**Scurr, Martin J, Greenshields-Watson, Alex, Campbell, Emma, Somerville, Michelle S, Chen, Yuan, Hulin-Curtis, Sarah L ORCID logoORCID: https://orcid.org/0000-0003-0889-964X, Burnell, Stephanie EA, Davies, James A, Davies, Michael M, Hargest, Rachel, Phillips, Simon, Christian, Adam D, Ashelford, Kevin E, Andrews, Robert, Parker, Alan L, Stanton, Richard J, Gallimore, Awen and Godkin, Andrew (2020) Cancer Antigen Discovery Is Enabled by RNA Sequencing of Highly Purified Malignant and Nonmalignant Cells. Clinical Cancer Research, 26 (13). pp. 3360-3370. doi:10.1158/1078-0432.CCR-19-3087**

Official URL: http://doi.org/10.1158/1078-0432.CCR-19-3087
DOI: http://dx.doi.org/10.1158/1078-0432.CCR-19-3087
EPrint URI: https://eprints.glos.ac.uk/id/eprint/14142

# Cancer antigen discovery is enabled by RNA-sequencing of highly purified malignant and non-malignant cells

Martin J. Scurr[1^], Alex Greenshields-Watson[1^], Emma Campbell[1], Michelle S. Somerville[1], Yuan Chen[1], Sarah L. Hulin-Curtis[1], Stephanie E. A. Burnell[1], James A. Davies[2], Michael M. Davies[3], Rachel Hargest[3], Simon Phillips[3], Adam Christian[4], Kevin E. Ashelford[2], Robert Andrews[1], Alan L. Parker[2], Richard J. Stanton[1], Awen Gallimore[1*^] and Andrew Godkin[1,5*^]

[1] Division of Infection & Immunity, Henry Wellcome Building, Cardiff University, Cardiff, UK.
[2] Division of Cancer & Genetics, Sir Geraint Evans Building, Cardiff University, Cardiff, UK.
[3] Dept. of Colorectal Surgery, University Hospital of Wales, Heath Park, Cardiff, UK.
[4] Dept. of Histopathology, University Hospital of Wales, Heath Park, Cardiff, UK.
[5] Dept. of Gastroenterology & Hepatology, University Hospital of Wales, Heath Park, Cardiff, UK.

[^] *These authors contributed equally to this work.*

* *Corresponding Authors:*
Division of Infection and Immunity
Henry Wellcome Building
Health Park
Cardiff
CF14 4XN
Tel: +442920687003 / +442920687012
Emails: godkinaj@cardiff.ac.uk
gallimoream@cardiff.ac.uk

**Running Title:** Identifying novel, immunogenic tumor antigens

**Conflict of Interest:** M. Scurr, A. Gallimore and A. Godkin are co-inventors of a patent regarding DNAJB7 and uses thereof, co-owned by Cardiff University. All other authors declare no potential conflicts of interest.

**Keywords:** Tumor antigens; RNA sequencing; DNAJB7; immunotherapy; cancer vaccine.

**Word count** (not including references): 4948

**Total number of figures** (not including supplementary): 6

## Translational Relevance

In order for cancer vaccination strategies to realise their potential, they must elicit effective anti-tumor immune responses in a broad patient population. Tumor cell purification dramatically improves RNA sequencing resolution to the point where novel, highly differentially expressed, immunogenic proteins become detectable. This novel methodology of tumor antigen identification could enhance future vaccine efficacy.

## Abstract

**Purpose:** Broadly expressed, highly differentiated tumor-associated antigens (TAA) can elicit anti-tumor immunity. However, vaccines targeting TAAs have demonstrated disappointing clinical results, reflecting poor antigen selection and/or immunosuppressive mechanisms.

**Experimental design:** Here, a panel of widely expressed, novel colorectal TAAs were identified by performing RNA sequencing of highly purified colorectal tumor cells in comparison to patient-matched colonic epithelial cells; tumor cell purification was essential to reveal these genes. Candidate TAA protein expression was confirmed by immunohistochemistry, and pre-existing T cell immunogenicity towards these antigens tested.

