JCI insight

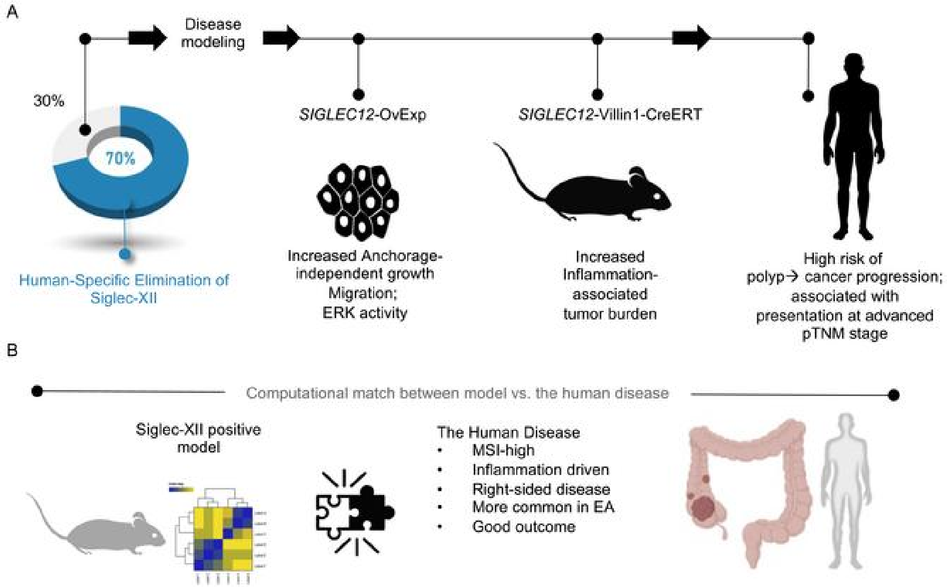
Human-specific elimination of epithelial Siglec-XII suppresses the risk of inflammation driven colorectal cancers

Hector A. Cuello, … , Ajit Varki, Pradipta Ghosh

JCI Insight. 2024. https://doi.org/10.1172/jci.insight.181539.

Research In-Press Preview Inflammation Oncology

Graphical abstract


A
Disease modeling
30%
70%
Human-Specific Elimination of Siglec-XII
SIGLEC12-OvExp
Increased Anchorage-independent growth Migration; ERK activity
SIGLEC12-Villin1-CreERT
Increased Inflammation-associated tumor burden
High risk of polyp→ cancer progression; associated with presentation at advanced pTNM stage
B
Computational match between model vs. the human disease
Siglec-XII positive model
The Human Disease
• MSI-high
• Inflammation driven
• Right-sided disease
• More common in EA
• Good outcome

Human-Specific Elimination of Epithelial Siglec-XII Suppresses the Risk of Inflammation Driven Colorectal Cancers

AUTHORS: Cuello Hector A[1, 2], Sinha Saptarshi[1, 3, 4], Verhagen Andrea L[1, 2], Varki Nissi[2, 5], Varki Ajit[1-3, 6*], Ghosh Pradipta[1, 3, 4, 7*].

AFFILIATIONS:

[1]Department of Cellular and Molecular Medicine, University of California San Diego, CA, USA;

[2]Glycobiology Research and Training Center, University of California San Diego, CA, USA;

[3]Department of Medicine, University of California San Diego, CA, USA;

[4]Moore's Comprehensive Cancer Center, University of California San Diego, CA, USA;

[5]Department of Pathology, University of California San Diego, CA, USA;

[6]Center for Academic Research and Training in Anthropogeny, University of California San Diego, CA, USA;

[7]HUMANOID Center of Research Excellence (CoRE), University of California San Diego, CA, USA;

Short Title: Siglec-XII expression drives inflammation-associated CRCs.

* **CORRESPONDING AUTHOR CONTACT INFORMATION**

Pradipta Ghosh, M.D.; Professor, Departments of Medicine and Cell and Molecular Medicine, University of California San Diego; 9500 Gilman Drive (MC 0651), George E. Palade Bldg, Rm 232; La Jolla, CA 92093. **Phone**: 858-822-7633. **Email:** prghosh@ucsd.edu

Ajit Varki; Professor, Department of Cell and Molecular Medicine, University of California San Diego; 9500 Gilman Drive, Biomedical Research Facility II; La Jolla, CA 92093. **Phone**: 858-534-2214. **Email:** a1varki@health.ucsd.edu.

STATEMENT ON CONFLICT OF INTERESTS

The authors declare no competing interest with the content of this manuscript.

ABSTRACT (200 WORDS)

Carcinomas are common in humans but rare among closely related “great apes”. Plausible explanations, including human-specific genomic alterations affecting the biology of sialic acids are proposed, but causality remains unproven. Here, an integrated evolutionary genetics-phenome-transcriptome approach studied the role of *SIGLEC12* gene (encodes Siglec-XII) on epithelial transformation and cancer. Exogenous expression of the protein in cell lines and genetically engineered mice recapitulated ~30% of the human population in whom the protein is expressed in a form that cannot bind ligand due to a fixed, homozygous, human-universal missense mutation. Siglec-XII null cells/mice recapitulated the remaining ~70% of the human population in whom an additional polymorphic frameshift mutation eliminates the entire protein. Siglec-XII expression drove several pro-oncogenic phenotypes in cell lines, and increased tumor burden in mice challenged with chemical carcinogen and inflammation. Transcriptomic studies yielded a 29-gene signature of Siglec-XII-positive disease and when used as a computational tool for navigating human datasets, pinpointed with surprising precision that *SIGLEC12* expression (model) recapitulates a very specific type of colorectal carcinomas (disease) that is associated with mismatch-repair defects and inflammation, disproportionately affects European-Americans, and carries a better prognosis. They revealed a hitherto unknown evolutionary genetic mechanism for an ethnic/environmental predisposition of carcinogenesis.

MAIN TEXT (6487 WORDS)

INTRODUCTION (403 WORDS)

Colorectal cancers (CRCs) are the third most commonly diagnosed cancers and the second leading cause of cancer-related deaths globally, with an alarmingly rising incidence (1). Such high prevalence and rising incidence among humans is particularly surprising, given that CRCs among other many carcinomas are a rarity among captive chimpanzees with whom we share >99% protein sequence homology (2–4). In fact, cancers are part of a list of common human diseases that may be partially or completely unique to our species compared to other primates (5–7).

Human-specific changes in specific Siglecs is one of the reasons put forth as molecular mechanisms that could explain human proneness to developing cancers (8, 9). Siglecs are a group of vertebrate lectins belonging to the immunoglobulin superfamily that recognize glycan bearing sialic acid (Sias). A subset of inhibitory CD33-related Siglecs (CD33rSiglecs) are prominently expressed in immune cells and are considered to have a regulatory role in suppressing the activation of innate immune cells via cytosolic immunoreceptor tyrosine-based inhibitory motifs (ITIMs) (10, 11). These ITIMs recruit protein phosphatases such as Src homology region 2-containing protein tyrosine phosphatases (PTPs) SHP-1 and SHP-2 (12, 13). Of relevance in the context of CRCs, both SHP1 (14) and SHP2 (15) have been found to serve as brakes that limit tumorigenesis via their ability to antagonistically inactivate pro-oncogenic tyrosine-based signals; mice lacking SHP1/2 in intestinal epithelial cells (IECs) develop higher tumor burden, associated with sustained activation of downstream pathways such as the PI3K/Akt, Wnt/β-catenin, NFκB and STAT3 signals. Thus, signaling via a functional Siglec→SHP1/2 axis in IECs is expected to inhibit tumorigenesis. Although multiple CD33-related Siglecs are reported to be upregulated in cancers (16, 17), no study has evaluated the role of Siglecs in IECs.

Here we explore a previously unforeseen human-specific association between CRCs and Siglec-XII, a member of the Siglec family of Sia-recognizing receptors that is primarily expressed in epithelial cells (18) and functionally inactivated from recognizing Sia ligands, thereby signaling aberrantly only in humans. Since mice do not have a *SIGLEC12* gene, we modeled the human disease *in vitro* and in mice and then used an unbiased computational approach to navigate human disease samples to unravel the implications of Siglec-XII expression and its impact on oncogenesis. Findings surprisingly reveal that expression of a functionally defective Siglec-XII in a subset of humans predisposes them to develop a specific type of CRCs that is environmentally influenced (higher inflammation) and ethnically predisposed.

RESULTS (2209 WORDS)

A study design rationalized by human-specific evolutionary genetics of the *SIGLEC12* gene.

To study the role of Siglec-XII in IECs and ensure that findings are relevant to the human disease (i.e., CRCs), we drew inspiration from the known uniquely human features of *SIGLEC12,* the gene that encodes Siglec-XII in humans. This gene harbors a human universal missense mutation in the Sia-binding domain that affects a critical arginine residue (the Arg/Cys mutation), resulting in an inability to recognize Sias (hence the protein is denoted using Roman numeral XII to differentiate it from functional Siglecs) (19) (**Figure 1A**). This inactivating mutation occurred prior to the common ancestor of all modern humans thus is absent in closely related "great apes", the latter express a functional Siglec-12 that preferentially recognizing Neu5Gc (a Sia lost from the human lineage) (18, 19). Furthermore, the human *SIGLEC12* locus is currently undergoing negative selection in humans that favors a null and/or truncated form of the protein, characterized by the excess of rare alleles and the presence of “selective sweep” acting on the gene throughout the overall human population (20). The most common frameshift mutation arises from the insertion of a guanine (G) in the *SIGLEC12* gene, resulting in the loss of expression of the complete protein in most humans (**Figure 1A**) (18, 21). Interestingly, among the minority of humans who possess the genomic ability to express it, the protein is detected in certain tissue macrophages, but is not found in other blood cell types and instead exhibits higher levels of expression on the surfaces of epithelial cells (18). Another peculiarity of this protein is that, even though Siglec-XII does not have the ability to recognize Sias, it still possesses ITIM and ITIM-like domains in its cytosolic tail that can undergo phosphorylation to recruit Shp1 and Shp2 phosphatases (22), raising the possibility of Siglec-XII serving as a dominant negative protein that can signal (via Shp) in the absence of binding the natural ligand (Sias).

We hypothesized that aberrant signaling via Siglec-XII supports human-specific mechanisms increasing cancer risk and progression. Because the ligand-binding defective mutant is expressed in only ~30% of the healthy human population, but enriched up to ~70-80% in all carcinomas and ~64% in CRCs (8) (**Figure 1B**), we hypothesized that the minority of humans who express full-length Siglec-XII may be at the highest risk for developing advanced carcinomas. The enrichment of Siglec-XII positivity from normal to cancer tissues suggested that Siglec-XII positivity, either alone or via its interaction with environmental factors predispose to cancers. To model these uniquely-human features of the *SIGLEC12* gene, we exogenously expressed the ligand-binding defective Siglec-XII in cell lines and in mice (which do not have a *SIGLEC12* gene (13)) and explored cellular phenotypes and tumorigenesis, respectively.

Exogenous expression of Siglec-XII in null human carcinoma cell lines enhances malignant features.

Previously, a flow cytometry screen of five human carcinoma cell lines showed that MDA-PCa-2b and LNCaP (prostate cancer) and MCF-7 (breast cancer) lines express Siglec-XII, whereas MDA-MB-231 (breast cancer) and PC-3 (prostate cancer) do not express it (18). In a second screen of another set of 4 colorectal carcinoma cell lines (Colo-320, Caco-2, LS-180 and HT-29), also by flow cytometry, we confirmed that 2 of 4 express the Siglec-XII protein (LS-180 and HT-29) (**Supplemental Figure S1A**). We confirmed this by checking for and confirming the absence of the polymorphic frameshift insertion mutation in the human *SIGLEC12* gene (**Supplemental Figure S1B**), the event which results in a premature stop codon and consequent loss of expression in most individuals. The higher frequency of expression in carcinoma cell lines (5/9) is in keeping with the prior conclusion (8) that human carcinomas have a higher incidence of Siglec-XII expression than expected in the general population.

