High-grade serous ovarian cancer (HGSC), the principal cause of death from gynecologic malignancies in the world, has not significantly benefited from advances in cancer immunotherapy. Although HGSC infiltration by lymphocytes correlates with superior survival, the nature of antigens that can elicit anti-HGSC immune responses is unknown. The goal of this study was to establish the global landscape of HGSC tumor-specific antigens (TSA) using a mass spectrometry pipeline that interrogated all reading frames of all genomic regions. In 23 HGSC tumors, we identified 103 TSAs. Classic TSA discovery approaches focusing only on mutated exonic sequences would have uncovered only three of these TSAs. Other mutated TSAs resulted from out-of-frame exonic translation (n = 2) or from noncoding sequences (n = 7). One group of TSAs (n = 91) derived from aberrantly expressed unmutated genomic sequences, which were not expressed in normal tissues. These aberrantly expressed TSAs (aTSA) originated primarily from nonexonic sequences, in particular intronic (29%) and intergenic (22%) sequences. Their expression was regulated at the transcriptional level by variations in gene copy number and DNA methylation. Although mutated TSAs were unique to individual tumors, aTSA were shared by a large proportion of HGSCs. Taking into account the frequency of aTSA expression and HLA allele frequencies, we calculated that, in Caucasians, the median number of aTSA per tumor would be five. We conclude that, in view of their number and the fact that they are shared by many tumors, aTSA may be the most attractive targets for HGSC immunotherapy.

Introduction

Ovarian cancer is the principal cause of death from gynecologic malignancies in the world and is responsible for over 14,000 deaths per year in the United States (1). High-grade serous ovarian cancer (HGSC) accounts for 70% to 80% of these deaths, and overall survival has not changed significantly for decades (2). The positive correlation between the abundance of tumor-infiltrating lymphocytes (TIL) and increased overall survival hints that T cells can recognize biologically relevant tumor antigens in HGSC (3, 4). Strong evidence suggests that HGSC TILs adjacent to tumor epithelial cells are actively engaged in local immune editing. Indeed, in a multimodality study of 212 HGSC samples from 38 patients, CD8+ TILs negatively associate with malignant cell diversity (5). In line with this, and given the therapeutic efficacy of immune-checkpoint inhibitors in several tumor types, clinical trials with one or more checkpoint inhibitors are currently under way in HGSC. However, initial trials with anti–PD-1 have shown limited activity in HGSC (6, 7). In view of this, there is a pressing need to identify the antigens that can elicit therapeutic anti-HGSC immune responses (3, 8). Such antigens could be used as vaccines (± immune-checkpoint inhibitors) or as targets for T-cell receptor–based approaches (cell therapy, bispecific biologics; ref. 9).

Rapid progress in the field and lack of a standardized nomenclature commonly lead to some confusion in the classification of tumor antigens. For the sake of clarity, we therefore used throughout this article a proposed classification of tumor antigens into three discrete categories: tumor-associated antigens (TAA), mutated tumor-specific antigens (mTSA), and aberrantly expressed tumor-specific antigens (aTSA; ref. 10). TAA are MHC-associated peptides (MAP) that show superior abundance on tumor cells but are nonetheless present on normal cells and, therefore, may induce central immune tolerance (11–13). mTSA derive from mutated DNA sequences that can be either exonic or nonexonic (10, 14). aTSA result from aberrant expression of unmutated transcripts that are not expressed in any normal somatic cell, including medullary thymic epithelial cells (mTEC), which orchestrate central immune tolerance. Expression of aTSA results from cancer-specific epigenetic changes. In one study, aTSA are found to represent the vast majority of TSAs in acute lymphoblastic leukemia and lung cancer (15). Finally, a peculiar antigen family, cancer-germline antigens (CGA), is astride the TAA and aTSA categories. CGAs are coded by canonical exons normally expressed only by germ cells, and their aberrant expression in cancer cells is driven mainly by epigenetic alterations. However, some CGAs...
Materials and Methods

Human HGSC samples

Tumor fragments of HGSCI-6 and matched normal adjacent tissues of HGSCI-3 were obtained from Tissue Solutions. High-grade serous tumor tissue (OV606) or ascites (OV633 and OV642) were obtained from the Princess Margaret Cancer Centre Biospecimen Program (Toronto, ON, Canada) under protocols approved by the UHN Research Ethics Board. All patients gave informed written consent. Snap-frozen samples described above were used for RNA extraction and MHCI-associated peptide isolation (described below). RNA sequencing (RNA-seq) data for OvCa48-114 were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive under project PRJNA398141, converted to FASTQ file using SRA Toolkit (https://github.com/ncbi/sra-tools), and processed like the other samples. MS raw data of the samples in this cohort were downloaded from ProteomeXchange Consortium via the PRIDE partner with identical accession numbers. Libraries were further amplified with TruSeq primers (Illumina) by 7 to 10 cycles, and 1 pmol/L library per sample was used for paired-end RNA-seq on the Illumina HiSeq 2000 or the NextSeq 500 platform, which yielded 150 to 300 million reads.