**Results:** The most promising candidate for further development is DNAJB7 [DnaJ heat shock protein family (Hsp40) member B7], identified here as a novel cancer-testis antigen. It is expressed in many tumors and is strongly immunogenic in patients with cancers originating from a variety of sites. DNAJB7-specific T cells were capable of killing colorectal tumor lines in vitro, and the IFN-$\gamma^+$ response was markedly magnified by control of immunosuppression with cyclophosphamide in cancer patients.

**Conclusion:** This study highlights how prior methods that sequence whole tumor fractions (i.e. inclusive of alive/dead stromal cells) for antigen identification may have limitations**.** Through tumor cell purification and sequencing, novel candidate TAAs have been identified for future immunotherapeutic targeting.

## Introduction

Despite understandable excitement surrounding results from cancer immunotherapy studies, actual outcomes are disappointing. We recently demonstrated the principle that immunological responses generated to the 5T4 oncofetal antigen through MVA-5T4 (TroVax) vaccination can positively influence the outcome of patients with advanced colorectal cancer (CRC)(1). Whilst survival was significantly prolonged for vaccinated patients mounting an anti-5T4 response, all patients had progressed within 10-months. Indeed, stand-alone anti-cancer vaccines are rarely effective in the advanced disease setting; this could be the result of inherent mechanisms of local immunosuppression, or sub-optimal antigenic targets.

Many upregulated tumor-associated antigen (TAA) targets are often readily detectable in healthy tissue, e.g. the autoantigen carcinoembryonic antigen (CEA)- and directing immune responses against such antigens can lead to side-effects (2) and poor survival outcomes post-surgery (3). Thus, identification of TAAs that can be targeted by immunotherapy is a balance between expression on tumor and healthy tissue. The challenge is further complicated by T cell cross-reactivity which can result in off-target effects in distant tissue with potentially fatal consequences (4).

Whilst immunotherapies targeting neoepitopes hold promise, they are highly focused to the individual and currently prohibitively expensive to develop. For therapies relevant to the wider population such as cancer vaccines, antigens must be broadly expressed in the same tumor types of multiple individuals and present at minimal levels in healthy tissue. Ideal discovery pipelines would involve large scale analysis of TAA candidates followed by selection based on immunogenicity and tissue-specific expression. Candidates that fit these criteria could be explored further with cancer vaccination and CAR-T cell therapy. Indeed, vaccinations targeting nonmutated tumor antigens are capable of inducing robust T cell responses in cancer patients (1,5).

The development of RNA-sequencing in differential expression analysis provides an attractive methodology to initiate TAA discovery pipelines. However, this technology is limited by the heterogeneity of the tissue in question and is not extensively used in TAA discovery. For the colon, a mixture of immune cells, epithelium and stroma complicates expression profiles, limiting identification of significantly differential expressed genes especially when tumor immune infiltrate varies highly between individuals and tumor location. Purification of epithelial and tumor cells prior to RNA-sequencing analysis is a novel methodology developed to overcome tissue heterogeneity. In this study, we used EpCAM purification of tumor and healthy colonic epithelium at two sites to improve the resolution between expression profiles and thus aid identification of differentially expressed genes (DEG). Gene lists were created based on expression profiles between all tissues, and significance levels in a DESeq2 comparison analysis. These lists were analysed, and several genes selected for further investigation. Immunogenic analysis and tissue expression of the protein products of these genes in healthy tissues were used to select the best candidate for cancer immunotherapy.

## Materials and Methods

### Excision of colonic and tumor tissue

Tumor and paired background (unaffected) colon specimens were obtained from three patients undergoing anterior resection for primary rectal cancer at the University Hospital of Wales, Cardiff (see Supplementary Table 1 for patient characteristics). Autologous colon samples were cut from macroscopically normal sections of the excised tissue, both “near” (within 2 cm) and “far” (at least 10 cm) from the tumor site. All fresh tumor samples were derived from the luminal aspect of the specimen, so as not to interfere with histopathological staging. All patients and participants gave written, informed consent personally prior to inclusion. This study was conducted in accordance with the Declaration of Helsinki. The Wales Research Ethics Committee granted ethical approval for this study.