To begin to explore the significance of Siglec-XII in the progression of CRCs, we used Siglec-XII non-expressing Caco-2 cell line. As positive control, and to provide continuity with our prior work (8, 18), we used the Siglec-XII non-expressing PC-3 human prostate carcinoma cells. These cells were previously characterized by flow cytometry to confirm cell surface expression of Siglec-XII exclusively after transfection. Both cell lines were transfected with either pCDNA-3.1-*SIGLEC12* or with empty vector (control), forcing them to exogenously express full-length Siglec-XII. Stable clones were selected (see methods) and used in a variety of assays, i.e., cell adhesion, spheroid formation and migration. Compared to controls, the Siglec-XII-expressing counterparts significantly and consistently showed decreased cell adhesion (**Figure 2A-B**), accelerated spheroid growth (**Figure 2C-E**) and increased Transwell® migration (**Figure 2F-G**). These phenotypes were accompanied also by increased ERK1/2 activity, as determined by immunoblotting for the phosphorylated kinase (**Figure 2H-K**). Siglec-XII expression was confirmed by immunoblotting (**Figure 2H-K**).

Creation and validation of a transgenic intestine-specific knock-in *SIGLEC12* murine model.

Because carcinogenesis in the colon requires the complex interplay between multiple factors (host genetics, gut microbes, and the immune system) that are hard to recapitulate in vitro in cell line models, mouse models have proven crucial in the identification of the role of genes responsible for CRC initiation and progression (23). Given that mice do not have any endogenous *SIGLEC12* gene (13), we developed a mouse model that allows conditional expression of the protein Siglec-XII (see legend, **Figure 3A**). The *SIGLEC12* knock-in mice presents the *egfp* gene (including stop codon) flanked between two loxP sites, and it is upstream of the Siglec-XII coding sequence. This stop codon prevents *SIGLEC12* gene expression in the absence of Cre recombinase. To selectively knock-in *SIGLEC12* in the intestine, we bred *SIGLEC12* transgenic mice with

Villin1-Cre-ERT mice; the latter restricts the Cre-ERT expression to the villi and crypts of the small and large intestine (24, 25). We collected small and large intestines at early and late time points after five consecutive days of tamoxifen administration (**Figure 3B**) and analyzed them for Siglec-XII expression by immunohistochemistry and immunoblotting. Both methodologies confirmed that Siglec-XII is expressed prominently and homogeneously exclusively in the transgenic mice (*SIGLEC12*-Villin1-Cre-ERT) but not their control littermates (Villin1-Cre-ERT) as early as day 12 (**Figure 3C-D**). We also confirmed that such conditional expression of Siglec-XII in the transgenic mice was sustained as late as day 87 (**Supplemental Figure S2**). Histopathological analysis ruled out any gross or microscopic changes on day 87 in various organs (colon, liver, kidney, lung, and spleen) (**Supplemental Figure S3**). More importantly, we did not observe any features suggestive of inflammation, metaplasia, dysplasia, or neoplasia.

***SIGLEC12*-knock-in mice display greater tumor burden in response to chemical carcinogenesis.**

Because the transgenic *SIGLEC12* mice do not develop spontaneous tumors, next we sought to use it in conjunction with chemically induced CRC models which recapitulate the progression from aberrant crypt foci and adenoma to adenocarcinoma and are commonly used to study the effects of diet, inflammation, and gut microbiota (23). More specifically, we subjected mice to chemical carcinogenesis using well established use of azoxymethane (AOM) and dextran sulfate sodium salt (DSS) (26–28) (**Figure 4A**; **Supplemental Figure S4**). While AOM mainly leads to the generation of adenomas, exposure to AOM/DSS is known to induce the formation of a complete process of colon oncogenesis, progressing from the initial proliferation of crypts to the final development of high-grade dysplasia and adenocarcinomas in ~25-50% of the C57BL/6 mice (29). Because many Siglecs are inhibitory receptors expressed in innate immune cells that regulate inflammation (30), the AOM/DSS model seemed furthermore appropriate as it is known to primarily recapitulate inflammation driven CRCs (31). The animals were followed for 87 days and examined for colorectal tumors at necropsy. Examination of the colons showed that Siglec-XII-expressing mice presented significantly increased tumor burden than controls (**Figure 4B-C**), and the base of the tumors were typically associated with immune cell infiltrates (arrow, compare **Figure 4D-E**). Animals with induced Siglec-XII and exposure to AOM/DSS also showed larger rectal tumors compared to control animals (**Figure 4D-E**).

Gene signatures uniquely induced due to Siglec-XII expression in human CRCs.

To ascertain which processes and/or drivers of human carcinogenesis are recapitulated in our chemically induced transgenic mice, we sought an unbiased computationally driven 2-step approach. First, we carried out RNA sequencing of the colons at baseline and after AOM/DSS challenge. A differential expression analysis (DEA) of genes between AOM/DSS treated controls (Villin1-Cre-ERT) and Siglec-XII-expressing mice led to the identification of a 29-gene signature (**Figure 5A-B**), which is upregulated in Siglec-XII mice.

This set of 29 genes was enriched for diverse bioenergetic processes (**Figure 5C**). As expected, in the absence of ligand recognition capabilities, the tyrosine-based signaling pathways, typically modulated antagonistically by Shp1/2 phosphatases, were lacking. These DEGs were not differentially expressed at early timepoints (baseline; **Figure 5D**) when Siglec-XII expression is strong (**Figure 3C-D**), indicating that Siglec-XII expression alone was insufficient. Instead, the gene signature captures the combinatorial effect of AOM/DSS and Siglec-XII. In fact, no significant DEGs were found between baseline samples. The DEGs were upregulated also in Siglec-XII expressing Caco-2 cells (**Figure 5E**).

Next, we used the gene set as a signature of CRC predisposition to navigate diverse CRC datasets. Because chemical induction models recapitulate some of the earliest steps for CRC initiation and progression (23), such as aberrant crypt foci, dysplasia, etc., we asked if the gene signature is differentially induced in different parts of the human colon and diverse subtypes of polyps that carry differential risk of progression to CRCs. We found that the 29-gene signature uniquely induced due to Siglec-XII was induced also in the right side of the human colon (compared to the left; see **Figure 5F**-*left*) regardless of whether these samples were from normal subjects (Control; **Figure 5F**-*left*) or from patients who had polyps (Adj. normal; **Figure 5F**-*left*). The signature was significantly induced in polyps that are known to carry risk for CRC progression (adenomatous and sessile serrated adenomas; SSAs; **Figure 5F**-*left*) but not in benign hyperplastic polyps. Induction of the signature in polyps was confirmed also in an independent cohort (**Figure 5F**-*right*).

We asked if Siglec-XII expression is associated with higher risk of polyp→CRC progression. To this end, we leveraged a publicly available dataset that represents a time-lapse model for CRC initiation and progression in humans (32) (**Figure 5G**-*left*). In that model, cancer adjacent polyps (CAPs) were used as a model to study cancer progression temporally because the precursor polyp of origin remains in direct contiguity with its related (33–35). Cancer-Free Polyp (CFP) cases, on the other hand, are polyps that have remained cancer free, despite sharing similar size, histologic features, and degrees of dysplasia as CAPs (**Figure 5G**-*left***).** Laser-dissected pre-neoplastic tissues from the CAPs represent polyps with a proven high risk of CRCs, CFPs represent polyps at low risk, and paired normal colons sampled ~8 cm away from the polyps served as controls. Our 29-gene signature is significantly induced in CAPs compared to CFPs and could classify them perfectly (ROC AUC 1.00; **Figure 5G**-*right*), indicating that Siglec-XII expression shares similar patterns of induction of gene expression that are encountered in polyps that are at highest risk for progression to CRCs.

Consistent with the fact that cancers that originate from right-sided polyps are often diagnosed at advanced stages (36), a CRC array and multivariate analyses showed that Siglec-XII positivity was significantly associated with presentation at advanced stages (pTNM; **Figure 6A-B**; **Supplemental Table 1**).

Siglec-XII expression is associated with a specific ethnic subtype of CRCs.

Next, we used the 29-gene model-derived signature as a computational tool to navigate human CRC datasets and objectively assess for a precise match in gene expression patterns in Siglec-XII model vs human CRC subtypes (**Figure 6C**). We found that the 29-gene signature was significantly induced in both tumors and adjacent normal tissues from self-identified European Americans [a.k.a "whites" and described in the dataset as "Caucasian Americans"] vs "African Americans" [a.k.a. blacks"] (**Figure 6D**; ROC AUC for each 1.00). This dataset was first used in a study (37) that showed differential contributions of immune cells and inflammation and mismatch repair defects among two ethnic groups; it is one of the studies that established what is now widely recognized as a key ethnic difference in the CRC subtypes (38). Furthermore, consistent with the fact that tumors in European Americans are more often right-sided with microsatellite instability (MSI) and carry an overall good prognosis (38), we found that high expression of the 29-gene signature was associated with a favorable outcome among all CRCs; both overall (**Figure 6E**) and progression-free (**Supplementary Figure S5A**) survival were prolonged. This favorable impact on outcome continued to hold true even when the analysis was repeated among just the MSI-high tumors (**Figure 6F**; **Supplementary Figure S5B**).

Having observed a match in the model-derived 29-gene expression pattern in the CRC subtype to which Caucasians are predisposed, we asked if the converse holds true, i.e., if the key disease-drivers reported in the European Americans tumors as key clinicopathological disease features are also recapitulated in our model. The study reported that European Americans, but not African Americans develop tumors that are characterized by inflammation (high *IL1B, IL8, NFKBIE* and *IL6ST*) and microsatellite instability (MSI-high) in the setting of altered expression of several key genes in the mismatch repair pathway (37). We found both these patterns to hold true in our mouse model (**Figure 6D**; **Supplementary Figure S5C-H**). These findings show that the Siglec-XII model faithfully recapitulates the pathological drivers believed to be frequently seen in one ethnicity (European Americans) but not the other (African Americans).

DISCUSSION (1120 WORDS)

The major discoveries we report here provide insights into the consequences of expression of the epithelial Sia-binding defective Siglec-XII in ~30% of the human population, and how that may put them at risk of developing inflammation-driven CRCs. We show that in model systems that recapitulate most individuals who lack expression of the Siglec-XII receptor versus those who do, the expression of the receptor that is unable to bind its natural ligand has 3 key effects (see summary of findings; **Figure 7**): i) cancer-associated cellular phenotypes are enhanced; (ii) tumor burden is increased in mice and is associated with advanced stages of disease at diagnosis and (iii) gene expression patterns changed in ways that mirrored with surprising degree of precision an inflammation-environmentally driven carcinogenesis process. Because phenotypic changes in CRC-derived Caco-2 cells generally held true in prostate cancer-derived PC-3 cells, aberrant functionally defective Siglec-XII expression in other epithelial linings may serve as a shared contributor to and/or predisposition for other inflammation-driven human carcinomas. We conclude that the persistence of Siglec-XII in humans predisposes to CRCs and likely other carcinomas, and its elimination could serve as a selection favoring survival.

It is noteworthy that besides *SIGLEC12*, there are other Siglecs that have undergone human-specific changes in functional gene status, expression, or ligand binding, which include: *SIGLEC1*, *SIGLEC5/14*, *SIGLEC6*, *SIGLEC7*, *SIGLEC9*, *SIGLEC11*, *SIGLEC13*, and *SIGLEC16* (39). As with Siglec-XII, only a minority of the human population (38.7 %) has a SIGLEC16 allele coding for functional protein expression, whereas the majority carries an inactive pseudogene, *SIGLEC16P*, product of a four-nucleotide deletion disrupting the open reading frame. Although in the vast majority of these cases we don't know how the human-specific changes impact oncogenesis, a positive association with survival in glioblastoma was found for the intact *SIGLEC16 (activating Siglec)*-positive cases (9). It is possible that activating and inactivating Siglecs, when aberrantly expressed in the human population as functional or non-functional variants, could alter the risk of initiation and/or progression of oncogenesis in diverse organs. What we established with certainly is that in the case of Siglec-XII, its expression did not cause spontaneous carcinogenesis, but predisposed to environmentally induced carcinogenesis. Its expression in established tumors, however, was associated with improved outcome.