Generation of customized reference databases for MS analyses

For each sample, we generated a customized “global cancer database” by concatenating two modules, the “cancer canonical proteome” and the “cancer-specific proteome” as previously described (15). Briefly, RNA-seq reads were trimmed by Trimomatic v0.35 (24) and aligned to the reference human genome version GRCh38.88 using STAR v2.5.1b (https://github.com/alexdobin/STAR/releases/tag/2.5.1b). Transcript expression was quantified in transcripts per million (tpm) with kallisto v0.43.0 (https://pachterlab.github.io/kallisto) using default parameters. A sample-specific exome for each sample was then built with pyGeno (25) by inserting single-base variants (quality >=20) identified using FreeBayes (https://github.com/ekg/freebayes). Annotated open reading frames (ORF) with tpm >0 were in silico translated from sample-specific exome using pyGeno and added into the canonical cancer proteome in a fasta format.

For a paired-end run, Illumina sequencing generates two FASTQ files: Read1 and Read2. To generate cancer-specific proteomes, trimmed reads from Read1 FASTQ file were reverse complemented and were used for 33- and 24-nucleotide-long k-mer database generation together with trimmed reads from Read2 FASTQ file. In order to exclude sequencing errors and to limit database size, sample-specific thresholds of minimal occurrence were applied for 33 nucleotides: 7 for HGSCI-3, 8 for OV642, 10 for OV633, 4 for HGSC4 and OV606, 6 for HGSC5, 5 for HGSC6, and 3 for OvCa48-114. Cancer-specific proteins were obtained by subtraction of 33-mers expressed in purified human thymic epithelial cells (TEC) harvested from six human thymus (15), then assembled into longer sequences (contigs) by the kmers_assembly tool from NEKTAR (https://github.com/iric-soft/nektar). Contigs >34 nucleotides long were 3-frame translated and split at internal stop codons. The resulting subsequences longer than eight amino acids were included in the relevant cancer-specific proteome.

Isolation of MAPs

Tumor fragments of HGSCI-6 and OV606 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 UltraTurrax T25 homogenizer (IKA-Labortechnik) set at speed 20,000 rpm and then 20 seconds using an UltraTurrax T8 homogenizer (IKA-Labortechnik) set at speed 25,000 rpm. Then, 550 μL of ice-cold 10× lysis buffer (5% w/v CHAPS, Sigma, cat. # C9426-5G) was added to each sample. After a 60-minute incubation with tumbling at 4°C, samples were spun at 10,000 × g for 30 minutes at 4°C. Supernatants were transferred into new tubes containing 1 mg of W6/32 antibody (Bio X Cell, cat. #BE0079) covalently cross-linked to 1 mL of protein A magnetic beads (Pure Proteome, cat. #LSKMAG10) using dimethyl-pimelidate (26), and MAPs were immunoprecipitated as previously described (27). MAP extracts were then dried using a Speed-Vac and kept frozen at −20°C until MS analyses.

Liquid chromatography-tandem MS analyses

Dried peptide extracts were resuspended in 0.2% formic acid and loaded on a homemade C18 analytical column (15 cm × 150 μm i.d. packed with C18 Jupiter Phenomenex) with a 56-minute (HGSCI-6) or 106-minute (OV606, OV642, and OV633) gradient from 0% to 30% acetonitrile (0.2% formic acid) and a 600 nL/min flow rate on an Easy-Q LC/MS 2 QTRAP. Libraries were further amplified with TruSeq primers (Illumina) by 7 to 10 cycles, and 1 pmol/L library per sample was used for paired-end RNA-seq on the Illumina HiSeq 2000 or the NextSeq 500 platform, which yielded 150 to 300 million reads.
nLC II system. Samples were analyzed with a Q-Exact mass spectrometer (Thermo Fisher Scientific) in positive ion mode with NanoSpray 2 source at 1.6 kV. Each full MS spectrum, acquired with a 60,000 resolution, was followed by 20 tandem MS (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 5 × 10^4, an injection time of 100 ms, and collision energy of 25%. For HGSC4-6, each full MS spectrum, acquired with a 60,000 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 2 × 10^4, an injection time of 800 ms and collision energy of 25%.

