. Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., and Jemal, A. (2020) Colorectal cancer statistics, 2020. *CA Cancer J Clin* 70, 145-164

8. Jones, S., Chen, W. D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M. A., Siegel, C., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Willis, J., and Markowitz, S. D. (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci U S A* 105, 4283-4288

9. Issa, I. A., and Noureddine, M. (2017) Colorectal cancer screening: An updated review of the available options. *World J Gastroenterol* 23, 5086-5096

10. Mitsala, A., Tsalikidis, C., Pitiakoudis, M., Simopoulos, C., and Tsaroucha, A. K. (2021) Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. A New Era. *Curr Oncol* 28, 1581-1607

11. Singh, H., Bernstein, C. N., Samadder, J. N., and Ahmed, R. (2015) Screening rates for colorectal cancer in Canada: a cross-sectional study. *CMAJ Open* 3, E149-157

12. Joseph, D. A., King, J. B., Dowling, N. F., Thomas, C. C., and Richardson, L. C. (2020) Vital Signs: Colorectal Cancer Screening Test Use - United States, 2018. *MMWR. Morbidity and mortality weekly report* 69, 253-259

13. Virostko, J., Capasso, A., Yankeelov, T. E., and Goodgame, B. (2019) Recent trends in the age at diagnosis of colorectal cancer in the US National Cancer Data Base, 2004-2015. *Cancer*

14. Force, U. S. P. S. T., Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., Donahue, K. E., Doubeni, C. A., Krist, A. H., Kubik, M., Li, L., Ogedegbe, G., Owens, D. K., Pbert, L., Silverstein, M., Stevermer, J., Tseng, C. W., and Wong, J. B. (2021) Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 325, 1965-1977

15. Laiyemo, A. O., Doubeni, C., Pinsky, P. F., Doria-Rose, V. P., Bresalier, R., Lamerato, L. E., Crawford, E. D., Kvale, P., Fouad, M., Hickey, T., Riley, T., Weissfeld, J., Schoen, R. E., Marcus, P. M., Prorok, P. C., and Berg, C. D. (2010) Race and colorectal cancer disparities: health-care utilization vs different cancer susceptibilities. *J Natl Cancer Inst* 102, 538-546

16. Mahabir, D. F., O'Campo, P., Lofters, A., Shankardass, K., Salmon, C., and Muntaner, C. (2021) Experiences of everyday racism in Toronto's health care system: a concept mapping study. *Int J Equity Health* 20, 74

17. Nelson, B. (2020) How structural racism can kill cancer patients. *Cancer Cytopathol* 128, 83-84

18. Meyer, B., and Are, C. (2017) Current Status and Future Directions in Colorectal Cancer. *Indian J Surg Oncol* 8, 455-456

19. Armaghany, T., Wilson, J. D., Chu, Q., and Mills, G. (2012) Genetic alterations in colorectal cancer. *Gastrointest Cancer Res* 5, 19-27

20. Day, D. W. (1984) The adenoma-carcinoma sequence. *Scand J Gastroenterol Suppl* 104, 99-107

21. Levine, J. S., and Ahnen, D. J. (2006) Clinical practice. Adenomatous polyps of the colon. *N Engl J Med* 355, 2551-2557

22. Edition, S., Edge, S., and Byrd, D. (2017) American Joint Committee on Cancer. Chapter 20 - Colon and Rectum. *AJCC cancer staging manual.*, 8th Ed., Springer, New York, NY

23. Karamchandani, D. M., Chetty, R., King, T. S., Liu, X., Westerhoff, M., Yang, Z., Yantiss, R. K., and Driman, D. K. (2020) Challenges with colorectal cancer staging: results of an international study. *Mod Pathol* 33, 153-163

24. Luchtenborg, M., Weijenberg, M. P., Roemen, G. M., de Bruine, A. P., van den Brandt, P. A., Lentjes, M. H., Brink, M., van Engeland, M., Goldbohm, R. A., and de Goeij, A. F. (2004) APC mutations in sporadic colorectal carcinomas from The Netherlands Cohort Study. *Carcinogenesis* 25, 1219-1226

25. Iacopetta, B. (2003) TP53 mutation in colorectal cancer. *Hum Mutat* 21, 271-276

26. Olivier, M., Hollstein, M., and Hainaut, P. (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* 2, a001008

27. Battaglin, F., Naseem, M., Lenz, H. J., and Salem, M. E. (2018) Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. *Clin Adv Hematol Oncol* 16, 735-745

28. Chen, M. L., Chen, J. Y., Hu, J., Chen, Q., Yu, L. X., Liu, B. R., Qian, X. P., and Yang, M. (2018) Comparison of microsatellite status detection methods in colorectal carcinoma. *Int J Clin Exp Pathol* 11, 1431-1438

29. Bonneville, R., Krook, M. A., Chen, H. Z., Smith, A., Samorodnitsky, E., Wing, M. R., Reeser, J. W., and Roychowdhury, S. (2020) Detection of Microsatellite Instability Biomarkers via Next-Generation Sequencing. *Methods Mol Biol* 2055, 119-132

30. Dinu, D., Dobre, M., Panaitescu, E., Birla, R., Iosif, C., Hoara, P., Caragui, A., Boeriu, M., Constantinoiu, S., and Ardeleanu, C. (2014) Prognostic significance of KRAS gene mutations in colorectal cancer--preliminary study. *J Med Life* 7, 581-587

31. Bang, Y. J., Kwon, J. H., Kang, S. H., Kim, J. W., and Yang, Y. C. (1998) Increased MAPK activity and MKP-1 overexpression in human gastric adenocarcinoma. *Biochem Biophys Res Commun* 250, 43-47