The specific mechanism of action of the Sia binding-defective Siglec-XII in those who express it is unknown. However, taking into consideration prior reports from us (8, 18) and others (22), it is possible to highlight a relation between this cell receptor and the risk to develop carcinomas. For example, cancer-related signaling pathways were enriched in PC-3 prostate cancer cells transfected with *SIGLEC12* (18) which was accompanied by enhanced tumor growth as xenografts in nude mice (8, 18). Similar efforts to transfect the chimpanzee version of *SIGLEC12* or the arginine-restored version of human *SIGLEC12* were not successful

(18), suggesting that additional components that may be critical for protein folding and targeting were also lost during evolution. The fact that still recruit PTPs trough phosphorylated ITIM and ITIM-like domains in its cytosolic tail (22) suggests that it retains the ability to transmit downstream signals; however, whether it does so constitutively or upon binding to hitherto unidentified ligands remains unknown and cannot be dismissed. We show that expression of human Siglec-XII increased at least one type of signaling pathway *in vitro* in the epithelial cells (ERK/MAPK) and inflammatory cytokine signals in vivo in the murine tumors (IL6, IL8, IL1β), demonstrating that despite defect in binding to its natural ligand, Siglec-XII may support some form of gain in signaling function that were associated with also gain in pro-oncogenic phenotypes. It is possible that expression of the inhibitory Siglec-XII served as a dominant negative receptor that sequestered tumor suppressive SHP1/2 phosphatases (14, 15), thereby contributing to the oncogenic risk. Although we did not observe aberrant co-expression of either SHP1 or SHP2, we noted the upregulation of another member in the PTP family, i.e., *Ptpn18*; upregulation of *Ptpn18* has been reported in yet another type of CRC, i.e., early onset CRC (EOCRC) and such upregulation carries worse prognosis (40).

Perhaps the most important finding of translational relevance is the degree of precision with which Siglec-XII positive tumors (our model) recapitulated the gene expression patterns encountered in normal human colonic mucosa and in diverse human polyps and CRCs (the disease). Expression of *SIGLEC12* captured the gene expression pattern that is detected at higher levels on the right side of the colonic mucosa compared to the left. Because the 29 gene signature largely reflected mitochondrial bioenergetic processes, we suspect that this difference is due to previously demonstrated striking differences in mitochondrial bioenergetics between the right vs left colonic mucosa (41). In fact, the bioenergetic status of the right colon has been shown to mimic that seen in that of the normal tissue adjacent to CRCs (41). We also found that elevated expression of the 29 gene signature is encountered in polyps that carry a higher risk of progression to CRCs. It also mirrored a distinct subtype of CRCs that are more often encountered in European Americans; these are right-sided, primarily driven by mismatch repair defects and IL1β/IL8/IL6-centric inflammation and are associated with improved outcome. Consistent with the form of disease in humans, we saw immune cell infiltrates in our mouse model. Given their active immune microenvironment and elevated expression of various checkpoint molecules, MSI-H, right-sided CRCs in “whites” present as promising candidates for targeted immunotherapy with immune checkpoint inhibitors (42, 43). It is possible that either Siglec-XII or the 29-gene signature could serve as a biomarker for both prognostication and prediction of therapeutic response to immunotherapy. On the therapeutic side, Siglec-XII is a promising candidate for targeted drug delivery to cancer cells expressing it due to its limited and specific expression in only a few cell types. For example, given its ability to internalize upon antibody binding (18), coupling a toxin to the antibody presents a potentially effective strategy for advancing cancer therapy. The simple assay we developed to rapidly screen for all mutations abrogating expression using patient-derived saliva and urine samples could help identify those who may benefit (8).

Despite the insights gained, there are a few limitations of this study. The use of a handful of CRC cell lines is one; analyzing a broader range of CRC cell lines is expected to yield how this Siglec-XII phenomenon intersects with other CRC-driving genetics. Additional validation studies are also needed for dissecting the signaling pathways in the animal model; such studies are expected to establish a clearer link between Siglec-XII and its role in cancer.

Taken together these data support the notion that Siglec-XII expression may facilitate CRC progression in humans. Similar studies need to be done with other carcinomas that also have very high incidence in humans compared with closely related apes.

METHODS (2755 WORDS)

Sex as a biological variant. In this study, we evaluated the impact of Siglec-XII expression in mice. Although only female mice were used, sex was not considered a biological variable, based on multivariate analyses of a Siglec-XII positive human cohort (**Figure 6B**).

Experimental Methodologies:

Cell lines. Prostate (PC-3) and colorectal (Caco-2) adenocarcinoma cell lines were purchased from the American Type Culture Collection (ATCC, Virginia, USA). PC-3 cells were grown in F-12K Medium (Kaighn's Modification of Ham's F-12 Medium) supplemented with 10% fetal bovine serum (FBS) (Gibco, New York, USA). Caco-2 cells were grown in Eagle's Minimum Essential Medium (EMEM) supplemented with 20% fetal bovine serum (FBS). Monolayers were routinely sub-cultured with Trypsin-EDTA solution (Gibco, New York, USA), following standard procedures. Cell cultures were maintained at 37°C in a humidified atmosphere of 5% CO_2 and tested for contamination with Mycoplasma with the kit DAPI (Vector, California, USA). The cell lines used for the described experiments had all been maintained in tissue culture for less than 20 passages.

Establishing Cell lines Stably Expressing Siglec-XII. PC-3 and Caco-2 cells were transfected with PvuI linearized h*SIGLEC12*-pcDNA3.1 or empty pcDNA3.1(-) in six-well plates using Lipofectamine 2000 (Invitrogen, California, USA). 48 h after transfection, the cells were trypsinized and grown with 800 µg/ml G418. After 1 month in culture, expression of Siglec-XII was determined using Western Blot. Cell adhesion, spontaneous spheroid formation, cell viability, cell migration and western blot studies were conducted using stable vector-transfected PC-3 and Caco-2 cells (h*SIGLEC12*-pcDNA3.1 or empty vector as negative control).

Flow cytometry. Cell lines were collected by enzyme-free cell dissociation buffer (Thermo Fisher Scientific, California, USA) and incubated with mouse anti Siglec-XII 276, anti Siglec-XII rabbit polyclonal antibody (AP18196PU-N, Origene, Maryland, USA), mouse IgG Isotype Control (MG1-45, BioLegend, California, USA) or rabbit IgG Isotype Control (X0936, Dako, Denmark) on ice for 30 min. Cells were washed with phosphate buffered saline (PBS) and incubated with anti-mouse Alexa-Fluor® 488 (A11001, Invitrogen, California, USA) or anti-rabbit Alexa-Fluor® 488 (A11053, Invitrogen, California, USA) on ice for 30 min. Acquisition of data was carried out using a FACSCalibur flow cytometer (Becton Dickinson, New Jersey, USA) and data analyzed using FlowJo® software.

***SIGLEC12* frameshift mutation**. The genomic DNA was isolated from cell lines. The frameshift deletion mutation of *SIGLEC12* was analyzed using PCR. The primers used to amplify the *SIGLEC12* locus were 5′-ACCCCTGCTCTGTGGGAGAGT-3′ (forward) and 5′AGGATCAGGAGGGGCATCCAAGGTGC-3′ (reverse). The PCR was performed using the Phusion High-Fidelity Polymerase kit (Thermo Scientific, California, USA). The amplified product was purified using the QIAquick PCR purification kit (Qiagen, Venlo, Netherlands) and sent for sequencing to EtonBio (San Diego, USA). The sequencing was performed using the primer: 5′-CTCTCTCTGGTGTCTCTGATGC-3′ (reverse).

Cell adhesion assay. Cell adhesion was measured using crystal violet staining. Cells were harvested with an enzyme-free cell dissociation buffer and seeded at a concentration of 4 × 10^4 cells/well in complete medium in a 96-well plate. After incubation at 37°C at 0.5, 1 and 1.5 h the cells were washed with PBS, and non-adherent cells were removed by aspiration. Adherent cells were stained with a 0.5% (w/v) crystal violet solution with 20% (v/v) methanol. After washes, the dye was solubilized with 10% (v/v) methanol and 5% (v/v) acetic acid, and the absorbance was measured at 595 nm by EnSpire® Multimode Plate Reader (PerkinElmer, Massachusetts, USA).

Spontaneous spheroid formation. Cells were harvested and passed through a 40 µM cell strainer. Cells were plated at a density of 3000 cells in 100 µL of growth media per well using 96-well spheroid microplates. Spheroid cultures were photographed, and cell viability was measured at day 4, 8 and 10. The same seeding methods were used for all cell lines.

Cell viability assay. The CellTiter-Glo® 3D Cell Viability Assay protocol was followed (Promega, Wisconsin, USA). The CellTiter-Glo® 3D Cell Viability Assay is a homogeneous method to determine the number of viable cells in 3D cell culture based on quantitation of the ATP present, which is a marker for the presence of metabolically active cells. Briefly, spheroids were transferred to white plates and CellTiter-Glo® 3D reagent was added directly into wells in a 1:1 dilution. The solutions were well mixed by shaking the plate for 5 minutes then incubated at room temperature for 25 minutes. After incubation, the generated luminescent signal was read and analyzed using the EnSpire® Multimode Plate Reader (PerkinElmer, Massachusetts, USA).

Cell migration assay. After overnight starvation, 1×104 PC-3 or 2x104 Caco-2 cells previously transfected with h*SIGLEC12*-pcDNA3.1or empty pcDNA3.1(-), were seeded into the Transwell® inserts (HTS Transwell®-96 Permeable Support with 8.0 µm Pore Polyester Membrane, Corning, Nueva York, USA) in serum-free medium. The lower chamber was filled with 10% FBS containing medium. Stationary cells were removed from the upper surface of the membranes with a cotton swab. Cells that migrated to the lower surface were fixed and stained with crystal violet. Migrating cells were counted in five randomly selected fields and normalized to control.

Western blotting and antibodies. Cells were homogenized on ice in RIPA lysis buffer (Cell Signaling Technology, CST, Massachusetts, USA) supplemented with protease and phosphatase inhibitors (CST, USA). Protein concentration was quantified using a Pierce bicinchoninic acid (BCA) protein assay kit (Thermo Scientific, California, USA). Equal amounts of proteins were resolved by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred onto polyvinylidene difluoride (PVDF) membranes (Bio-Rad, California, USA). The membranes were blocked in Tris Buffered Saline with 0.1% Tween® 20 (TBST, CST, Massachusetts, USA) and 0.5% Bovine Serum Albumin (BSA) (Sigma-Aldrich, Missouri, USA) for 1 h at room temperature and then incubated with the primary antibodies at 4 °C overnight. The primary antibodies used for immunoblotting were anti-β-Actin (#4970, 1/10000, CST, Massachusetts, USA), anti-phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) antibody (#9101, 1/1000, CST, Massachusetts, USA), anti-p44/42 MAPK (Erk1/2) antibody (#9102, 1/1000, CST, Massachusetts, USA), and anti-Siglec-XII (AP18196PU-N, 1/2000, Origene, Maryland, USA). Then, membranes were incubated with IRDye® 800CW Goat anti-Rabbit IgG secondary antibody (1/15000, LI-COR Biosciences, Nebraska, USA). Protein bands were visualized using Odyssey® Imager (LI-COR Biosciences, Nebraska, USA).