Identification of MAPs

All liquid chromatography (LC)-MS/MS (LC-MS/MS) data were searched against the relevant global cancer database using PEAKS 8.5 or PEAKS X (Biorbitmics Solution Inc.). For peptide identification, tolerance was set at 10 ppm and 0.01 Da for precursor and fragment ions, respectively. For the reanalysis of data from Schuster and colleagues (13), tolerance was set at 5 ppm and 0.5 Da for precursor and fragment ions, respectively. The occurrences of oxidation (M) and deamidation (NQ) were set as variable modifications. The occurrences of carbamyl (C) and methionine oxidation (M) were set as fixed modifications. Tolerance was set at 10 ppm and 0.01 Da for precursor and fragment ions, respectively. The occurrences of oxidation (M) and deamidation (NQ) were set as variable modifications. The tolerance was set at 5 ppm and 0.01 Da for precursor and fragment ions, respectively.

For the identification of endogenous peptides, a Laboratories (13), a Tribrid Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) coupled to an Easy nLC1000 was used. Peptides were separated on a custom C18 reversed phase column (150 μm i.d. × 100 mm, C18 Jupiter, Phenomenex) using a flow rate of 600 nL/min and a linear gradient of 5% to 30% acetonitrile (0.2% FA) in 56 minutes. Survey scan (MS1) were acquired with the Orbitrap at a resolving power of 120,000 (at m/z 200) over a scan range of 350 to 1200 m/z with a target values of 5 × 10^6 and a maximum injection time of 50 ms. MS/MS spectra were acquired in collision-induced dissociation with a normalized collision energy of 30. Up to 20 precursor ions were accumulated with a precursor isolation window of 1.3 m/z, an advanced gain control of 2 × 10^4 with a maximum injection time of 50 ms.

Spectra correlations were computed using a script written in Python v3.7.2 using the following steps. The first step reads the list of peptides and computes expected peptide fragments (b/y ions, singly and doubly charged, water and ammonia losses) with pyteomics v4.0.1 (https://github.com/levitas/pyteomics). The next step searches reproducibly detected peptide fragments using multiple synthetic peptide scans. The list of peptide scans is provided by PEAKS X search result file and MS raw files are read using Thermo MSFileReader Python bindings (pre-release version, https://github.com/mobiusklein/MSFileReader-Python-bindings). Fragment tolerance was set to 0.5 Da for collision-induced dissociation and 0.02 Da for higher-energy collisional dissociation scans. Thereafter, root scaled intensities of reproducible fragments were correlated with between all corresponding endogenous and synthetic peptide scan pairs. The Pearson correlation coefficient, P value, and confidence interval were computed using SciPy v1.2.1 (https://www.scipy.org/). Finally, the scan pair of each peptide with the lowest P value was retained to generate a mirror plot using spectra-utils v0.2.1 (https://github.com/bittremieux/spectrum_utils). Peptides with poor Pearson correlation score were inspected manually. This occurred for peptides identified from the data reanalysis of Schuster and colleagues (13), due to the usage of different mass spectrometers.

TSA validation with synthetic peptides

We confirmed eight TSAs from HGSC1-3 using an isobaric peptide labeling approach with corresponding specimen (Supplementary Fig. S2). With remaining peptide extracts from HGSC1-3, we used an isobaric peptide labeling approach in confirmation of 8 expected TSAs (TASDLNLKV, TEINSNQQA, RTATPLTMK, RTATPLTMKK, SVYMATTLK, SQGFSHSQM, STQMTITFTQK, and VTIDTTQTK) in the peptide extracts of the corresponding specimens. Endogenous peptide extracts were labeled with TMT6-127 (Thermo Fisher Scientific) following the manufacturer's protocol, whereas the synthetic peptides (IPT PeptideTechnologies) were derivatized with TMT6-129 and TMT6-131 for spiked levels of 10 fmol and 200 fmol, respectively. Peptides and corresponding extract were loaded on a homemade C18 analytical column (15 cm × 150 μm i.d. packed with C18
The Global Landscape of Ovarian Cancer-Specific Antigens

Jupiter Phenomenex) with a 56-minute gradient from 0% to 30% acetonitrile (0.2% formic acid) at 600 nL/min flow rate on a nEasy-LC II system. Samples were analyzed with an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific). Full MS spectra were acquired with a 120,000 resolution, an automatic gain control of $5 \times 10^5$ and maximum injection time of 100 ms. Ions were selected for fragmentation, based on inclusion lists, using higher-energy collisional dissociation with 35% collision energy and an isolation window of 1.6 Th. The automatic gain control target was set to $2 \times 10^5$ with a maximum injection time of 500 ms and a resolution of 50,000.