32. Hemmings, B. A., and Restuccia, D. F. (2012) PI3K-PKB/Akt pathway. *Cold Spring Harb Perspect Biol* 4, a011189

33. Huang, X. F., and Chen, J. Z. (2009) Obesity, the PI3K/Akt signal pathway and colon cancer. *Obes Rev* 10, 610-616

34. Idos, G., and Valle, L. (1993) Lynch Syndrome. In: Adam, M. P., Ardinger, H. H., Pagon, R. A., Wallace, S. E., Bean, L. J. H., Mirzaa, G., and Amemiya, A., eds. *GeneReviews((R))*, Seattle (WA)

35. Bhattacharya, P., and McHugh, T. W. (2021) Lynch Syndrome. *StatPearls*, Treasure Island (FL)

36. Fearnhead, N. S., Britton, M. P., and Bodmer, W. F. (2001) The ABC of APC. *Hum Mol Genet* 10, 721-733

37. Jasperson, K. W., Patel, S. G., and Ahnen, D. J. (1993) APC-Associated Polyposis Conditions. In: Adam, M. P., Ardinger, H. H., Pagon, R. A., Wallace, S. E., Bean, L. J. H., Mirzaa, G., and Amemiya, A., eds. *GeneReviews((R))*, Seattle (WA)

38. Abel, A. M., Yang, C., Thakar, M. S., and Malarkannan, S. (2018) Natural Killer Cells: Development, Maturation, and Clinical Utilization. *Front Immunol* 9, 1869

39. Murphy, K., Travers, P., Walport, M., and Janeway, C. (2012) *Janeway's immunobiology*, 8th Ed., Garland Science, New York

40. Wieczorek, M., Abualrous, E. T., Sticht, J., Alvaro-Benito, M., Stolzenberg, S., Noe, F., and Freund, C. (2017) Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front Immunol* 8, 292

41. Ferrington, D. A., and Gregerson, D. S. (2012) Immunoproteasomes: structure, function, and antigen presentation. *Prog Mol Biol Transl Sci* 109, 75-112

42. Bourdetsky, D., Schmelzer, C. E., and Admon, A. (2014) The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proc Natl Acad Sci U S A* 111, E1591-1599

43. Embgenbroich, M., and Burgdorf, S. (2018) Current Concepts of Antigen Cross-Presentation. *Front Immunol* 9, 1643

44. Joffre, O. P., Segura, E., Savina, A., and Amigorena, S. (2012) Cross-presentation by dendritic cells. *Nat Rev Immunol* 12, 557-569

45. Krangel, M. S. (2009) Mechanics of T cell receptor gene rearrangement. *Curr Opin Immunol* 21, 133-139

46. Yang, X., and Mariuzza, R. A. (2015) Pre-T-cell receptor binds MHC: Implications for thymocyte signaling and selection. *Proc Natl Acad Sci U S A* 112, 8166-8167

47. Takaba, H., and Takayanagi, H. (2017) The Mechanisms of T Cell Selection in the Thymus. *Trends Immunol* 38, 805-816

48. Smith-Garvin, J. E., Koretzky, G. A., and Jordan, M. S. (2009) T cell activation. *Annu Rev Immunol* 27, 591-619

49. Buchbinder, E. I., and Desai, A. (2016) CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *Am J Clin Oncol* 39, 98-106

50. Stirling, E. R., Bronson, S. M., Mackert, J. D., Cook, K. L., Triozzi, P. L., and Soto-Pantoja, D. R. (2022) Metabolic Implications of Immune Checkpoint Proteins in Cancer. *Cells* 11

51. Patsoukis, N., Weaver, J. D., Strauss, L., Herbel, C., Seth, P., and Boussiotis, V. A. (2017) Immunometabolic Regulations Mediated by Coinhibitory Receptors and Their Impact on T Cell Immune Responses. *Front Immunol* 8, 330

52. Parry, R. V., Chemnitz, J. M., Frauwirth, K. A., Lanfranco, A. R., Braunstein, I., Kobayashi, S. V., Linsley, P. S., Thompson, C. B., and Riley, J. L. (2005) CTLA-4 and PD-1 receptors inhibit T-cell activation by distinct mechanisms. *Mol Cell Biol* 25, 9543-9553

53. Seidel, J. A., Otsuka, A., and Kabashima, K. (2018) Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. *Front Oncol* 8, 86

54. Han, Y., Liu, D., and Li, L. (2020) PD-1/PD-L1 pathway: current researches in cancer. *Am J Cancer Res* 10, 727-742

55. Hanahan, D., and Weinberg, R. A. (2000) The hallmarks of cancer. *Cell* 100, 57-70

56. Hanahan, D., and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* 144, 646-674

57. Gonzalez, H., Hagerling, C., and Werb, Z. (2018) Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev* 32, 1267-1284

58. Rezaei, N. (2015) Cancer Immunology : A Translational Medicine Context. 1st Ed., pp. 1 online resource (XLVII, 597 pages 269 illustrations, 246 illustrations in color, Springer Berlin Heidelberg : Imprint: Springer,, Berlin, Heidelberg

59. Baghban, R., Roshangar, L., Jahanban-Esfahlan, R., Seidi, K., Ebrahimi-Kalan, A., Jaymand, M., Kolahian, S., Javaheri, T., and Zare, P. (2020) Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun Signal* 18, 59