Mouse Strains. Villin1-Cre transgenic mice, with a Cre recombinase gene introduced under the promoter of the Villin1 gene (24) were acquired from the Jackson Laboratory (Maine, USA). Human *SIGLEC12* conditional knock-in was produced by Cyagen (California, USA). The gRNA to mouse ROSA26 gene, the donor vector containing BGH pA-Kozak-human *SIGLEC12*CDSloxP-SV40 early pA-EGFP-loxP-CAG promoter cassette, and Cas9 mRNA were co-injected into fertilized mouse eggs to generate targeted conditional knock-in offspring. F0 founder animals were identified by PCR followed by sequence analysis, which were bred to wildtype mice to test germline transmission and F1 animal generation. The *SIGLEC12* knock-in mice presents the *egfp* gene (including stop codon) flanked between two loxP sites, and it is upstream of the Siglec-XII coding sequence. Mice with one floxed allele for *SIGLEC12* were crossed with Villin1Cre, to generate Villin1Cre with heterozygous floxed *SIGLEC12* progeny. The littermates containing only Villin1Cre were used as controls.

Tamoxifen Preparation and Administration. Tamoxifen (Sigma-Aldrich, Missouri, USA) was prepared as described previously (44). Eight-week-old mice were i.p. injected with 100 μL of tamoxifen stock solution (10 mg/ml) for 5 consecutive days and sacrificed either after 12 or 87 days of the first injection.

Immunohistochemistry studies. Slides with multi-tissue arrays of different human carcinomas were obtained from Super Bio Chips (Gagnum-gu, Korea). The sections were deparaffinized and blocked for endogenous biotin and peroxidase. Citrate buffer pH 6.0 was used for heat-induced epitope retrieval. A five-step signal amplification method was used, consisting of the application of mouse monoclonal anti-Siglec-XII antibody (clone 276, which has been described earlier (18)), followed by biotinylated donkey anti-mouse, horseradish peroxidase (HRP) (Jackson ImmunoResearch Laboratories, Pennsylvania, USA), Streptavidin, followed by application of the enzyme biotinyl tyramide, and then, labeled Streptavidin (Jackson ImmunoResearch Laboratories, Pennsylvania, USA).

For mouse samples, tissues were frozen in Optimum Cutting Temperature compound (OCT) and processed for frozen sections using the Leica cryostat. Slides were fixed for 1 min in acetone and, after 3 washes with TBST, incubated with anti-Siglec-XII (AP18196PU-N, Origene, Origene, Maryland, USA) antibody for 30 min at room temperature. After three washes with TBST, slides were incubated with Peroxidase AffiniPure™ Goat Anti-Rabbit IgG (Jackson ImmunoResearch Laboratories, Pennsylvania, USA) for 30 min at room temperature. For both human tissue array and frozen mouse tissue sections, the AEC kit (Vector, California, USA) was used as substrate, nuclear counterstain was carried out with Mayer's hematoxylin, and the slides were aqueous mounted for digital photographs, taken using the Olympus BH2 microscope.

Modeling colorectal carcinogenesis. Eight-week-old female mice received 5 days treatment (10 mg/ml) with tamoxifen followed by an intraperitoneal injection (10 mg/kg body weight) of AOM (Sigma-Aldrich, Missouri, USA) followed by 5 days of DSS (MP biomedicals, California, USA) treatment (2.0%) and 14 days of recovery, as described previously (28). This cycle was repeated three times. After the fourth DSS cycle (87 days), mice were sacrificed, and organs harvested for various analyses. This included small intestines, colons, kidneys, livers, lungs, and spleens. The intestines were opened and examined for the presence of tumors and the number of intestinal tumors was assessed. The size of the tumors was determined by ImagJ software.

Tissue histology. Colon, kidney, liver, spleen, and lung samples were immediately fixed in 10% neutral buffered formalin and processed into paraffin blocks and sectioned at 3 μm using a microtome and placed on slides. These slides were used for hematoxylin and eosin (H&E) staining. Digital photographs of H&E were taken using the Keyence BZ-9000E microscope. The Keyence microscope system was used to capture digital images at low power and the images were merged to obtain the final image of the roll of mouse intestine.

Study approval. Mice were housed at an animal facility of the University of California San Diego (UCSD). All mouse procedures were approved by The Institutional Animal Care and Use Committee (IACUC).

Computational Methodologies:

Curation of Publicly available Datasets: Several publicly available microarrays and RNASeq databases were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) server. Gene expression summarization was performed by normalizing Affymetrix platforms by RMA (Robust Multichip Average) and RNASeq platforms by computing TPM (Transcripts Per Millions) values whenever normalized data were not available in GEO. We used log2(TPM +1) as the final gene expression value for analyses. GEO accession numbers are reported in figures and text. For the dataset (GSE146009; 15 African American and 18 Caucasian American samples) containing RNA-seq data generated by TruSeq Stranded mRNA Library Prep Kit, we obtained it from NCBI GEO and subsequently cleaned to exclude paired tumor and normal mucosa samples with the mapping rate >90% (all 15 African American and 9 Caucasian American samples passed the QC check). Caucasians were defined as Americans with European ancestry; African Americans were defined as having any amount of ancestry contribution from Africa.

RNA seq on cells and mouse colons. Caco-2 cells and mouse tissues (distal colon after tamoxifen administration for baseline, and mouse colon tumors from AOM-DSS carcinogenesis protocol) were subjected to mRNA extraction using RNeasy plus mini kit (Qiagen, Venlo, Netherlands). Sample quality control was evaluated by the TapeStation system (Agilent, California, USA). Transcriptomic analysis was performed on RNA libraries prepared from samples not expressing or expressing Siglec-XII using the Illumina Stranded mRNA Prep. Each sample was used to prepare three separate technical replicate libraries for sequencing. Libraries were sequenced at 2 × 100 bp on NovaSeq 6000 (Illumina, California, USA). Reads were mapped to human reference genome Hg19 using kallisto 0.44.0 pipeline. Mapped reads were counted at the gene level using featureCounts v1.5.220 and counts were analyzed using DESeq2 v1.14.1.21. Sample clustering was confirmed by principal component analysis (PCA), which is an unsupervised learning algorithm technique used to examine the interrelations among a set of variables. Differentially expressed genes with a p value ≤0.05 and fold change ≥2 was then selected for further examination.

Gene Expression Analyses. The expression levels of all genes in these datasets were converted to binary values (high or low) using the StepMiner algorithm (45) which undergoes an adaptive regression scheme to verify the best possible up and down steps based on sum-of-square errors. The steps are placed between data points at the sharpest change between expression levels, which gives us the information about threshold of the gene expression-switching event. To fit a step function, the algorithm evaluates all possible steps for each position and computes the average of the values on both sides of a step for the constant segments. An adaptive regression scheme is used that chooses the step positions that minimize the square error with the fitted data. Finally, a regression test statistic is computed as follows:

$$F\ stat = \frac{\sum_{i=1}^{n}(\widehat{X_\iota} - \bar{X})^2/(m-1)}{\sum_{i=1}^{n}(X_i - \widehat{X_\iota})^2/(n-m)}$$

Where X_i for $i = 1$ to n are the values, $\widehat{X_\iota}$ for $i = 1$ to n are fitted values. M is the degrees of freedom used for the adaptive regression analysis. $\bar{X}$ is the average of all the values:

$$\bar{X} = \frac{1}{n} * \sum\nolimits_{j=1}^{n} X_j$$

For a step position at k, the fitted values $\widehat{X_\iota}$ are computed by using:

$$\frac{1}{k} * \sum\nolimits_{j=1}^{n} X_j$$

for $i = 1$ to k and

$$\frac{1}{(n-k)} * \sum\nolimits_{j=k+1}^{n} X_j$$

for $i = k + 1$ to n.

Gene expression values were normalized according to a modified Z-score approach centered around StepMiner threshold (formula = (expr – SThr)/3*stddev). The normalized expression values for all genes were added together to create the final score for the gene signature. The samples were ordered based on the final signature score. Differentially expressed genes are identified using DESeq2 package in R. Standard t-tests were performed using Python scipy.stats.ttest_ind package (version 0.19.0) with Welch's two-sample t-test (unpaired, unequal variance (equal_var = False), and unequal sample size) parameters. Multiple hypothesis correction was performed by adjusting p-values with statsmodels.stats.multitest.multipletests (fdr_bh: Benjamini/Hochberg principles). Pathway analysis of gene lists was carried out via the Reactome database and GO Biological processes.

Measurement of classification strength or prediction accuracy. Receiver operating characteristic (ROC) curves were computed by simulating a score based on the ordering of samples that illustrates the diagnostic ability of binary classifier system as its discrimination threshold is varied along with the sample order. The ROC curves were created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the curve (often referred to as simply the AUC) is equal to the probability that a classifier will rank a randomly chosen CRC samples higher than a randomly chosen healthy samples. In addition to ROC AUC, other classification metrics such as accuracy ((TP + TN)/N; TP: True Positive; TN: True Negative; N: Total Number), precision (TP/(TP+FP); FP: False Positive), recall (TP/(TP+FN); FN: False Negative) and f1 (2 * (precision * recall)/(precision + recall)) scores were computed.

Precision score represents how many selected items are relevant and recall score represents how many relevant items are selected. Python Scikitlearn package was used to compute the ROC-AUC values. Fisher exact test is used to examine the significance of the association (contingency) between two different classification systems (one of them can be ground truth as a reference).

Unsupervised clustering and Heatmap. Expression patterns of the genes that are differentially expressed in African American Caucasian American samples (in GSE146009) and *SIGLEC-12* expressing and control groups, before or after AOM/DSS challenge are clustered without bias based on their z-normalized cpm expression values, in all the samples. The data is visualized using the seaborn clustermap package (v 0.12) in python.

Multivariate Analyses. To assess which demographic and clinicopathologic factor(s) may influence Siglec-XII expression in CRCs, multivariate regression has been performed on a tumor microarray dataset. Multivariate analysis models the *SIGLEC-12* expression in samples (base variable) as a linear combination of all other metadata that was associated with each tumor, i.e., clinical (stage, pTNM, location), demographic (age/gender), or histopathological parameters. Here, the stats models module from python has been used to perform Ordinary least-squares (OLS) regression analysis of each of the variables. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect).

Kaplan-Meier Survival Plots. Survival analysis was performed using "Use multiple genes" options on Kaplan-Meier Plotter (46) and running the analysis on the *SIGLEC12* gene signature using the ‘default setting’ using the mean expression of the genes.

Statistics. Statistical significance was calculated using Prism 10 statistical software (GraphPad, Inc. California, USA). The data presented in this study is expressed as mean values ± SD. Normality test was performed prior to statistical test. For comparisons between two independent samples, T-Test was used. For multiple comparisons ANOVA, followed by Tukey’s multiple comparisons test, was performed. The data correspond to at least three independent experiments. A statistically significant value was defined as $p < 0.05$.

Data availability. RNA sequencing data have been made available publicly through the NCBI GEO repository (GSE262088), and in the “Supporting data values” XLS file.

AUTHOR CONTRIBUTIONS TO MANUSCRIPT

HAC, NV, AV, and PG conceptualized the project. HAC performed the experiments and analyzed the results. SS and PG conducted all computational analyses in this work. HAC, NV, AV, SS and PG prepared display items for data visualization. HAC, NV, AV, and PG wrote the original draft of the manuscript. All authors provided input and edited and revised the manuscript. All co-authors approved the final version of the manuscript. AV and PG coordinated and supervised all parts of the project.

ACKNOWLEDGEMENTS

This study was initiated with support from R01GM32373 (AV). This work was supported by the National Institutes for Health (NIH) grants UG3TR002968, UH3 TR002968, R01-CA238042, R01-AI155696 and R01-AI141630 (to PG). Other sources of support include Padres Pedal the Cause awards #PTC2021 and #PTC2022, Curebound Foundation awards #23DG04 and #21DG03. S.S was also supported in part through funds from the American Association of Immunologists (AAI) Intersect Fellowship Program for Computational Scientists and Immunologists. This manuscript includes data generated at the UC San Diego Institute of Genomic Medicine (IGM) using an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929). HAC was supported by the Pew Latin American Fellows program. We are grateful for access to the computational resources at the Center for Precision Computational Systems Network (PreCSN), UC San Diego Institute for Network Medicine.