**Computation of hydrophobicity index**

For additional validation, we assessed the correlation between retention time and hydrophobicity for TSAs (Supplementary Fig S3). The hydrophobicity values of the putative TSAs and other MAPs were calculated using an online available tool SSRCalc (https://bio.tools/ssrcalc). Only unmodified peptides were included, and predictions were based on the 100Å C18 column, 0.1% formic acid separation system. We correlated the observed mean retention time of a given peptide against the predicted hydrophobicity index, which corresponded to the percentage of acetonitrile at which the peptide elutes from the chromatographic system. For samples analyzed on our LC-MS/MS system, we computed a hydrophobicity index from a linear regression using retention time of our peptide quality control standard sample injected before endogenous and synthetic peptides.

**Tissue expression of sequences coding for aeTSA candidates**

RNA-seq data for 27 different tissues were downloaded from the Genotype-Tissue Expression (GTex) through dbGaP accession number phs000424.v7.p2 on April 16, 2018, and were used to assess the expression of the coding sequence of aeTSA candidates, as previously described (15). RNA-seq data from 50 donors were obtained for each tissue, except for the cervix (n = 6), fallopian tube (n = 7), bladder (n = 12), kidney (n = 38), and adipose tissue (n = 49; Supplementary Table S3). For each aeTSA candidate, the number of reads containing the MAP-coding sequence was normalized to reads per thousand million reads sequenced (rphm), log-transformed (log10(rphm + 1)), and averaged across all samples available for each tissue. aeTSA candidates exhibiting no peripheral expression at rphm >10 in tissues other than the MHC classes I and II tissues (brain cortex, nerve, and testis) were considered as genuine aeTSAs.

**Expression of TSA-coding regions**

The RNA expression of the TSA-coding region was quantified from mapped bam files using the "qCount" function of R package "QuasR" (29) with parameter "orientation = 'same'," as counts of all reads that overlapped with TSA-coding region on the TSA-coding strand. The counts were normalized to reads per hundred million reads mapped. In aeTSA expression analyses, we specifically analyzed the unique region to which individual aeTSAs were mapped because the context of the surrounding sequence may influence aeTSA expression.

**Clinical and genomic data from The Cancer Genome Atlas**

Processed and normalized level 3 data with hg38 for HM27 methylation, DNA copy-number variation (CNV), RNA-seq gene expression (FPKM normalized), as well as the clinical data, were downloaded from The Cancer Genome Atlas (TCGA) open-access database using the R package "TCGAbiolinks" (30). Matched RNA-seq and CNV (n = 376) or methylation data (n = 378) from TCGA-OV patients were used for Spearman correlation to assess associations between aeTSA RNA expression and CNV or DNA methylation changes. Case IDs of included samples were listed in Supplementary Table S4. To adjust for multiple tests, the Benjamini–Hochberg method was applied. Arm-level DNA copy-number alterations were downloaded from the Broad Institute TCGA Genome Data Analysis Center (doi:10.7908/C1P64B9Q). For each tumor, immune cell scores representing immune cell populations were estimated based on RNA-seq data as described by Danaher and colleagues (31).

**Frequency of aeTSA presentation**

To estimate the number of aeTSAs presented by individual patients, we performed a bioinformatic simulation based on two parameters. First, the likelihood of aeTSA expression was based on the proportion of tumors expressing the corresponding RNA in the TCGA-OV cohort. For aeTSA containing a germline SNP, the expression likelihood was calculated as follows: (proportion of TCGA tumors expressing aeTSA) × (SNP frequency in the given population). The SNP frequencies were obtained from the Genome Aggregation Database. Second, the HLA allele frequencies were retrieved from USA National Marrow Donor Program for the European–American population (n = 1,242,890), African American (n = 416,581), and Chinese (n = 99,672). Next, we simulated patients’ HLA genotype with six HLA class I alleles based on the reported frequencies in the given population. Because we assumed that the six HLA alleles were independent events, some HLA loci were homozygous in simulated patients. An aeTSA was considered to be presented in a simulated patient when both the aeTSA and the relevant HLA allele were expressed. Expression of each aeTSA was considered to be an independent event, except for the overlapping aeTSAs whose expression status was only simulated once for the same overlapping group. One million simulated patients and their aeTSA presentation status were generated for each of the three populations and used to plot the distribution.