### Patient treatment schedule

Orally administered 50mg cyclophosphamide was taken twice-a-day on treatment days 1–7 and 15–21; no cyclophosphamide was taken on treatment days 8–14 or 22–106, or until patient relapsed. Peripheral blood samples (40ml) were taken at regular intervals during therapy.

### Purification of tissue samples

Background colon and tumor specimens were transported and washed in extraction medium supplemented with 2% human AB serum (Welsh Blood Service), gentamicin and Fungizone (ThermoFisher). Within 30-minutes of resection from a patient, samples were minced and forced through 70μm cell strainers to collect a single cell suspension. In no instances were collagenase or DNase treatments used. Dissociated cell preparations of tumor, near and far healthy colonic tissue were stained with Live/Dead fixable Aqua (ThermoFisher) followed by surface marker staining with CD3-APC (BioLegend) and EpCAM-PE (Miltenyi Biotec) antibodies. Samples were resuspended in FACS buffer (PBS, 2% BSA) prior to sorting into Live/Dead$^-$EpCAM$^+$CD3$^-$ populations on a FACS Aria III (BD). Tumor tissue also stained with CD3 and EpCAM antibodies was additionally passed through the cell sorter without gating, and used as an unsorted control for RNA-sequencing. All samples were sorted directly into RLT buffer (Qiagen) with β-mercaptoethanol (Sigma Aldrich) and frozen at -80˚C. Frozen samples were thawed and RNA isolated using an RNeasy Micro kit (Qiagen).

### RNA sequencing

Library preparation and RNA sequencing was carried out by VGTI-FL (Florida, USA). Purified RNA was used to make libraries using an Illumina TruSeq kit. Libraries were sequenced to a depth of 37-63 M read pairs on an Illumina HiSeq platform. Paired end reads were processed on a Cardiff University pipeline. Reads were trimmed with Trimmomatic (28) and assessed for quality using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) using the default parameters. Reads were mapped to Ensembl human genome build GRCh38.89 downloaded from the Ensembl FTP site (http://www.ensembl.org/info/data/ftp/index.html/) using STAR (29). Counts were assigned to transcripts using featureCounts with the GRCh38.89 gene build GTF (30). RNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-8803.

### Differential expression analysis

Aligned reads were normalised using DESeq2 in R (31). Differentially expressed genes were identified between purified tumor samples, purified near, and purified far epithelium. Differential expression analysis was carried out using DESeq2 between sample types for all donors in a paired analysis. Comparisons of tumor and near or far normal epithelial tissue were carried out. For the three-donor expression analysis, genes with a Log 2-fold change greater than 3.5, FPKM values in healthy tissue less than 3.5, and FPKM values in tumor greater than 4.0 in any two of three donors and p-adjusted < 0.05 (Benjamini and Hochberg (8), in three donors) were taken forward for further analysis.

The analysis was expanded to genes which were significantly differentially expressed in separate comparisons for data in two of the three donors. Higher expression cut off values were used with FPKM greater than 5.0 in both donor's tumor tissue, and less than 1.0 in healthy tissues, with a log 2-fold change greater than 6 and p-adjusted < 0.05.

### Analysis of TCGA RNA-seq data

Level 3 (raw counts, htseq.counts.gz) RNA-seq data, and sample meta data (GDC sample sheet and clinical cart files) were downloaded from the TCGA GDC portal (https://portal.gdc.cancer.gov/) on 15-May-2018 for colon and rectal adenocarcinoma patients (TCGA-COAD, TCGA-READ). All datasets were normalized as one matrix using DESeq2 in R and output normalized counts were used for all analyses, with data visualized using the pheatmap R package.