FOOTNOTES

Abbreviations: AOM, azoxymethane; CRC, colorectal cancer; CAP, polyps that progress to CRC; CFP, cancer-free polyps; DSS, dextran sulfate sodium salt; DEGs, differentially expressed genes; H&E, hematoxylin and eosin staining; IECs, intestinal epithelial cells; Sia, Sialic acid; ITIMs, immunoreceptor tyrosine-based inhibitory motifs; pTNM Tumor, Node, Metastasis; MSI-high, microsatellite instability-high; PTPs, protein tyrosine phosphatases.

KEY WORDS

Siglecs, colorectal cancer, inflammation.

REFERENCES

1. Sung H, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3). https://doi.org/10.3322/caac.21660.

2. Varki NM, Varki A. On the apparent rarity of epithelial cancers in captive chimpanzees. *Philos Trans R Soc B Biol Sci*. 2015. https://doi.org/10.1098/rstb.2014.0225.

3. Goodman M, Grossman LI, Wildman DE. Moving primate genomics beyond the chimpanzee genome. *Trends Genet*. 2005;21(9). https://doi.org/10.1016/j.tig.2005.06.012.

4. Enard W. Functional primate genomics - Leveraging the medical potential. *J Mol Med*. 2012;90(5). https://doi.org/10.1007/s00109-012-0901-4.

5. Vaill M, et al. COMPARATIVE PHYSIOLOGICAL ANTHROPOGENY: EXPLORING MOLECULAR UNDERPINNINGS OF DISTINCTLY HUMAN PHENOTYPES. *Physiol Rev*. 2023;103(3). https://doi.org/10.1152/physrev.00040.2021.

6. Varki N, et al. Heart disease is common in humans and chimpanzees, but is caused by different pathological processes. *Evol Appl*. 2009;2(1). https://doi.org/10.1111/j.1752-4571.2008.00064.x.

7. Varki A. Nothing in medicine makes sense, except in the light of evolution. *J Mol Med*. 2012;90(5). https://doi.org/10.1007/s00109-012-0900-5.

8. Siddiqui SS, et al. Human-specific polymorphic pseudogenization of SIGLEC12 protects against advanced cancer progression. *FASEB BioAdvances*. 2021;3(2):69–82.

9. Thiesler H, et al. Proinflammatory Macrophage Activation by the Polysialic Acid-Siglec-16 Axis Is Linked to Increased Survival of Patients with Glioblastoma. *Clin Cancer Res*. 2023;29(12). https://doi.org/10.1158/1078-0432.ccr-22-1488.

10. Daëron M, et al. Immunoreceptor tyrosine-based inhibition motifs: A quest in the past and future. *Immunol Rev*. 2008;224(1). https://doi.org/10.1111/j.1600-065X.2008.00666.x.

11. Barrow AD, Trowsdale J. You say ITAM and I say ITIM, let’s call the whole thing off: The ambiguity of immunoreceptor signalling. *Eur J Immunol*. 2006;36(7). https://doi.org/10.1002/eji.200636195.

12. Varki A, Angata T. Siglecs - The major subfamily of I-type lectins. *Glycobiology*. 2006. https://doi.org/10.1093/glycob/cwj008.

13. Varki A, Schnaar RL, Crocker PR. I-Type Lectins -Essentials of Glycobiology -NCBI Bookshelf Chapter 35 I-Type Lectins. *Cold Spring Harb (NY*. 2015.

14. Leblanc C, et al. Epithelial Src homology region 2 domain–containing phosphatase-1 restrains intestinal growth, secretory cell differentiation, and tumorigenesis. *FASEB J*. 2017;31(8). https://doi.org/10.1096/fj.201601378R.

15. Gagné-Sansfaçon J, et al. SHP-2 phosphatase contributes to KRAS-driven intestinal oncogenesis but prevents colitis-associated cancer development. *Oncotarget*. 2016;7(40). https://doi.org/10.18632/oncotarget.11601.

16. Li B, et al. Expression signature, prognosis value, and immune characteristics of Siglec-15 identified by pan-cancer analysis. *Oncoimmunology*. 2020;9(1). https://doi.org/10.1080/2162402X.2020.1807291.

17. Leaubli H, Nalle SC, Maslyar D. Targeting the Siglec–Sialic Acid Immune Axis in Cancer: Current and Future Approaches. *Cancer Immunol Res*. 2022;10(12). https://doi.org/10.1158/2326-6066.CIR-22-0366.

18. Mitra N, et al. SIGLEC12, a human-specific segregating (pseudo)gene, encodes a signaling molecule expressed in prostate carcinomas. *J Biol Chem*. [published online ahead of print: 2011]. https://doi.org/10.1074/jbc.M111.244152.

19. Angata T, Varki NM, Varki A. A Second Uniquely Human Mutation Affecting Sialic Acid Biology. *J Biol Chem*. [published online ahead of print: 2001]. https://doi.org/10.1074/jbc.M105926200.

20. Yngvadottir B, et al. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am J Hum Genet*. [published online ahead of print: 2008]. https://doi.org/10.1016/j.ajhg.2009.01.008.

21. Flores R, et al. Siglec genes confer resistance to systemic lupus erythematosus in humans and mice. *Cell Mol Immunol*. [published online ahead of print: 2019]. https://doi.org/10.1038/cmi.2017.160.

22. Yu Z, et al. Identification and Characterization of S2V, a Novel Putative Siglec That Contains Two V Set

Ig-like Domains and Recruits Protein-tyrosine Phosphatases SHPs. *J Biol Chem*. [published online ahead of print: 2001]. https://doi.org/10.1074/jbc.M102394200.

23. Kucherlapati MH. Mouse models in colon cancer, inferences, and implications. *iScience*. 2023;26(6). https://doi.org/10.1016/j.isci.2023.106958.

24. El Marjou F, et al. Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis*. [published online ahead of print: 2004]. https://doi.org/10.1002/gene.20042.

25. Madison BB, et al. cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J Biol Chem*. [published online ahead of print: 2002]. https://doi.org/10.1074/jbc.M204935200.

26. Arnesen H, et al. Induction of colorectal carcinogenesis in the C57BL/6J and A/J mouse strains with a reduced DSS dose in the AOM/DSS model. *Lab Anim Res*. 2021;37(1). https://doi.org/10.1186/s42826-021-00096-y.

27. Neufert C, et al. Inducible mouse models of colon cancer for the analysis of sporadic and inflammation-driven tumor progression and lymph node metastasis. *Nat Protoc*. 2021;16(1). https://doi.org/10.1038/s41596-020-00412-1.

28. Allen IC, et al. The NLRP3 inflammasome functions as a negative regulator of tumorigenesis during colitis-associated cancer. *J Exp Med*. 2010;207(5). https://doi.org/10.1084/jem.20100050.

29. Dzhalilova D, et al. Murine models of colorectal cancer: the azoxymethane (AOM)/dextran sulfate sodium (DSS) model of colitis-associated cancer. *PeerJ*. 2023;11. https://doi.org/10.7717/PEERJ.16159.

30. Pillai S, et al. Siglecs and immune regulation. *Annu Rev Immunol*. 2012;30. https://doi.org/10.1146/annurev-immunol-020711-075018.

31. Okayasu I, et al. A novel method in the induction of reliable experimental acute and chronic ulcerative colitis in mice. *Gastroenterology*. 1990;98(3). https://doi.org/10.1016/0016-5085(90)90290-H.

32. Druliner BR, et al. Molecular characterization of colorectal adenomas with and without malignancy reveals distinguishing genome, transcriptome and methylome alterations. *Sci Rep*. 2018;8(1). https://doi.org/10.1038/s41598-018-21525-4.

33. Druliner BR, et al. Colorectal cancer with residual polyp of origin: A model of malignant transformation. *Transl Oncol*. 2016;9(4). https://doi.org/10.1016/j.tranon.2016.06.002.

34. Druliner BR, et al. Time Lapse to Colorectal Cancer: Telomere Dynamics Define the Malignant Potential of Polyps. *Clin Transl Gastroenterol*. 2016;7(9). https://doi.org/10.1038/ctg.2016.48.

35. Kim TM, et al. Clonal origins and parallel evolution of regionally synchronous colorectal adenoma and carcinoma. *Oncotarget*. 2015;6(29). https://doi.org/10.18632/oncotarget.4834.

36. Baran B, et al. Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. *Gastroenterol Res*. 2018;11(4). https://doi.org/10.14740/gr1062w.

37. Paredes J, et al. Immune-Related Gene Expression and Cytokine Secretion Is Reduced Among African American Colon Cancer Patients. *Front Oncol*. 2020;10. https://doi.org/10.3389/fonc.2020.01498.

38. Carethers JM. Racial and ethnic disparities in colorectal cancer incidence and mortality. *Advances in Cancer Research*. 2021.

39. Varki A. Uniquely human evolution of sialic acid genetics and biology. *Proc Natl Acad Sci U S A*. 2010;107(SUPPL. 2). https://doi.org/10.1073/pnas.0914634107.

40. Li C, et al. PTPN18 promotes colorectal cancer progression by regulating the c-MYC-CDK4 axis. *Genes Dis*. 2021;8(6). https://doi.org/10.1016/j.gendis.2020.08.001.

41. Lin WR, et al. Dynamic bioenergetic alterations in colorectal adenomatous polyps and adenocarcinomas. *EBioMedicine*. 2019;44. https://doi.org/10.1016/j.ebiom.2019.05.031.

42. Hewish M, et al. Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nat Rev Clin Oncol*. 2010;7(4). https://doi.org/10.1038/nrclinonc.2010.18.

43. Gatalica Z, et al. High microsatellite instability (MSI-H) colorectal carcinoma: a brief review of predictive biomarkers in the era of personalized medicine. *Fam Cancer*. 2016;15(3). https://doi.org/10.1007/s10689-016-9884-6.

44. Metzger D, Chambon P. Site- and time-specific gene targeting in the mouse. *Methods*. 2001;24(1). https://doi.org/10.1006/meth.2001.1159.

45. Sahoo D, et al. Boolean implication networks derived from large scale, whole genome microarray

datasets. *Genome Biol*. 2008;9(10). https://doi.org/10.1186/gb-2008-9-10-r157.

46. Nagy Á, et al. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci Rep*. 2018;8(1). https://doi.org/10.1038/s41598-018-27521-y.