60. Whiteside, T. L. (2008) The tumor microenvironment and its role in promoting tumor growth. *Oncogene* 27, 5904-5912

61. Gold, P., and Freedman, S. O. (1965) Specific carcinoembryonic antigens of the human digestive system. *J Exp Med* 122, 467-481

62. Gold, P., and Freedman, S. O. (1965) Demonstration of Tumor-Specific Antigens in Human Colonic Carcinomata by Immunological Tolerance and Absorption Techniques. *J Exp Med* 121, 439-462

63. Locker, G. Y., Hamilton, S., Harris, J., Jessup, J. M., Kemeny, N., Macdonald, J. S., Somerfield, M. R., Hayes, D. F., Bast, R. C., Jr., and Asco (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *J Clin Oncol* 24, 5313-5327

64. Thirunavukarasu, P., Sukumar, S., Sathaiah, M., Mahan, M., Pragatheeshwar, K. D., Pingpank, J. F., Zeh, H., 3rd, Bartels, C. J., Lee, K. K., and Bartlett, D. L. (2011) C-stage in colon cancer: implications of carcinoembryonic antigen biomarker in staging, prognosis, and management. *J Natl Cancer Inst* 103, 689-697

65. McDermott, D., Haanen, J., Chen, T. T., Lorigan, P., O'Day, S., and investigators, M. D. X. (2013) Efficacy and safety of ipilimumab in metastatic melanoma patients surviving more than 2 years following treatment in a phase III trial (MDX010-20). *Ann Oncol* 24, 2694-2698

66. Darvin, P., Toor, S. M., Sasidharan Nair, V., and Elkord, E. (2018) Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp Mol Med* 50, 1-11

67. Cameron, F., Whiteside, G., and Perry, C. (2011) Ipilimumab: first global approval. *Drugs* 71, 1093-1104

68. Vaddepally, R. K., Kharel, P., Pandey, R., Garje, R., and Chandra, A. B. (2020) Review of Indications of FDA-Approved Immune Checkpoint Inhibitors per NCCN Guidelines with the Level of Evidence. *Cancers (Basel)* 12

69. Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M., Anders, R. A., Eshleman, J. R., Vogelstein, B., and Diaz, L. A., Jr. (2015) PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* 372, 2509-2520

70. Harper, J. W., and Bennett, E. J. (2016) Proteome complexity and the forces that drive proteome imbalance. *Nature* 537, 328-338

71. Ho, C. S., Lam, C. W., Chan, M. H., Cheung, R. C., Law, L. K., Lit, L. C., Ng, K. F., Suen, M. W., and Tai, H. L. (2003) Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* 24, 3-12

72. Bruins, A. P. (1998) Mechanistic aspects of electrospray ionization. *Journal of Chromatography A* 794, 345-357

73. Pappireddi, N., Martin, L., and Wuhr, M. (2019) A Review on Quantitative Multiplexed Proteomics. *Chembiochem* 20, 1210-1224

74. Yates, J. R., 3rd (2015) Pivotal role of computers and software in mass spectrometry - SEQUEST and 20 years of tandem MS database searching. *J Am Soc Mass Spectrom* 26, 1804-1813

75. (2017) The problem with neoantigen prediction. *Nat Biotechnol* 35, 97

76. Li, J., Cai, Z., Bomgarden, R. D., Pike, I., Kuhn, K., Rogers, J. C., Roberts, T. M., Gygi, S. P., and Paulo, J. A. (2021) TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J Proteome Res* 20, 2964-2972

77. Rauniyar, N. (2015) Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. *Int J Mol Sci* 16, 28566-28581

78. Deutsch, E. W. (2011) Tandem mass spectrometry spectral libraries and library searching. *Methods Mol Biol* 696, 225-232

79. O'Bryon, I., Jenson, S. C., and Merkley, E. D. (2020) Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide identification. *Protein Sci* 29, 1864-1878

80. Kanaseki, T., and Torigoe, T. (2019) Proteogenomics: advances in cancer antigen research. *Immunol Med* 42, 65-70

81. Hunt, D. F., Michel, H., Dickinson, T. A., Shabanowitz, J., Cox, A. L., Sakaguchi, K., Appella, E., Grey, H. M., and Sette, A. (1992) Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* 256, 1817-1820

82. Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E., and Engelhard, V. H. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255, 1261-1263

83. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976-989

84. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 17, 2337-2342

85. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 11, M111 010587

86. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372

87. Lanoix, J., Durette, C., Courcelles, M., Cossette, E., Comtois-Marotte, S., Hardy, M. P., Cote, C., Perreault, C., and Thibault, P. (2018) Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics* 18, e1700251

88. Pfammatter, S., Bonneil, E., Lanoix, J., Vincent, K., Hardy, M. P., Courcelles, M., Perreault, C., and Thibault, P. (2020) Extending the Comprehensiveness of Immunopeptidome Analyses Using Isobaric Peptide Labeling. *Anal Chem* 92, 9194-9204

89. De Groot, A. S., McMurry, J., and Moise, L. (2008) Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates. *Curr Opin Pharmacol* 8, 620-626

90. Ehx, G., and Perreault, C. (2019) Discovery and characterization of actionable tumor antigens. *Genome Med* 11, 29

91. Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., and Chen, Y. T. (2002) Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol Rev* 188, 22-32

92. Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., Javeri, I., Nellaiappan, K., Salazar, A. M., Daley, H., Seaman, M., Buchbinder, E. I., Yoon, C. H., Harden, M., Lennon, N., Gabriel, S., Rodig, S. J., Barouch, D. H., Aster, J. C., Getz, G., Wucherpfennig, K., Neuberg, D., Ritz, J., Lander, E. S., Fritsch, E. F., Hacohen, N., and Wu, C. J. (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217-221

93. Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., and Holt, R. A. (2014) Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res* 24, 743-750

94. Tran, E., Robbins, P. F., Lu, Y. C., Prickett, T. D., Gartner, J. J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E. M., Kriley, I. R., and Rosenberg, S. A. (2016) T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med* 375, 2255-2262

95. Loffler, M. W., Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., Schroeder, C. M., Trautwein, N., Hilke, F. J., Zinser, R. S., Muhlenbruch, L., Kowalewski, D. J., Schuster, H., Sturm, M., Matthes, J., Riess, O., Czemmel, S., Nahnsen, S., Konigsrainer, I., Thiel, K., Nadalin, S., Beckert, S., Bosmuller, H., Fend, F., Velic, A., Macek, B., Haen, S. P., Buonaguro, L., Kohlbacher, O., Stevanovic, S., Konigsrainer, A., Consortium, H., and Rammensee, H. G. (2019) Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med* 11, 28

96. Saeterdal, I., Bjorheim, J., Lislerud, K., Gjertsen, M. K., Bukholm, I. K., Olsen, O. C., Nesland, J. M., Eriksen, J. A., Moller, M., Lindblom, A., and Gaudernack, G. (2001) Frameshift-mutation-derived peptides as tumor-specific antigens in inherited and spontaneous colorectal cancer. *Proc Natl Acad Sci U S A* 98, 13255-13260

97. Linnebacher, M., Gebert, J., Rudy, W., Woerner, S., Yuan, Y. P., Bork, P., and von Knebel Doeberitz, M. (2001) Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens. *Int J Cancer* 93, 6-11

98. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T. R. (2012) Landscape of transcription in human cells. *Nature* 489, 101-108

99. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, E., Caron-Lizotte, O., Hardy, M. P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P., and Perreault, C. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* 7, 10238

100. Minati, R., Perreault, C., and Thibault, P. (2020) A Roadmap Toward the Definition of Actionable Tumor-Specific Antigens. *Front Immunol* 11, 583287

101. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, E., Bonneil, E., Laverdure, J. P., Gendron, P., Courcelles, M., Hardy, M. P., Cote, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P., and Perreault, C. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* 10

102. Zhao, Q., Laverdure, J. P., Lanoix, J., Durette, C., Cote, C., Bonneil, E., Laumont, C. M., Gendron, P., Vincent, K., Courcelles, M., Lemieux, S., Millar, D. G., Ohashi, P. S., Thibault, P., and Perreault, C. (2020) Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. *Cancer Immunol Res* 8, 544-555

103. Ehx, G., Larouche, J. D., Durette, C., Laverdure, J. P., Hesnard, L., Vincent, K., Hardy, M. P., Theriault, C., Rulleau, C., Lanoix, J., Bonneil, E., Feghaly, A., Apavaloaei, A., Noronha, N., Laumont, C. M., Delisle, J. S., Vago, L., Hebert, J., Sauvageau, G., Lemieux, S., Thibault, P., and Perreault, C. (2021) Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity*

104. Newey, A., Griffiths, B., Michaux, J., Pak, H. S., Stevenson, B. J., Woolston, A., Semiannikova, M., Spain, G., Barber, L. J., Matthews, N., Rao, S., Watkins, D., Chau, I., Coukos, G., Racle, J., Gfeller, D., Starling, N., Cunningham, D., Bassani-Sternberg, M., and Gerlinger, M. (2019) Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. *J Immunother Cancer* 7, 309

105. Hirama, T., Tokita, S., Nakatsugawa, M., Murata, K., Nannya, Y., Matsuo, K., Inoko, H., Hirohashi, Y., Hashimoto, S., Ogawa, S., Takemasa, I., Sato, N., Hata, F., Kanaseki, T., and Torigoe, T. (2021) Proteogenomic identification of an immunogenic HLA class I neoantigen in mismatch repair-deficient colorectal cancer tissue. *JCI Insight* 6

106. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (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

107. Eriksen, A. C., Sorensen, F. B., Lindebjerg, J., Hager, H., dePont Christensen, R., Kjaer-Frifeldt, S., and Hansen, T. F. (2018) The Prognostic Value of Tumor-Infiltrating lymphocytes in Stage II Colon Cancer. A Nationwide Population-Based Study. *Transl Oncol* 11, 979-987

108. Zhao, Y., Ge, X., He, J., Cheng, Y., Wang, Z., Wang, J., and Sun, L. (2019) The prognostic value of tumor-infiltrating lymphocytes in colorectal cancer differs by anatomical subsite: a systematic review and meta-analysis. *World J Surg Oncol* 17, 85

109. Fabrizio, D. A., George, T. J., Jr., Dunne, R. F., Frampton, G., Sun, J., Gowen, K., Kennedy, M., Greenbowe, J., Schrock, A. B., Hezel, A. F., Ross, J. S., Stephens, P. J., Ali, S. M., Miller, V. A., Fakih, M., and Klempner, S. J. (2018) Beyond microsatellite testing: assessment of tumor mutational burden identifies subsets of colorectal cancer who may respond to immune checkpoint inhibition. *J Gastrointest Oncol* 9, 610-617

110. Wagner, S., Mullins, C. S., and Linnebacher, M. (2018) Colorectal cancer vaccines: Tumor-associated antigens vs neoantigens. *World J Gastroenterol* 24, 5418-5432