**Statistical analyses and data visualization**

Analyses and figures were performed using the R v3.5.1 or Python v2.7.6. The “ggplot” package in R was used to generate heat maps of TSA-coding region expression in tumors. Correlation tests were done using the R function "cor.test" with the Spearman method unless otherwise indicated. Tests involving comparisons of distributions were performed with the one-way ANOVA test, and pairwise comparisons between groups were performed with the Wilcoxon rank sum test. Log-rank P values for survival analysis were calculated with the "survival" R package.

**Data and materials availability**

MS raw data and associated databases were deposited to the ProteomeXchange Consortium via the PRIDE (32) partner repository with the following data set identifiers: PXD014062 and 10.6019/PXD014062. RNA-seq data were deposited in the NCBI Sequence Read Archive and Gene Expression Omnibus (GEO) under accession code GSE131880.

**Results**

Proteogenomic analyses identify 103 TSAs in 23 HGScs

To get a systems-level characterization of the TSA landscape, we performed direct MAP identification with high-throughput MS/MS analyses (33–35). Current search engines rely on user-defined protein databases to match each acquired MS/MS spectrum to a peptide sequence (36). Hence, a peptide in a test sample can only be identified by the search engine when its sequence is included in the reference database. Because generic reference protein databases, such as...
UniProt, do not contain sample-specific mutations, out-of-frame translation events, and nonexonic sequences, our quest to capture TSAs coded by all genomic regions required construction of customized databases containing tumor-specific translation products for each tumor. Therefore, we used our previously described proteogenomic approach (15) to build a customized search database from RNA-seq reads of each sample analyzed. Such customized databases contain two modules: the canonical proteome (in-frame translation of exons) and the cancer-specific proteome, which is a 3-frame translation of cancer-specific RNA sequences after subtraction of normal RNA sequences (from TECs; Fig. 1). As in previous studies, we used TECs as a normal control for two reasons: (i) their key role in establishing immune tolerance during the development of immature T cells (i.e., central tolerance) and (ii) their remarkable ability to promiscuously express more transcripts than other types of somatic cells (37). MAPs from nine primary HGSC samples were obtained by immunoprecipitation of MHC I molecules, then analyzed by LC-MS/MS (15, 27). We also reanalyzed the immunopeptidomic data of an additional cohort of 14 HGSCs reported by Schuster and colleagues (13), by applying our proteogenomic approach to the samples for which matched RNA-seq and MS data were available.

TSAs that overlap genomic variants absent in dbSNP were unlikely to represent germline polymorphisms and were therefore labeled as mTSAs. Such classification yielded 12 mTSAs from the 23 samples analyzed (Supplementary Fig. S4; Supplementary Table S5). None of these mTSAs have been previously reported. Of the 12 mTSAs, three resulted from in-frame exonic translation, four from out-of-frame exonic translation, and eight from noncoding sequences. For three tumors, matched normal tissue was available, and RNA-seq analyses confirmed that mTSA variants were not germline polymorphisms (Supplementary Fig. S5). For tumors without matched normal tissues, we could not formally exclude that some mTSAs might have corresponded to rare polymorphisms absent in dbSNP. The possibility that we may have slightly overestimated the number of mTSAs would only reinforce our conclusion that, with less than one mTSA per tumor (12 mTSAs/23 tumors), mTSAs are so rare that most mTSAs are not CGAs.

For the unmutated TSA candidates, we applied stringent criteria to identify those that were genuine aetSAs, that is, whose expression was cancer specific. To this end, we analyzed RNA expression in 27 peripheral tissues across the human body. Due to their diverse genomic origins and association with alternative splicing events, we reasoned that quantification of peptide-coding sequence expression provided a uniform and more accurate measurement than quantification of whole gene expression. We thereby excluded from our aetSA list all candidates whose coding RNA were expressed (rpm >10) in any peripheral tissue other than MHC-I tissue (brain, nerve, and testis). Three aetSAs expressed in brain and nerve, whose application should be case dependent, as MHC-I have been shown to increase under inflammatory conditions (38). When the coding sequence of an aetSA candidate contained a germline single polymorphism (reported in dbSNP), this candidate was labeled as a valid aetSA only when the SNP-containing sequence and the reference sequence fulfilled the above criteria (Supplementary Fig. S6). Overall, 91 aetSA candidates satisfied our stringent criteria (Fig. 2; Supplementary Table S5). Five of 91 aetSAs were expressed in the testis, showing that some CGAs are aetSAs but that most aetSAs are not CGAs.