FIGURES AND FIGURE LEGENDS

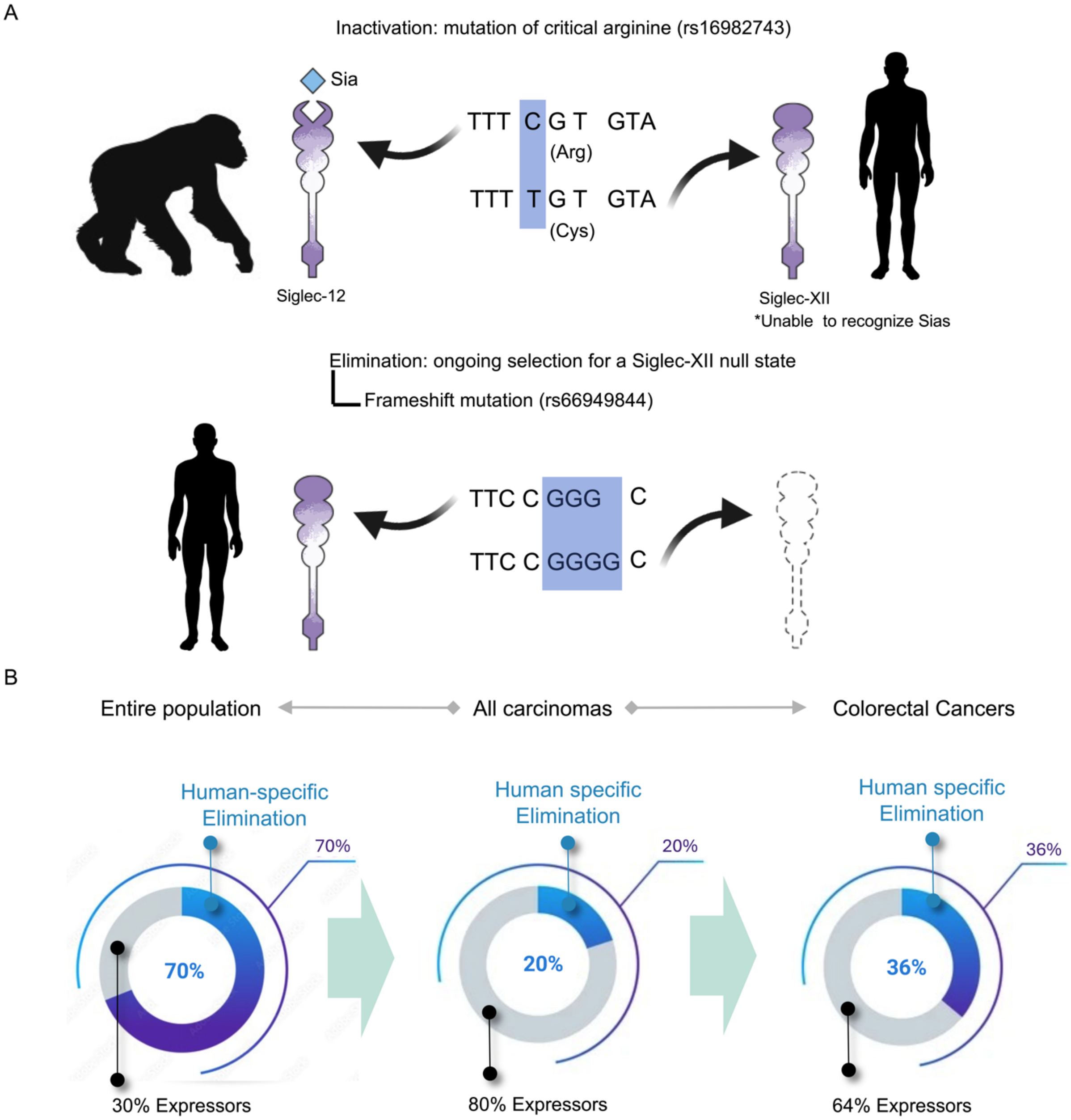

A
Inactivation: mutation of critical arginine (rs16982743)
Sia
TTT C G T GTA
(Arg)
TTT T G T GTA
(Cys)
Siglec-12
Siglec-XII
*Unable to recognize Sias
Elimination: ongoing selection for a Siglec-XII null state
Frameshift mutation (rs66949844)
TTC C GGG C
TTC C GGGG C
B
Entire population
All carcinomas
Colorectal Cancers
Human-specific Elimination
70%
70%
30% Expressors
Human specific Elimination
20%
20%
80% Expressors
Human specific Elimination
36%
36%
64% Expressors

Figure 1. Mechanism and the observed prevalence of human-specific inactivation and elimination of the protein product of *SIGLEC12*. **A.** Schematic (top) summarizes the impact of the human universal mutation (rs16982743) of the gene *SIGLEC12* which results in a loss of an essential arginine which abolishes the ability of the Siglec-XII protein to bind/recognize sialic acids (Sias). This functionally inactivating mutation occurred prior to the common ancestor of all modern humans, *SIGLEC12* is intact and functional in great apes. Schematic (bottom) summarizes the ongoing selection for the Siglec-XII null state that continues in the current world-wide human population. The most common polymorphic mutation is a frameshift mutation (rs66949844), guanine (G) insertion, that in the homozygous state eliminates the protein expression in most humans. **B.** Pie charts indicate the restricted prevalence of Siglec-

XII expression (~30%) in the entire human population (left), but the enrichment of such expressors among all (middle) and colorectal (right) carcinomas.

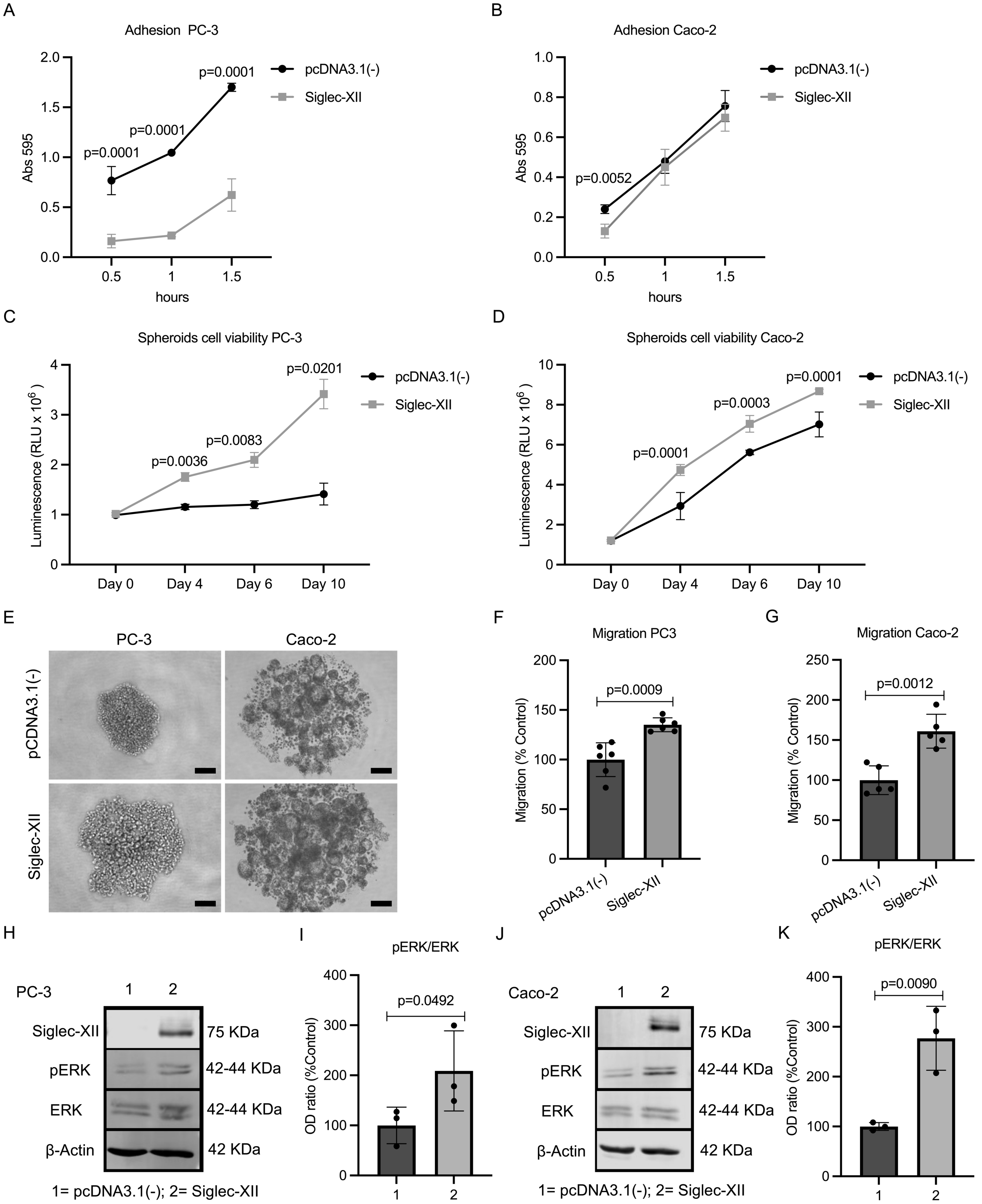


Figure 2. Forced expression of Siglec-XII in null human carcinoma cell lines enhances cellular processes associated with tumor aggressiveness. **A-B.** Graphs display cell adhesion on 2D surface for PC3 (A) and Caco-2 (B) cells, as measured by crystal

violet staining. **C-E.** Graphs (C, PC3; D, Caco-2) display cellular viability of the same cells in 3D tumoroid cultures. Representative images are displayed (E). Scale bar: 100 µm**. F-G.** Graphs display % migration of PC-3 (F) and Caco-2 (G) cells, as determined by Transwell® assays (0-10% serum gradient). **H-K.** Quantitative immunoblotting on equal aliquots of whole cell lysates of PC3 (H-I) or Caco-e (J-K) cells to assess ERK1/2 activity. Blots results were set up in parallel, run contemporaneously and normalized to loading controls (β-Actin). OD, optical density. Representative immunoblots are shown in H and J, and quantification of 3 independent repeats are shown as bar graphs in I and K. Error bars indicate ± S.D. See also **Supplementary Figure S1** for approaches used to confirm the Siglec-XII null state.

Statistics: P values were calculated using GraphPad Prism, p value <0.05 was considered as significant. A, B, C and D, ANOVA followed by Tukey's multiple comparisons post hoc test. F, G, and K, 2- tailed t-Test. I, 1- tailed t-Test.