111. Loffler, M. W., Kowalewski, D. J., Backert, L., Bernhardt, J., Adam, P., Schuster, H., Dengler, F., Backes, D., Kopp, H. G., Beckert, S., Wagner, S., Konigsrainer, I., Kohlbacher, O., Kanz, L., Konigsrainer, A., Rammensee, H. G., Stevanovic, S., and Haen, S. P. (2018) Mapping the HLA Ligandome of Colorectal Cancer Reveals an Imprint of Malignant Cell Transformation. *Cancer Res* 78, 4627-4641

112. Picard, E., Verschoor, C. P., Ma, G. W., and Pawelec, G. (2020) Relationships Between Immune Landscapes, Genetic Subtypes and Responses to Immunotherapy in Colorectal Cancer. *Front Immunol* 11, 369

113. Parkhurst, M. R., Yang, J. C., Langan, R. C., Dudley, M. E., Nathan, D. A., Feldman, S. A., Davis, J. L., Morgan, R. A., Merino, M. J., Sherry, R. M., Hughes, M. S., Kammula, U. S., Phan, G. Q., Lim, R. M., Wank, S. A., Restifo, N. P., Robbins, P. F., Laurencot, C. M., and Rosenberg, S. A. (2011) T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Mol Ther* 19, 620-626

114. Smith, C. C., Selitsky, S. R., Chai, S., Armistead, P. M., Vincent, B. G., and Serody, J. S. (2019) Alternative tumour-specific antigens. *Nat Rev Cancer* 19, 465-478

115. Kloor, M., Reuschenbach, M., Karbach, J., Rafiyan, M., Al-Batran, S.-E., Pauligk, C., Jaeger, E., and Doeberitz, M. v. K. (2015) Vaccination of MSI-H colorectal cancer patients with frameshift peptide antigens: A phase I/IIa clinical trial. *Journal of Clinical Oncology* 33, 3020-3020

116. van den Bulk, J., Verdegaal, E. M. E., Ruano, D., Ijsselsteijn, M. E., Visser, M., van der Breggen, R., Duhen, T., van der Ploeg, M., de Vries, N. L., Oosting, J., Peeters, K., Weinberg, A. D., Farina-Sarasqueta, A., van der Burg, S. H., and de Miranda, N. (2019) Neoantigen-specific immunity in low mutation burden colorectal cancers of the consensus molecular subtype 4. *Genome Med* 11, 87

117. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114-2120

118. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21

119. Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525-527

120. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014) OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 30, 3310-3316

121. Jia, P., Yang, X., Guo, L., Liu, B., Lin, J., Liang, H., Sun, J., Zhang, C., and Ye, K. (2020) MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. *Genomics Proteomics Bioinformatics* 18, 65-71

122. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550

123. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., and Chanda, S. K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 10, 1523

124. Hardy, M. P., Audemard, E., Migneault, F., Feghaly, A., Brochu, S., Gendron, P., Boilard, E., Major, F., Dieude, M., Hebert, M. J., and Perreault, C. (2019) Apoptotic endothelial cells release small extracellular vesicles loaded with immunostimulatory viral-like RNAs. *Sci Rep* 9, 7203

125. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32, D493-496

126. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80-92

127. Daouda, T., Perreault, C., and Lemieux, S. (2016) pyGeno: A Python package for precision medicine and proteogenomics. *F1000Res* 5, 381

128. Lanoix, J., Durette, C., Courcelles, M., Cossette, E., Comtois-Marotte, S., Hardy, M. P., Côté, C., Perreault, C., and Thibault, P. (2018) Comparison of the MHC I immunopeptidome repertoir of B-cell lymphoblasts using two isolation methods. *Proteomics* 18, e1700251

129. Courcelles, M., Durette, C., Daouda, T., Laverdure, J. P., Vincent, K., Lemieux, S., Perreault, C., and Thibault, P. (2020) MAPDP: A Cloud-Based Computational Platform for Immunopeptidomics Analyses. *J Proteome Res* 19, 1873-1881

130. Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016) GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods Mol Biol* 1418, 283-334

131. Ogishi, M., and Yotsuyanagi, H. (2019) Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front Immunol* 10, 827

132. Conway, J. R., Lex, A., and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938-2940

133. Hanzelmann, S., Castelo, R., and Guinney, J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14, 7

134. Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Trevino, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., and Verhaak, R. G. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 4, 2612

135. Chan, A. Y., and Anderson, M. S. (2015) Central tolerance to self revealed by the autoimmune regulator. *Ann N Y Acad Sci* 1356, 80-89

136. Pira, G., Uva, P., Scanu, A. M., Rocca, P. C., Murgia, L., Uleri, E., Piu, C., Porcu, A., Carru, C., Manca, A., Persico, I., Muroni, M. R., Sanges, F., Serra, C., Dolei, A., Angius, A., and De Miglio, M. R. (2020) Landscape of transcriptome variations uncovering known and novel driver events in colorectal carcinoma. *Sci Rep* 10, 432

137. Kawakami, H., Zaanan, A., and Sinicrope, F. A. (2015) Microsatellite instability testing and its role in the management of colorectal cancer. *Curr Treat Options Oncol* 16, 30

138. Jimeno, A., Messersmith, W. A., Hirsch, F. R., Franklin, W. A., and Eckhardt, S. G. (2009) KRAS mutations and sensitivity to epidermal growth factor receptor inhibitors in colorectal cancer: practical application of patient selection. *J Clin Oncol* 27, 1130-1136

139. Van Cutsem, E., Kohne, C. H., Lang, I., Folprecht, G., Nowacki, M. P., Cascinu, S., Shchepotin, I., Maurel, J., Cunningham, D., Tejpar, S., Schlichting, M., Zubel, A., Celik, I., Rougier,