### Antigens

20mer peptides overlapping by 10 amino acids, covering the entire protein sequence of each identified TAA, were synthesized by Fmoc chemistry to >95% purity (GL Biochem), and divided into pools, as shown (Supplementary Tables 2-8). Individual 9mer peptides (to measure HLA-A*02-restricted DNAJB7-specific CD8$^+$ T cell responses) were synthesised by Fmoc chemistry to >90% purity (Peptide Synthetics). The recall antigens tuberculin purified protein derivative (PPD; Statens

Serum Institut) and hemagglutinin (HA; gift from Dr. John Skehel, National Institute of Medical Research) and the T cell mitogen phytohemagglutinin (PHA; Sigma) were used as positive controls. All antigens were used at a final concentration of 5μg/ml.

### Peripheral blood mononuclear cell (PBMC) culture

Peripheral blood samples were obtained from pre-operative colorectal adenocarcinoma patients (n=17), hepatocellular carcinoma patients (n=2), cholangiocarcinoma patient (n=1), head and neck carcinoma patients (n=2) and age-matched non-tumor-bearing donors (n=10). Blood samples were collected in 10ml heparin tubes (BD) no more than 7-days prior to surgery. PBMCs were isolated by centrifugation of heparinized blood over Lymphoprep (Axis-Shield). Cells were washed and re-suspended in CTL Test Plus media (CTL Europe), L-glutamine and penicillin / streptomycin. PBMC were plated in 96-well plates (Nunc) and cultured in duplicate wells with specific antigens for 14-days, supplemented with fresh media containing 20 IU/ml IL-2 on days 3, 7 and 10.

### Generation of DNAJB7 9mer epitope-specific CD8$^+$ T cell lines

PBMCs from two healthy donors and one CRC patient, who were known HLA-A*02 positive, were used to generate CD8$^+$ T cell lines. The HLA class I epitope prediction algorithms NetMHC 4.0 (32) and SYFPEITHI (33) were used to identify HLA-A*02-restricted DNAJB7 9mers predicted to bind with the highest affinity. PBMCs were stimulated with the top five scoring DNAJB7-derived 9mer epitopes in the presence of 40 IU/ml IL-2, 2 ng/ml IL-7, and 5 ng/ml IL-15, in CTL Test Plus culture media. Cells were incubated in 37°C, 5% $CO_2$ for 9-12 days before testing their specificity by IFN-γ ELISPOT or intracellular cytokine staining. CD8$^+$ T cells from positive lines were sorted by MojoSort CD8 T cell isolation kits (Biolegend, UK), and further expanded using irradiated T2 cells, loaded with relevant peptides, and irradiated autologous PBMCs as feeders. Cells were rested for a minimum of three days in media containing no cytokines before use in downstream assays.

### Adenovirus vectors and cell lines

The replication deficient (ΔE1/ΔE3) adenovirus vectors Ad5-DNAJB7 (expressing the entire DNAJB7 ORF (open reading frame)) and Ad5-EMPTY (lacking an inserted ORF) were generated in the AdZ vector system, using homologous recombineering methods as previously described (34,35). The DNAJB7 ORF was gene synthesized by GeneArt (ThermoFisher Scientific, UK), prior to being inserted into the AdZ vector. Viruses were produced in T-REx-293 and purified as previously described (36).

The colorectal tumor cell lines SW480 (European Collection of Authenticated Cell Cultures (ECACC) 87092801) and Caco-2 (ECACC 86010202) were used in this study, given their confirmed expression of HLA-A*02 and the coxsackie and adenovirus receptor (data not shown). Cell lines were transduced using 2500 virus particles (vp) per cell (SW480) or 10,000 vp/cell (Caco-2) of Ad5 vector expressing DNAJB7 under the control of a CMV promoter, or control empty Ad5 vector for 3 hours at 37°C 5% $CO_2$ in media containing FBS. After the incubation period, cells were washed with fresh media and cultivated according to culture collection protocols (ECACC), before use in downstream assays.  Ad5-GFP titration and subsequent flow cytometric analyses had been performed prior to Ad5-DNAJB7 transduction to determine the optimal vp/cell dose for each cell line. Vp/ml of each adenovirus was quantified using a Pierce Micro BCA Protein Assay Kit.