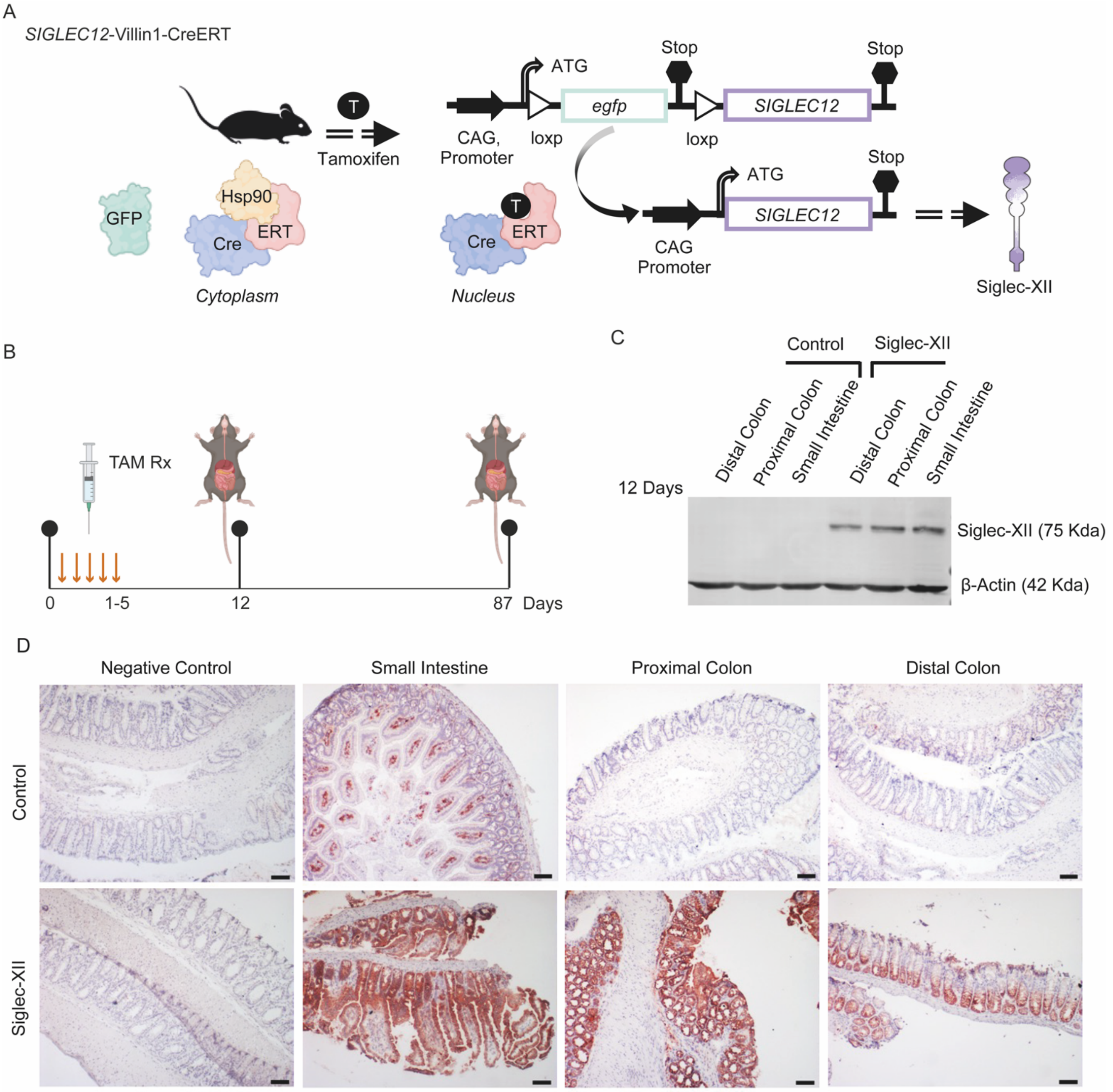
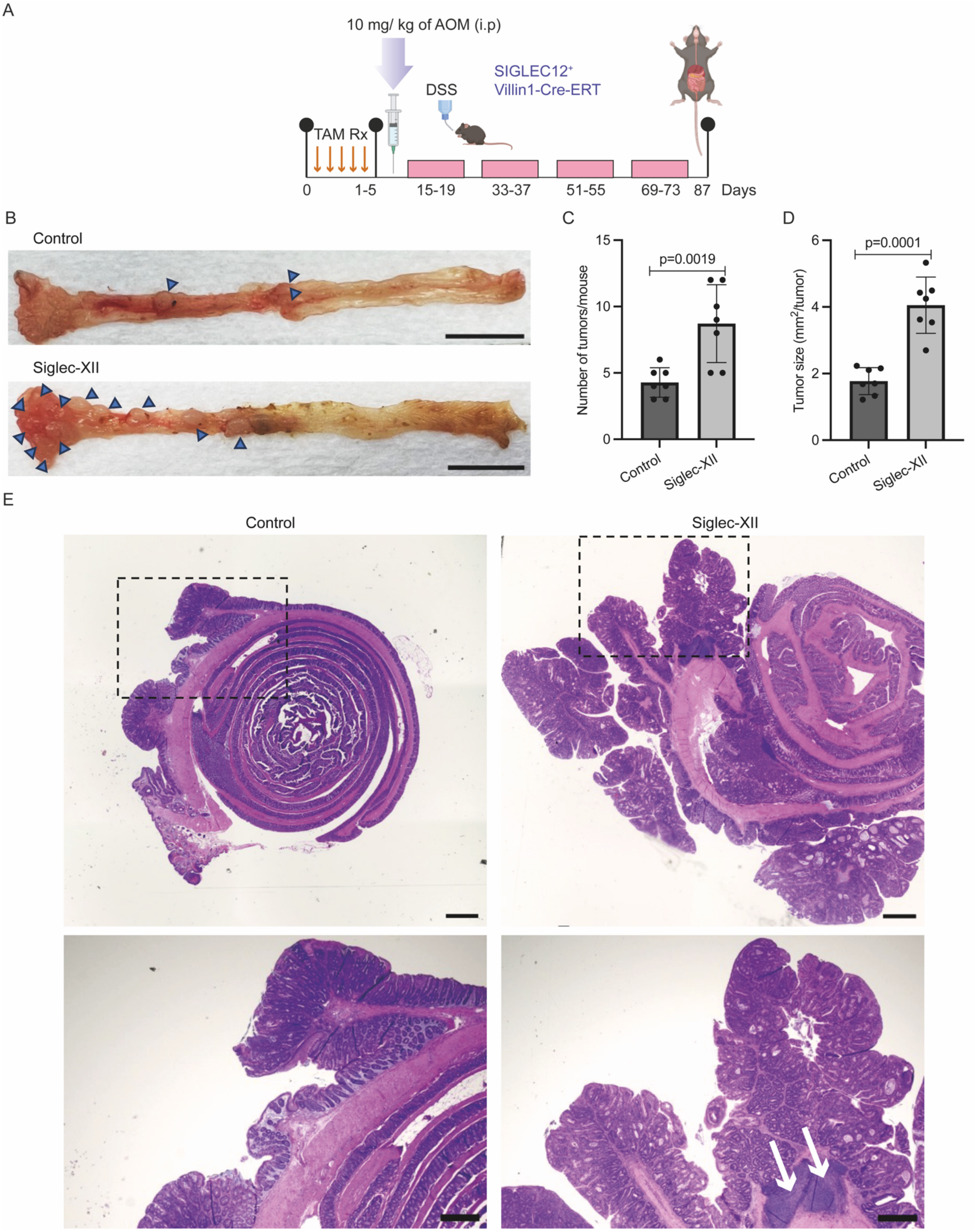


Figure 3. Creation and validation of a transgenic knock-in *SIGLEC12* murine model that expresses Siglec-XII in the small and large intestine. **A.** Schematic displays the cloning strategy for creation of the *SIGLEC12* knock-in mice (wild-type mice do not harbor a *SIGLEC12* gene). Tissue specific Cre-driver mice express a tamoxifen-inducible System of estrogen receptor fused to Cre (Cre-ERT). In absence of Tamoxifen (T), Hsp90 binds to Cre-ERT and maintains its cytoplasmic retention. Nuclear translocation of Cre-ERT by tamoxifen. In the nucleus, Cre-ERT recognizes loxP sites and allows tissue-specific expression of Siglec-XII. **B.** An overview of experimental design for the induction of Siglec-XII expression by serial administration of Tamoxifen on five consecutive days, followed by harvesting of tissues to confirm early (day 12) and sustained (day 87) expression of Siglec-XII. **C.** Western blot for Siglec-XII and β-Actin on transgenic mouse and control tissues at day 12 post induction using Tamoxifen. **D.** Expression of Siglec-XII in mouse tissue evaluated by immunohistochemistry at day 12 post induction using Tamoxifen. Scale bar: 100 μm. See also **Supplementary Figure S2A** (for immunoblots) and **S2B** (for immunohistochemistry) on samples at day 87 post induction.


A
10 mg/ kg of AOM (i.p)
DSS
SIGLEC12+
Villin1-Cre-ERT
TAM Rx
0
1-5
15-19
33-37
51-55
69-73
87
Days
B
Control
Siglec-XII
C
Number of tumors/mouse
15
10
5
0
p=0.0019
Control
Siglec-XII
D
Tumor size (mm2/tumor)
6
4
2
0
p=0.0001
Control
Siglec-XII
E
Control
Siglec-XII

687

Figure 4. Transgenic knock-in *SIGLEC12* mice are at greater risk of inflammation-associated colorectal cancers. A. An overview of experimental design for the induction of Siglec-XII expression by serial administration of Tamoxifen followed by carcinogenesis protocol consisting of a single administration of azoxymethane (AOM) and four cycles of dextran sodium sulfate (DSS). **B.** Representative pictures of colonic tissue from control and Siglec-XII-expressing mice subjected to tamoxifen administration and carcinogenesis protocol (AOM/DSS). Scale bar: 1 cm. The complete panel of pictures of colonic tissue is shown in **Supplemental Figure S4**. **C-D.** Comparison of the number of tumors (C) and tumor size (D) in control (N=7) and Siglec-XII-expressing mice (N=7) subjected to tamoxifen administration and carcinogenesis protocol. Error bars indicate ± S.D. **E.** Representative pictures of H&E-stained colonic tissue from control and Siglec-XII-expressing mice subjected to tamoxifen administration and carcinogenesis protocol, the boxed areas (top) are shown at higher magnification (bottom). Arrows indicate immune cell infiltrates. Scale bars: 100 μm (top), 50 μm (bottom). *Statistics*: P values were calculated by 2-tailed t-test using GraphPad Prism, p value <0.05 was considered as significant.

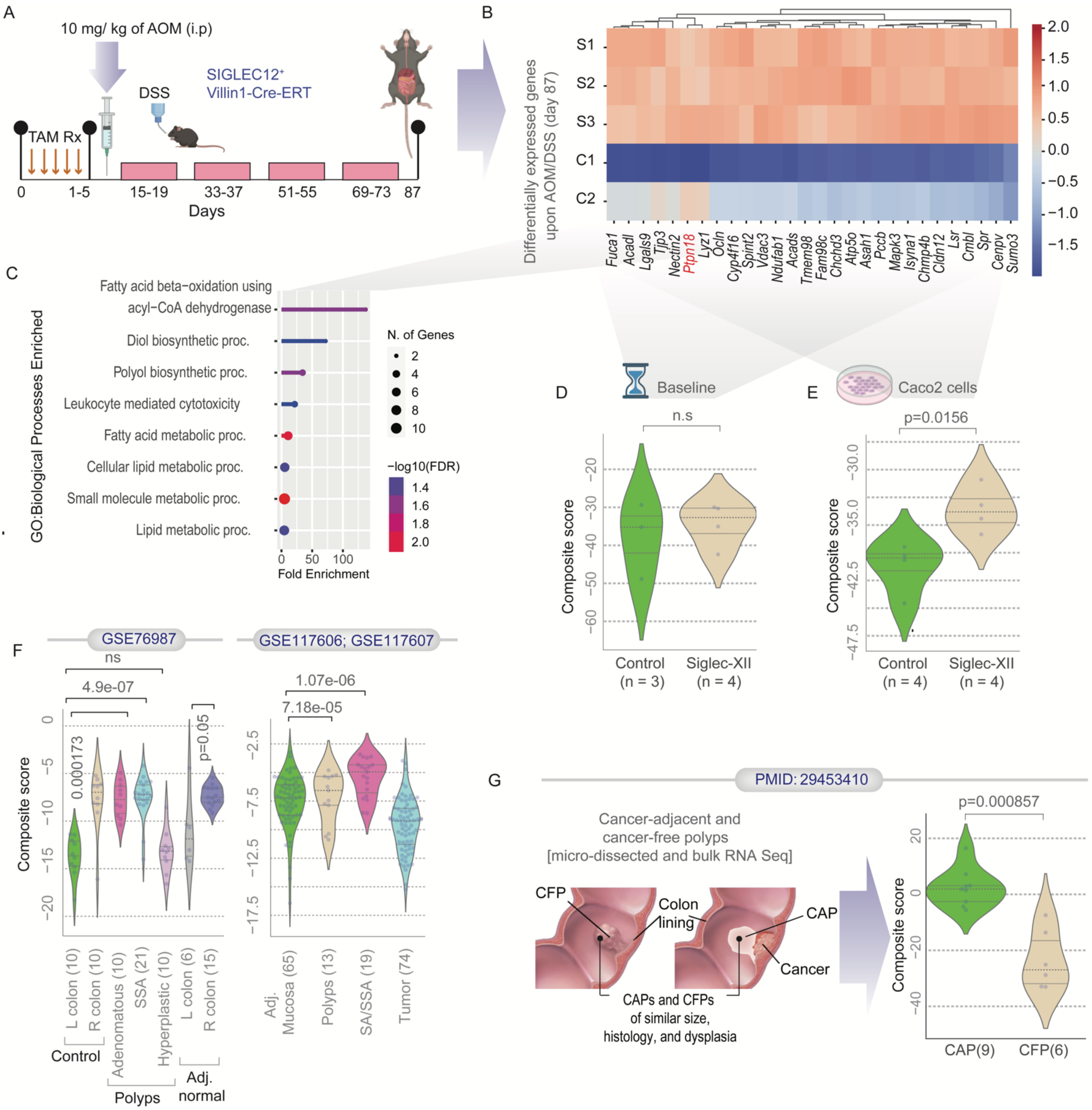


Figure 5. *SIGLEC12* expression induces gene expression in polyps at risk of progression to CRCs. **A**. An overview of experimental design for the induction of Siglec-XII expression by serial administration of Tamoxifen followed by carcinogenesis protocol consisting of a single administration of AOM and four cycles of DSS. On day 87, mouse colon tumors were harvested for RNA Seq analysis. **B**. Heatmap shows the z normalized expression pattern of upregulated differentially expressed genes (DEGs) in Siglec-XII-expressing cohort. **C**. Plot showing the fold enrichment of various biological processes from the Gene Ontology (GO) database. **D-E**. Violin plots display the *StepMiner* normalized composite scores of the DEGs (in B) in control vs Siglec XII samples at

baseline (day 12; D) (collected within 1 week after tamoxifen administration) and Caco-2 pcDNA3.1(-) and Siglec-XII-Caco-2 (E**). F.** Violin plots display the *StepMiner* normalized composite scores of the DEGs (in B) in patient tissues from three independent cohorts (GSE76987, GSE117606, and GSE117607). R, right; L, left; SSA, sessile serrated adenoma; Adj., adjacent. **G**. Violin plots display the *StepMiner* normalized composite scores of the DEGs (in B) in laser microdissected adenomatous tissues from polyps that progressed to cancers [cancer-adjacent polyps (CAP)] vs. those which did not [cancer-free polyps (CFP)]. *Statistics*: *p* values in each violin plot (D-H) are based on Welch's T-test between comparator groups. *p* values for survival plots were determined by log rank test.