Each TSA was assigned a genomic location. When multiple locations were possible, the one with the highest occurrence of matching RNA reads was selected. Features of all TSAs are reported in Supplementary Table S5. It is formally possible that our stringent approach may have underestimated the total number of aetSAs resulting from atypical translation [5′untranslated region (UTR), 3′UTR, intergenic, frameshift]. Indeed, although we knew the reading frame used to generate MAPs in tumors when their coding RNA was expressed in some normal tissue, we could not infer which reading frame was translated. Such aetSAs candidates were therefore excluded in order to avoid inclusion of false positives in our TSA list.

Most HGSC TSAs are unmutated MAPs resulting from noncanonical translation

With an average of 2,200 unique MAPs identified per sample, we found a total of 103 unique TSAs (Fig. 3A; Supplementary Table S2). Each TSA was validated with its corresponding synthetic peptide (Supplementary Figs. S1 and S2). The retention characteristics of our TSAs were consistent with the distribution of MAPs and were correlated with their hydrophobicity indexes, supporting their correct identification (Supplementary Fig. S3). The number of TSAs identified...
per sample significantly correlated with the number of MAPs (Fig. 3B), and a correlation between the number of MAPs per HLA allele and tumor sample size was seen (Pearson correlation coefficient $r = 0.31$; Supplementary Fig. S7). This is consistent with the notion that tumor sample size is a limiting factor in MS analyses (27). In principle, TSAs that originate from mutated noncoding sequences could be designated as both mTSAs and aeTSAs. Arbitrarily, we decided to label them as mTSAs. The rationale was that irrespective of their genomic origin (exonic or not), mTSAs are expected to be "private TSAs," that is, not to be shared by a large proportion of tumors. In contrast, unmutated aeTSAs can theoretically be shared by a significant proportion of HGSCs. The features of TSAs identified in samples initially processed by us or by Schuster and colleagues (13) were similar (Fig. 3C). This suggested that our proteogenomic approach could be applied to RNA-seq and MS data, in general, and was not ostensibly affected by interlaboratory variability. In both cohorts, about 89% of TSAs were unmutated, and the majority of TSAs resulted from noncanonical translation, primarily from noncoding regions, and to a lesser extent from out-of-frame exonic translation (Fig. 3C). Two features of aeTSAs were observed: (i) 79% derived from noncoding sequences, in particular intronic (29%) and intergenic (22%) sequences and (ii) 91% were novel MAPs (Fig. 3D; Supplementary Table S5). Previously reported MAPs derived from in-frame exonic translation, except for...
Expression of aeTSA-coding transcripts in ovarian cancer samples

We then asked whether cancer-specific expression of aeTSA-coding transcripts resulted from random transcriptional noise or from recurrent transcriptional aberrations. To address this, we analyzed the RNA expression of genomic regions coding for our 91 aeTSAs in samples from this study and from the ovarian cancer cohort from TCGA. Regions coding for aeTSAs were expressed in a substantial proportion of ovarian cancers: 71 (78%) were expressed in at least 10% of samples and 16 (18%) were expressed in at least 80% of samples (Fig. 4; Supplementary Table S5). These commonly expressed regions have high potential to generate shared TSAs between patients. We conclude that expression of this set of 91 aeTSA-coding transcripts in HGSC is not a rare or random event but rather a common feature of HGSCs.

Genomic correlates of aeTSA expression

To understand the mechanisms of aeTSA expression, we took advantage of multiomic data from the data set TCGA-OV to explore the relationship between aeTSA RNA expression and local genetic or epigenetic aberrations. Correlations were tested between focal DNA copy-number changes, DNA methylation on the gene promoter regions, and the RNA expression for each aeTSA when applicable (Fig. 5A). When an aeTSA derived from a genomic region that was part of a gene (exon, intron, or UTR), we also analyzed the correlation between expression of the relevant gene and aeTSA expression. In the latter situation, we observed a conspicuous correlation between gene and aeTSA expression (Fig. 5A). This suggested that for aeTSAs whose coding region was in a gene, regulation of aeTSA expression generally affected the whole gene. Changes in DNA copy number showed a
positive correlation with RNA expression of aeTSAs. This was the case for both within-gene aeTSAs (left) and out-of-gene aeTSAs (antisense and intergenic, right). This suggested that DNA copy-number alterations had a substantial effect on aeTSA expression, especially for the within-gene aeTSAs expressed in a larger proportion of tumors (Supplementary Fig. S8A). When we examined the chromosomal distribution of aeTSA-coding regions, we found that several chromosome arms, frequently amplified in HGSC, yielded many aeTSAs (Fig. 5B). For example, the long arm of chromosome 3, which is commonly amplified in ovarian cancer (45), was the source of eight...
aeTSAs. As one of the top amplified regions, MECOM located at 3q26.2 (45) generated three overlapping exonic out-of-frame aeTSAs (Supplementary Table S5). However, amplification of chromosome arms was not always necessary (e.g., 15q) nor sufficient (e.g., 8q) to generate aeTSAs (Fig. 5B; Supplementary Fig. S8B).