P., and Ciardiello, F. (2011) Cetuximab plus irinotecan, fluorouracil, and leucovorin as first-line treatment for metastatic colorectal cancer: updated analysis of overall survival according to tumor KRAS and BRAF mutation status. *J Clin Oncol* 29, 2011-2019

140. Ahmed, D., Eide, P. W., Eilertsen, I. A., Danielsen, S. A., Eknaes, M., Hektoen, M., Lind, G. E., and Lothe, R. A. (2013) Epigenetic and genetic features of 24 colon cancer cell lines. *Oncogenesis* 2, e71

141. Berg, K. C. G., Eide, P. W., Eilertsen, I. A., Johannessen, B., Bruun, J., Danielsen, S. A., Bjornslett, M., Meza-Zepeda, L. A., Eknaes, M., Lind, G. E., Myklebost, O., Skotheim, R. I., Sveen, A., and Lothe, R. A. (2017) Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. *Mol Cancer* 16, 116

142. Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M. D., Wendl, M. C., and Ding, L. (2014) MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* 30, 1015-1016

143. Aaltonen, L. A., Peltomaki, P., Mecklin, J. P., Jarvinen, H., Jass, J. R., Green, J. S., Lynch, H. T., Watson, P., Tallqvist, G., Juhola, M., and et al. (1994) Replication errors in benign and malignant tumors from hereditary nonpolyposis colorectal cancer patients. *Cancer Res* 54, 1645-1648

144. Llosa, N. J., Cruise, M., Tam, A., Wicks, E. C., Hechenbleikner, E. M., Taube, J. M., Blosser, R. L., Fan, H., Wang, H., Luber, B. S., Zhang, M., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Sears, C. L., Anders, R. A., Pardoll, D. M., and Housseau, F. (2015) The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov* 5, 43-51

145. Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., Lu, S., Kemberling, H., Wilt, C., Luber, B. S., Wong, F., Azad, N. S., Rucki, A. A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Greten, T. F., Duffy, A. G., Ciombor, K. K., Eyring, A. D., Lam, B. H., Joe, A., Kang, S. P., Holdhoff, M., Danilova, L., Cope, L., Meyer, C., Zhou, S., Goldberg, R. M., Armstrong, D. K., Bever, K. M., Fader, A. N., Taube, J., Housseau, F., Spetzler, D., Xiao, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Eshleman, J. R., Vogelstein, B., Anders, R. A., and Diaz, L. A., Jr. (2017) Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 357, 409-413

146. Prossomariti, A., Piazzi, G., Alquati, C., and Ricciardiello, L. (2020) Are Wnt/beta-Catenin and PI3K/AKT/mTORC1 Distinct Pathways in Colorectal Cancer? *Cell Mol Gastroenterol Hepatol* 10, 491-506

147. Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., Geller, M. A., Odunsi, K., Beechem, J., and Fling, S. P. (2017) Gene expression markers of Tumor Infiltrating Leukocytes. *J Immunother Cancer* 5, 18

148. Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Frohling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T., and Hahn, W. C. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108-112

149. Kim, H., Jen, J., Vogelstein, B., and Hamilton, S. R. (1994) Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences. *Am J Pathol* 145, 148-156

150. Smyrk, T. C., Watson, P., Kaul, K., and Lynch, H. T. (2001) Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. *Cancer* 91, 2417-2422

151. Dolcetti, R., Viel, A., Doglioni, C., Russo, A., Guidoboni, M., Capozzi, E., Vecchiato, N., Macri, E., Fornasarig, M., and Boiocchi, M. (1999) High prevalence of activated intraepithelial cytotoxic T lymphocytes and increased neoplastic cell apoptosis in colorectal carcinomas with microsatellite instability. *Am J Pathol* 154, 1805-1813

152. Phillips, S. M., Banerjea, A., Feakins, R., Li, S. R., Bustin, S. A., and Dorudi, S. (2004) Tumour-infiltrating lymphocytes in colorectal cancer with microsatellite instability are activated and cytotoxic. *Br J Surg* 91, 469-475

153. Boland, C. R., Koi, M., Chang, D. K., and Carethers, J. M. (2008) The biochemical basis of microsatellite instability and abnormal immunohistochemistry and clinical behavior in Lynch syndrome: from bench to bedside. *Fam Cancer* 7, 41-52

154. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M.,

Ouspenskaia, T., Law, T., Justesen, S., Stevens, J., Lane, W. J., Eisenhaure, T., Lan Zhang, G., Clauser, K. R., Hacohen, N., Carr, S. A., Wu, C. J., and Keskin, D. B. (2020) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* 38, 199-209

155. Larouche, J. D., Trofimov, A., Hesnard, L., Ehx, G., Zhao, Q., Vincent, K., Durette, C., Gendron, P., Laverdure, J. P., Bonneil, E., Cote, C., Lemieux, S., Thibault, P., and Perreault, C. (2020) Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med* 12, 40

156. Cherkasova, E., Scrivani, C., Doh, S., Weisman, Q., Takahashi, Y., Harashima, N., Yokoyama, H., Srinivasan, R., Linehan, W. M., Lerman, M. I., and Childs, R. W. (2016) Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer. *Cancer Res* 76, 2177-2185

157. Patra, R., Das, N. C., and Mukherjee, S. (2021) Exploring the Differential Expression and Prognostic Significance of the COL11A1 Gene in Human Colorectal Carcinoma: An Integrated Bioinformatics Approach. *Front Genet* 12, 608313