### Western blot

Protein was extracted from SW480 and Caco-2 cells using 4x NUPAGE LDS sample buffer (ThermoFisher Scientific). Protein (6-9μg) was resolved using 4-12% sodium dodecyl sulphate polyacrylamide electrophoresis gels at 120V for 90 minutes and transferred to PVDF membrane at 30V for 80 minutes at room temperature. Membranes were blocked using 5% milk in PBS plus 0.05% Tween 20 (PBST) and incubated with DNAJB7 antibody (1:1000, rabbit IgG, polyclonal, Bio-Techne) or β-actin antibody (1:2000, rabbit IgG, monoclonal, Sigma-Aldrich) overnight at 4°C. Antibodies were prepared in 0.05% milk in PBST. Secondary antibody (1:3000, Anti-rabbit IgG, Bio-Rad) was applied for 1-hour at room temperature. Images of the bands were visualized using SuperSignal Pico PLUS chemiluminescent substrate (ThermoFisher Scientific).

**Real-time cytotoxicity assay (xCelligence)**

Target SW480 or Caco-2 cells were harvested and plated at 20,000 and 12,000 cells per well, respectively, in a 96-well xCelligence E-plate (ACEA Biosciences). Transduction was performed 48 hours prior to plating where appropriate. Suitable cell densities were determined by previous titration experiments. Cell attachment was monitored using the xCelligence Real-Time Cell Analysis (RTCA) instrument until the plateau phase was reached. DNAJB7-specific $CD8^+$ T cell lines were added at an effector to target ratio of 5:1 and impedance measurements performed every 10 minutes for up to 72 hours. All experiments were performed in duplicate. Changes in electrical impedance were expressed as a dimensionless cell index value, normalized to impedance values immediately preceding the addition of effector T cells.

**ELISpot / FluoroSpot Assays**

IFN-γ ELISpot and IFN-γ/Granzyme B FluoroSpot assays were performed as previously described (13). Briefly, PVDF 96-well filtration plates were coated with 50μl antibody (Mabtech). Cells were washed, plated, and stimulated with 5μg/ml antigen in duplicate wells. Plates were incubated at 37$^{o}$C, 5% $CO_2$ for 24-hours before developing spots. Spot-forming cells (SFC), i.e. cytokine-producing T cells, were enumerated using SmartCount settings on an automated plate reader (ImmunoSpot S6 Ultra; CTL Europe). Positive responses were identified as having at least 20 SFC/$10^5$ cultured PBMCs, and at least double that of the negative (no antigen) control. Wells with spot counts >1000 were deemed too numerous to count and capped at this level.

**Flow cytometry**

To perform T cell counts, 15µl of human TBNK 6-colour cocktail (BioLegend) was added to 50µl of whole heparinized blood using a reverse pipetting technique. Red blood cells were lyzed and samples run on a NovoCyte 3000 (ACEA Biosciences) to obtain absolute cell counts. To calculate the proportion of proliferating $CD4^+$ regulatory T cells, fresh PBMCs were stained with Live/Dead-Aqua (ThermoFisher Scientific), surface stained with CD3-FITC, CD4-BV605 and CD25-BV421 (BioLegend), followed by fixation / permeabilization and intracellular staining with Foxp3-APC (ThermoFisher Scientific) and Ki67-PE (BD Biosciences).

**Immunohistochemistry**

The identified TAAs from this study were evaluated for protein expression characteristics on healthy tissue and a range of tumor samples by utilising the Human Protein Atlas resource (12). In addition, DNAJB7 expression was assessed on formalin-fixed paraffin embedded blocks of colorectal tumor and healthy colon tissue (see Supplementary Table 1 for patient characteristics), and testis tissue as a positive control for DNAJB7 expression. Immunohistochemistry was performed on the Leica Bond RX Automated Research Stainer. Dewaxing/hydration of 5μm sections was performed according to manufacturer's instructions (Leica). Antigen retrieval was performed using Bond Epitope Retrieval Solution 2. DNAJB7 antibody (HPA000534, Atlas Antibodies) was used at a dilution of 1:100 and incubated for 105 minutes. Antibody detection was performed using Bond Polymer Refine Detection Kit, followed by hematoxylin counter staining. Following this, samples were dehydrated, mounted then scanned using Slide Scanner Axio Scan.Z1 (Zeiss); representative images were taken using Zen Blue software.