Due to the technology used by TCGA for analysis of DNA methylation (HM27 arrays), methylation data were unavailable for the out-of-gene aeTSAs and the promoters of some aeTSA source genes. Our analysis of promoter methylation was therefore limited to a subset of 17 aeTSAs. Still, for six aeTSAs, we observed a significant correlation between DNA methylation and aeTSA expression (Fig. 5A). The correlation was negative in five cases and positive in one case. This is consistent with the notion that promoter demethylation frequently results with enhanced transcription. The two genes showing the highest negative correlation were MAGEC1 (\(r = -0.53, P_{\text{adj}} = 1.6 \times 10^{-3}\)) and MAGEA4 (\(r = -0.51, P_{\text{adj}} = 6.7 \times 10^{-3}\); dark blue bars, Fig. 5A). Genes of the MAGE family are CGAs that are over-expressed in several cancer types, including HGSC (3). Overall, we conclude that aeTSA expression is regulated, at least in part, at the transcriptional level by variations in gene copy number and DNA methylation.

The expression of three aeTSAs correlates with improved survival

Next, we sought to evaluate whether some aeTSAs might elicit spontaneous protective immune responses. Addressing this question is complicated by the fact that expression of aeTSAs at the peptide level requires, in addition to expression of aeTSA RNA, the presence of the relevant HLA allotype. We therefore allocated patients from TCGA into four subgroups based on the expression (or not) of individual aeTSA RNA and the presence (or not) of the relevant HLA allotype. Presentation of three aeTSAs correlated with a more favorable clinical outcome (Fig. 6A–C). The polymorphism of HLA alleles considerably reduced the size of each group, and therefore the statistical power of this analysis. Accordingly, the log-rank \(P\) value for the three aeTSAs ranged from 0.013 to 0.076 (Fig. 6A–C). Nonetheless, two observations suggested that these correlations were biologically meaningful. First, the “protective effect” of these aeTSAs appeared to be HLA restricted. In patients expressing aeTSA RNA, survival was superior when they also expressed the relevant HLA allele. Second, expression of the RTHQMNTFQR aeTSA and its relevant HLA allotype showed a positive correlation with tumor infiltration by T cells and cytotoxic T cells (Fig. 6D and E).

What is the median number of aeTSAs presented by individual tumors?

With the list of 91 aeTSAs, we estimated to what extent our study may benefit TSA-targeted immunotherapy. We randomly simulated the presentation status for 91 aeTSAs in one million patients. To estimate HLA allele frequencies, we used the three largest data sets from the USA National Marrow Donor Program: European Americans, African Americans, and Chinese (46). Six HLA alleles and...
aeTSA expression status were randomly generated independently using the allele frequencies in the given population, the expression proportion in TCGA-OV tumors, and the SNP frequencies when applicable. The number of aeTSAs per individual tumor was calculated as the sum of expressed HLA-aeTSA pairs. Based on these simulations, we found that at least one aeTSA could be found in 98% European Caucasians, 74% African Americans, and 77% Chinese (Fig. 7). The median number of aeTSAs per tumor was five in European Caucasians, two in African Americans, and four in Chinese. Differences among these populations resulted from variations in HLA allele frequencies and the fact that our tumor samples were mainly from European Caucasians. We suspect that our calculations underestimated the number of aeTSAs per tumor, mainly for three reasons. First, we did not take into account that more than 50% of MAPs bind two or more HLA allotypes, often across supertypes or even loci (47). Second, genomic regions that code for a given MAP frequently generate overlapping MAPs presented by different HLA allotypes (21). Third, for the five aeTSAs that include nonsynonymous SNPs listed in dbSNP, we assumed that only the SNP variant generating MAPs in our samples was valid and that the other SNP variant did not generate MAPs. We adopted this cautious strategy because changes in a single amino acid are sufficient to abrogate MAP presentation (48). We conclude that vaccines, including the current set of 91 aeTSAs, would cover practically all Caucasians with HGSC. Proteogenomic analyses of HGSC samples obtained from other groups should be performed to attain the same level of population coverage.