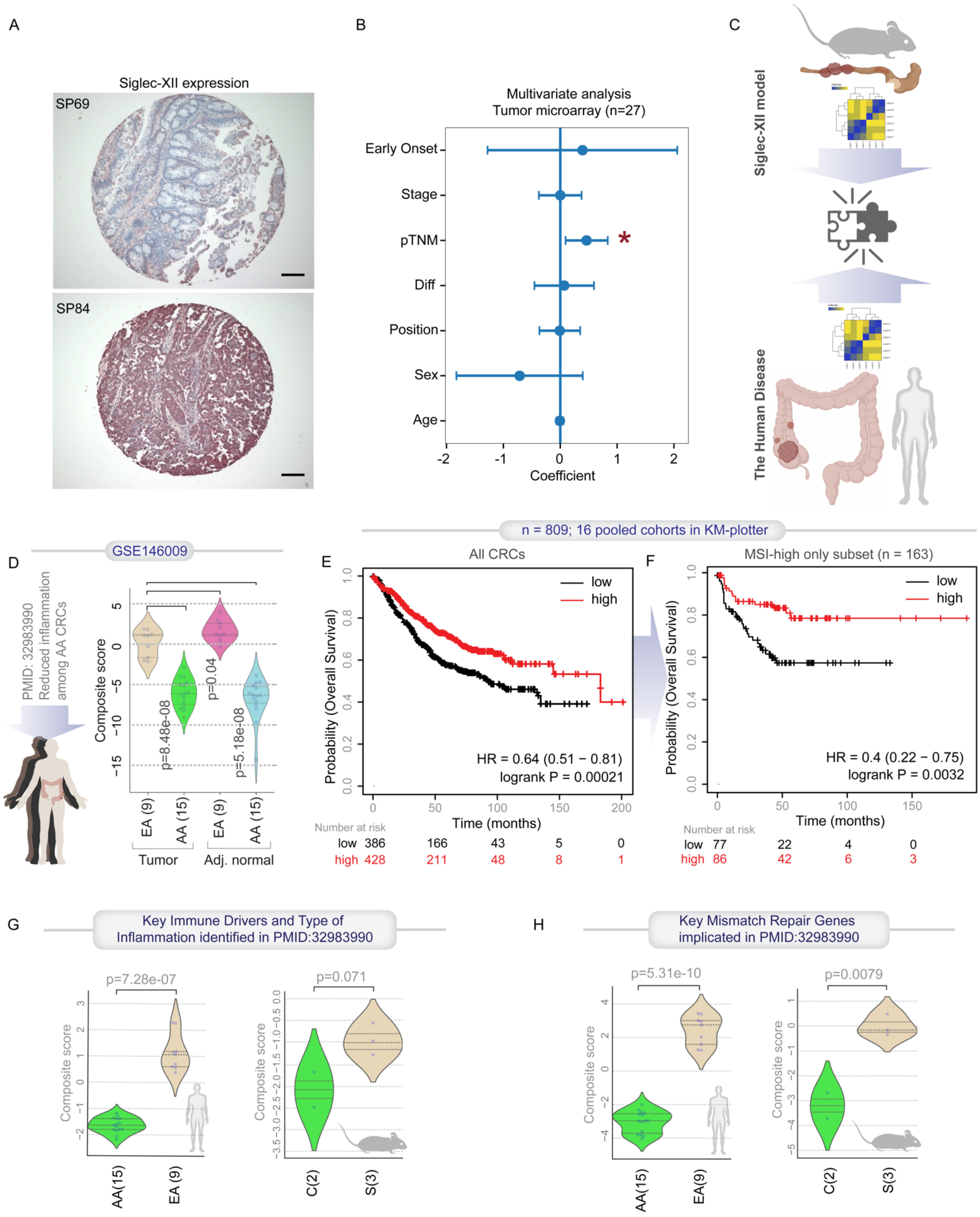

A
Siglec-XII expression
SP69
SP84
B
Multivariate analysis
Tumor microarray (n=27)
Early Onset
Stage
pTNM
Diff
Position
Sex
Age
*
-2
-1
0
1
2
Coefficient
C
Siglec-XII model
The Human Disease
D
GSE146009
PMID: 32983990
Reduced inflammation
among AA CRCs
Composite score
5
0
−5
−10
−15
p=8.48e-08
p=0.04
p=5.18e-08
EA (9)
AA (15)
EA (9)
AA (15)
Tumor
Adj. normal
n = 809; 16 pooled cohorts in KM-plotter
E
All CRCs
Probability (Overall Survival)
1.0
0.8
0.6
0.4
0.2
0.0
low
high
HR = 0.64 (0.51 − 0.81)
logrank P = 0.00021
0
50
100
150
200
Time (months)
Number at risk
low 386 166 43 5 0
high 428 211 48 8 1
F
MSI-high only subset (n = 163)
Probability (Overall Survival)
1.0
0.8
0.6
0.4
0.2
0.0
low
high
HR = 0.4 (0.22 − 0.75)
logrank P = 0.0032
0
50
100
150
Time (months)
Number at risk
low 77 22 4 0
high 86 42 6 3
G
Key Immune Drivers and Type of
Inflammation identified in PMID:32983990
p=7.28e-07
Composite score
3
2
1
0
−1
−2
AA(15)
EA (9)
p=0.071
Composite score
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
C(2)
S(3)
H
Key Mismatch Repair Genes
implicated in PMID:32983990
p=5.31e-10
Composite score
4
2
0
−2
−4
AA(15)
EA(9)
p=0.0079
Composite score
1
0
−1
−2
−3
−4
−5
C(2)
S(3)

Figure 6. Computational Siglec-XII positive CRCs. **A.** Expression of Siglec-XII in human colorectal tumors was evaluated by immunohistochemistry. Representative images of tumors that were scored as negative (specimen 69) or positive (specimen 84) are shown. Scale bar: 100 µm. **B.** Multivariate analysis of Siglec-XII positivity as a linear combination of all variables in the tumors used in this study. The coefficient of each variable (at the center) with their upper and lower bounds of 95% confidence interval (as error bars) and the p-values from t-tests are illustrated in the bar plot. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). Asterisk = significant co-variate. *p 0.018. See also **Supplementary Table 1** for source data. **C**. Schematic summarizes the transcriptomics-based computational approach to find a match between model (Siglec-XII murine tumors) vs disease (human CRCs). **D**. Violin plots display the *StepMiner* normalized composite scores of the DEGs (in B) in tumor and matched adjacent normal colon tissues in 15 African American (AA) and 9 European American (EA) patients. **E-F**. Kaplan-Meier curves for overall (I-J) and progression-free survival (**Supplementary Figure S5A-B**) in patients with all CRCs (E) or just the MSI-high subset (F), stratified based on high vs low mean expression values of the DEGs in B. **G-H.** Violin plots of the *StepMiner* normalized composite scores of key immune (G) and mismatch repair (H) genes that were found to be differentially expressed between the two ethnic groups in GSE146009 (left) and in the control (C) vs Siglec-XII (S) mouse tumors (right). See also **Supplementary Figure S5C-F** for the patterns of expression of the individual genes displayed as heatmaps. *Statistics*: *p* values in each violin plot (D, G-H) are based on Welch’s T-test between comparator groups. *p* values for survival plots were determined by log rank test.

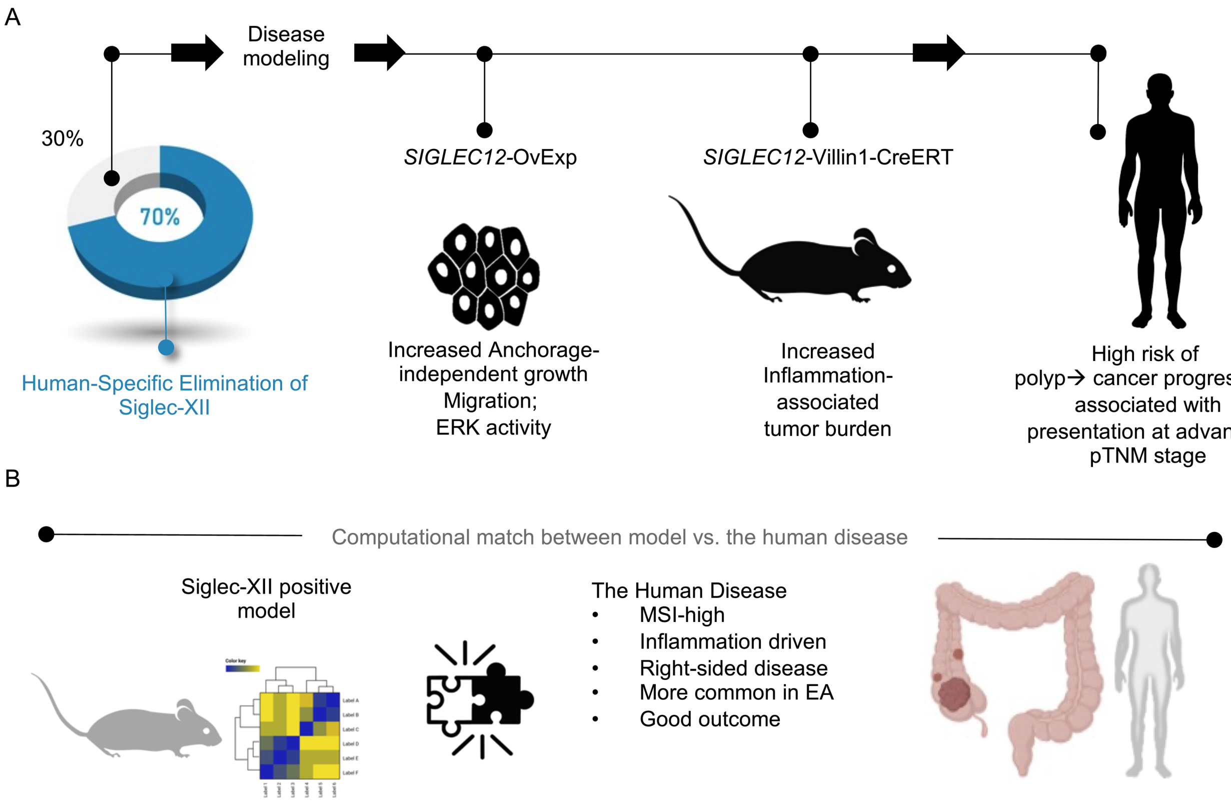

A
Disease modeling
30%
70%
Human-Specific Elimination of Siglec-XII
SIGLEC12-OvExp
Increased Anchorage-independent growth
Migration;
ERK activity
SIGLEC12-Villin1-CreERT
Increased Inflammation-associated tumor burden
High risk of polyp→ cancer progression; associated with presentation at advanced pTNM stage
B
Computational match between model vs. the human disease
Siglec-XII positive model
The Human Disease
• MSI-high
• Inflammation driven
• Right-sided disease
• More common in EA
• Good outcome

Figure 7. Summary of findings. A. Schematic summarizes the major goal, key model systems, and the key findings made using each model system in the current study. Three model systems were used, each seeking to model the oncogenic risk posed by continued Siglec-XII expression in humans (~30% of the population) despite evolutionary loss in most of the population. **B.** Schematic summarizes the key conclusions drawn from an unbiased navigation of the human disease, performed using an objective assessment of transcriptomic datasets using a model-derived gene signature. EA: European American.