158. Yi, X., Liao, Y., Wen, B., Li, K., Dou, Y., Savage, S. R., and Zhang, B. (2021) caAtlas: An immunopeptidome atlas of human cancer. *iScience* 24, 103107

159. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D. J., Freudenmann, L. K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., Engler, T., Matovina, S., Wang, J., Hauri-Hohl, M., Martin, R., Kapolou, K., Walz, J. S., Velz, J., Moch, H., Regli, L., Silginer, M., Weller, M., Löffler, M. W., Erhard, F., Schlosser, A., Kohlbacher, O., Stevanović, S., Rammensee, H.-G., and Neidert, M. C. (2020) The HLA Ligand Atlas - A resource of natural HLA ligands presented on benign tissues. *bioRxiv*, 778944

160. Gjerstorff, M. F., Andersen, M. H., and Ditzel, H. J. (2015) Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* 6, 15772-15787

161. Ishikawa, N., Takano, A., Yasui, W., Inai, K., Nishimura, H., Ito, H., Miyagi, Y., Nakayama, H., Fujita, M., Hosokawa, M., Tsuchiya, E., Kohno, N., Nakamura, Y., and Daigo, Y. (2007) Cancer-testis antigen lymphocyte antigen 6 complex locus K is a serologic biomarker and a therapeutic target for lung and esophageal carcinomas. *Cancer Res* 67, 11601-11611

162. Adamopoulou, E., Tenzer, S., Hillen, N., Klug, P., Rota, I. A., Tietz, S., Gebhardt, M., Stevanovic, S., Schild, H., Tolosa, E., Melms, A., and Stoeckle, C. (2013) Exploring the MHC-peptide matrix of central tolerance in the human thymus. *Nat Commun* 4, 2039

163. Kote, S., Pirog, A., Bedran, G., Alfaro, J., and Dapic, I. (2020) Mass Spectrometry-Based Identification of MHC-Associated Peptides. *Cancers (Basel)* 12

164. Lin, A., Zhang, J., and Luo, P. (2020) Crosstalk Between the MSI Status and Tumor Microenvironment in Colorectal Cancer. *Front Immunol* 11, 2039

165. Bonaventura, P., Shekarian, T., Alcazer, V., Valladeau-Guilemond, J., Valsesia-Wittmann, S., Amigorena, S., Caux, C., and Depil, S. (2019) Cold Tumors: A Therapeutic Challenge for Immunotherapy. *Front Immunol* 10, 168

166. Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., and Gaunt, T. R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34, 57-65

167. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 47, D886-D894

168. Sherry, S. T., Ward, M., and Sirotkin, K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 9, 677-679

169. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J., and Forbes, S. A. (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 47, D941-D947

170. Vogel, C., and Marcotte, E. M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 13, 227-232

171. Xiang, B., Snook, A. E., Magee, M. S., and Waldman, S. A. (2013) Colorectal cancer immunotherapy. *Discov Med* 15, 301-308

172. Zhou, F. (2009) Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int Rev Immunol* 28, 239-260

173. Du, W., Frankel, T. L., Green, M., and Zou, W. (2022) IFNγ signaling integrity in colorectal cancer immunity and immunotherapy. *Cellular & Molecular Immunology* 19, 23-32

174. Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., Garcia-Seisdedos, D., Jarnuczak, A. F., Hewapathirana, S., Pullman, B. S., Wertz, J., Sun, Z., Kawano, S., Okuda, S., Watanabe, Y., Hermjakob, H., MacLean, B., MacCoss, M. J., Zhu, Y., Ishihama, Y., and Vizcaino, J. A. (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res* 48, D1145-D1152

175. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Perez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., and Vizcaino, J. A. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47, D442-D450

176. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210

177. Crick, F. H. (1958) On protein synthesis. *Symp Soc Exp Biol* 12, 138-163

178. (1970) Central dogma reversed. *Nature* 226, 1198-1199

179. Liyanage, V. R., Jarmasz, J. S., Murugeshan, N., Del Bigio, M. R., Rastegar, M., and Davie, J. R. (2014) DNA modifications: function and applications in normal and disease States. *Biology (Basel)* 3, 670-723

180. Lelli, K. M., Slattery, M., and Mann, R. S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu Rev Genet* 46, 43-68

181. Ramazi, S., and Zahiri, J. (2021) Posttranslational modifications in proteins: resources, tools and prediction methods. *Database (Oxford)* 2021

182. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst* 5, 1512-1526

183. Riley, T. P., Keller, G. L. J., Smith, A. R., Davancaze, L. M., Arbuiso, A. G., Devlin, J. R., and Baker, B. M. (2019) Structure Based Prediction of Neoantigen Immunogenicity. *Front Immunol* 10, 2047

184. Calis, J. J., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., Kesmir, C., and Peters, B. (2013) Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* 9, e1003266

185. Jawa, V., Cousens, L. P., Awwad, M., Wakshull, E., Kropshofer, H., and De Groot, A. S. (2013) T-cell dependent immunogenicity of protein therapeutics: Preclinical assessment and mitigation. *Clin Immunol* 149, 534-555

186. Rius, C., Attaf, M., Tungatt, K., Bianchi, V., Legut, M., Bovay, A., Donia, M., Thor Straten, P., Peakman, M., Svane, I. M., Ott, S., Connor, T., Szomolay, B., Dolton, G., and Sewell, A. K. (2018) Peptide-MHC Class I Tetramers Can Fail To Detect Relevant Functional T Cell Clonotypes and Underestimate Antigen-Reactive T Cell Populations. *J Immunol* 200, 2263-2279