Discussion

The main conclusion from this study is that expanding the search space to all reading frames of all genomic regions has a transformative impact on the discovery of HGSC TSAs. Out of 103 TSAs reported herein, only three would have been picked up by strategies focusing solely on mutated exonic sequences. Limited numbers of mTSAs also resulted from out-of-frame exonic translation (n = 2) or from noncoding sequences (n = 7). In addition to their relative scarcity, mTSAs were unique to individual tumors. We therefore believe that, at least for HGSCs, mTSAs are not particularly attractive targets. In contrast, aeTSAs, which derived primarily from noncoding sequences, are considerably more attractive because they were more numerous (91 of our 103 TSAs) and were shared by a substantial proportion of HGSC tumors.

aeTSAs are not expressed in normal tissues, and their expression in HGSC was regulated at the transcriptional level by variations in somatic copy-number alterations and DNA methylation. In HGSC, the bulk of genetic changes is not somatic point mutations, but rather somatic copy-number alterations (49). These copy-number alterations drive cancer through defects in DNA repair, losses of tumor suppressors (such as TP53), and amplification of oncogenes (45, 49). Because somatic DNA copy-number alterations are common in various cancer types, it will be interesting to explore the impact of DNA copy number on aeTSA expression in other cancers. DNA hypomethylation also contributed to the aberrant expression of six aeTSAs, including two members of the MAGE family that are silenced by methylation in normal cells (3). However, because most aeTSAs derive from allegedly noncoding regions, it will be necessary to combine whole-genome methylation analyses with proteogenomic studies of HGSCs in order to evaluate more comprehensively the relation between DNA methylation and aeTSA expression. Other regulatory mechanisms could also be relevant to aeTSA biogenesis. In particular, recurrent alternative splicing events have been identified in various cancers (50) and could be instrumental to the appearance of intronic aeTSAs. Finally, the potential impact of trans factors, such as miRNAs and transcription factors, on aeTSAs has yet to be explored. Our observation that aeTSA expression was a common feature of HGSCs should provide impetus to perform such analyses.

Perhaps the next most critical question to address in future studies is the immunogenicity of individual aeTSAs. Our bioinformatic analyses of TCGA-OV provided hints that some aeTSAs might elicit spontaneous protective immune responses. However, the question of immunogenicity has to be addressed experimentally by studying responses of blood and tumor-infiltrating lymphocytes to aeTSAs. Testing the immunogenicity of 91 antigens is far from being trivial but is certainly doable using state-of-the-art high-throughput pipelines (51). We speculate that a high proportion of our aeTSAs will prove to be immunogenic. Indeed, when we previously performed comprehensive (ex vivo and in vivo) immunogenicity analyses of aeTSAs discovered in mice with the method used in the present article, we found that they were all immunogenic (15). In contrast to what is seen with “predicted MAPs” (i.e., MAPs not validated by MS analyses), a study has revealed
Zhao et al.

that more than 80% of MS-validated mouse MAPs are immunogenic (52). Finally, our bioinformatic predictions suggest that the group of aeTSAs reported herein would be sufficient to cover HGSC in almost all Caucasians. We intend to extend our proteogenomic analyses to HGSC samples obtained from other populations with different HLA allele frequencies. Many epitopes can be included in a vaccine. Hence, for HGSC treatment, it might be possible to generate aeTSA-based off-the-shelf vaccines. In order to increase the level of “precision,” different vaccines could be created that contain aeTSAs presented by one or a few HLA allotypes (e.g., HLA-A*02:01 aeTSAs).

Disclosure of Potential Conflicts of Interest

Q. Zhao and C. Perreault have ownership interest in a patent application filed by Université de Montréal. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: Q. Zhao, P. Thibault, C. Perreault
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): Q. Zhao, J. Lanoix, C. Durette, C. Côté, P. Gendron, K. Vincent, D.G. Millar, P.S. Ohashi

References


Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer

Qingchuan Zhao, Jean-Philippe Laverdure, Joël Lanoix, et al.


Updated version

Access the most recent version of this article at:
doi:10.1158/2326-6066.CIR-19-0541

Supplementary Material

Access the most recent supplemental material at:
http://cancerimmunolres.aacrjournals.org/content/suppl/2020/02/11/2326-6066.CIR-19-0541.DC1
http://cancerimmunolres.aacrjournals.org/content/suppl/2020/03/14/2326-6066.CIR-19-0541.DC2

Cited articles

This article cites 51 articles, 12 of which you can access for free at:
http://cancerimmunolres.aacrjournals.org/content/8/4/544.full#ref-list-1

E-mail alerts

Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions

To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions

To request permission to re-use all or part of this article, use this link http://cancerimmunolres.aacrjournals.org/content/8/4/544.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.