187. Simoni, Y., Becht, E., Fehlings, M., Loh, C. Y., Koo, S.-L., Teng, K. W. W., Yeong, J. P. S., Nahar, R., Zhang, T., Kared, H., Duan, K., Ang, N., Poidinger, M., Lee, Y. Y., Larbi, A., Khng, A. J., Tan, E., Fu, C., Mathew, R., Teo, M., Lim, W. T., Toh, C. K., Ong, B.-H., Koh, T., Hillmer, A. M., Takano, A., Lim, T. K. H., Tan, E. H., Zhai, W., Tan, D. S. W., Tan, I. B., and Newell, E. W. (2018) Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature* 557, 575-579

188. Scheper, W., Kelderman, S., Fanchi, L. F., Linnemann, C., Bendle, G., de Rooij, M. A. J., Hirt, C., Mezzadra, R., Slagter, M., Dijkstra, K., Kluin, R. J. C., Snaebjornsson, P., Milne, K., Nelson, B. H., Zijlmans, H., Kenter, G., Voest, E. E., Haanen, J., and Schumacher, T. N. (2019) Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nat Med* 25, 89-94

189. Brennick, C. A., George, M. M., Moussa, M. M., Hagymasi, A. T., Seesi, S. A., Shcheglova, T. V., Englander, R. P., Keller, G. L., Balsbaugh, J. L., Baker, B. M., Schietinger, A., Mandoiu, II, and Srivastava, P. K. (2021) An unbiased approach to defining bona fide cancer neoepitopes that elicit immune-mediated cancer rejection. *J Clin Invest* 131

190. Guil-Luna, S., Sedlik, C., and Piaggio, E. (2021) Humanized Mouse Models to Evaluate Cancer Immunotherapeutics. *Annual Review of Cancer Biology* 5, 119-136

191. Morotti, M., Albukhari, A., Alsaadi, A., Artibani, M., Brenton, J. D., Curbishley, S. M., Dong, T., Dustin, M. L., Hu, Z., McGranahan, N., Miller, M. L., Santana-Gonzalez, L., Seymour, L. W., Shi, T., Van Loo, P., Yau, C., White, H., Wietek, N., Church, D. N., Wedge, D. C., and Ahmed, A. A. (2021)

Promises and challenges of adoptive T-cell therapies for solid tumours. *Br J Cancer* 124, 1759-1776

192. Hollingsworth, R. E., and Jansen, K. (2019) Turning the corner on therapeutic cancer vaccines. *NPJ Vaccines* 4, 7

193. Santos, P. M., and Butterfield, L. H. (2018) Dendritic Cell-Based Cancer Vaccines. *J Immunol* 200, 443-449

194. Kantoff, P. W., Higano, C. S., Shore, N. D., Berger, E. R., Small, E. J., Penson, D. F., Redfern, C. H., Ferrari, A. C., Dreicer, R., Sims, R. B., Xu, Y., Frohlich, M. W., Schellhammer, P. F., and Investigators, I. S. (2010) Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 363, 411-422

195. Jobsis, A. C., De Vries, G. P., Meijer, A. E., and Ploem, J. S. (1981) The immunohistochemical detection of prostatic acid phosphatase: its possibilities and limitations in tumour histochemistry. *Histochem J* 13, 961-973

196. Bezu, L., Kepp, O., Cerrato, G., Pol, J., Fucikova, J., Spisek, R., Zitvogel, L., Kroemer, G., and Galluzzi, L. (2018) Trial watch: Peptide-based vaccines in anticancer therapy. *Oncoimmunology* 7, e1511506

197. Kim, V. M., Pan, X., Soares, K. C., Azad, N. S., Ahuja, N., Gamper, C. J., Blair, A. B., Muth, S., Ding, D., Ladle, B. H., and Zheng, L. (2020) Neoantigen-based EpiGVAX vaccine initiates antitumor immunity in colorectal cancer. *JCI Insight* 5

198. Apavaloaei, A., Hardy, M. P., Thibault, P., and Perreault, C. (2020) The Origin and Immune Recognition of Tumor-Specific Antigens. *Cancers (Basel)* 12

199. Benhammadi, M., Mathe, J., Dumont-Lagace, M., Kobayashi, K. S., Gaboury, L., Brochu, S., and Perreault, C. (2020) IFN-lambda Enhances Constitutive Expression of MHC Class I Molecules on Thymic Epithelial Cells. *J Immunol* 205, 1268-1280

200. Shen, L., and Rock, K. L. (2004) Cellular protein is the source of cross-priming antigen in vivo. *Proc Natl Acad Sci U S A* 101, 3035-3040

201. Paston, S. J., Brentville, V. A., Symonds, P., and Durrant, L. G. (2021) Cancer Vaccines, Adjuvants, and Delivery Systems. *Front Immunol* 12, 627932

202. Wang, M., and Claesson, M. H. (2014) Classification of human leukocyte antigen (HLA) supertypes. *Methods Mol Biol* 1184, 309-317

203. Fernandez Vina, M. A., Hollenbach, J. A., Lyke, K. E., Sztein, M. B., Maiers, M., Klitz, W., Cano, P., Mack, S., Single, R., Brautbar, C., Israel, S., Raimondi, E., Khoriaty, E., Inati, A., Andreani, M., Testi, M., Moraes, M. E., Thomson, G., Stastny, P., and Cao, K. (2012) Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci* 367, 820-829

204. Mohme, M., Neidert, M. C., Regli, L., Weller, M., and Martin, R. (2014) Immunological challenges for peptide-based immunotherapy in glioblastoma. *Cancer Treat Rev* 40, 248-258

205. Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021) The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 127, 3029-3030