## Results

### Purification of samples prior to RNA-sequencing provided enhanced resolution of differentially expressed genes

Rectal tumor and paired background (unaffected) colon specimens were obtained from three patients undergoing resection. Autologous colon samples were cut from macroscopically normal sections of the excised tissue, both “near” (within 2 cm) and “far” (at least 10 cm) from the tumor site (Figure 1A). Dissociated single cells were sorted into Live/Dead$^{-}$EpCAM$^{+}$CD3$^{-}$ populations (Figure 1B and C). EpCAM was chosen as it would enable preferential isolation of epithelial populations over stromal tissue and immune populations (6,7).

RNA-sequencing datasets were comparable following several normalization procedures. Differential expression comparisons were run using DESeq2 of healthy tissues (“near” and “far”) against purified tumor tissue in all three patients, and then separate analyses for each combination of two patients. An additional comparison of non-purified tumor tissue against healthy tissues was run to investigate the impact of EpCAM sorting. To find relevant genes that could be targeted by immunotherapy, we applied criteria that specified very low levels of expression in healthy tissue combined with high expression in tumor tissue (based on FPKM and log 2-fold change). Only genes assigned a p-adjusted < 0.05 (Benjamini and Hochberg (8)) were taken forward for further analysis.

Initial gene lists gave 83 significant genes showing differential expression between tumor and far colon tissue, while 92 genes between tumor and near colon tissue. Cross referencing of these gene lists resulted in five genes that satisfied significant criteria in both comparisons (including four of those taken forward; ARSJ, CENPQ, ZC3H12B and CEACAM3). To expand our analysis, we looked at DEGs which were significantly expressed in tumor tissue of two of three patients to a higher level (increased expression cut-offs and lower threshold of healthy tissue expression). These gene lists were combined with three donor lists, and then near and far tissue cross referenced (Figure 2A). This gave an initial set of 54 genes which were cut to 23 based on levels of expression in healthy tissue of all three donors (Supplementary Figure 1 and Supplementary Table 9). Of these 23 genes, 18 were protein coding. We inspected these 18 genes and selected those which were most suitable for further analysis, eliminating those involved in the central nervous system, or which exhibited an inconsistent expression or read mapping profile in three donors gauged by visual curation of mapped reads in IGV (Integrative Genomics Viewer, Broad Institute) (9).

The final genes selected were DNAJB7, CENPQ, ZC3H12B, ZSWIM1, CEACAM3, ARSJ and CYP2B6, based on their ideal expression profile for therapeutic exploitation (Figure 2B). Inspection of expression profiles in non-purified tumor tissue (Figure 2B and 2D) exemplified the difficulty in detecting these genes in the absence of purification, with all expression levels lower than purified tissue.

To further assess the impact of tissue purification, we looked at the DEGs between bulk and EpCAM sorted tumor samples, focusing on genes which were expressed in at least two of the three samples (Figure 2C). This emphasised the advantage of purification resulting in enhancement of the most relevant gene expression patterns and demonstrates how our novel antigens could not have been identified from bulk tumor sequencing alone (Figure 2D).

### Comparison of Common Cancer Antigen Expression

We next wanted to assess the expression patterns of the seven novel antigens identified in the context of other antigens commonly classified in the literature as TAAs (10,11). We compiled a representative list which included antigens such as GP100, 5T4, LAGE3 and MART and looked at how their expression levels compared within the patient samples used in this study (Figure 3A) and across both colon and rectal tumor data from the TCGA (Figure 3B). This analysis showed a clear distinction in antigen expression level, with CYP2B6, ZSWIM1 and 5T4 expression comparable to the highly expressed LAGE3. However, beyond these four genes the expression levels of other antigens were heterogeneous across tumor samples, and some appeared to be relatively low in the TCGA data, in particular CEACAM3 and DNAJB7.

A critical criterion for the analysis in Figure 2 was differential expression of genes between healthy tissue and tumor tissue. The TCGA data also has several paired datasets of colon and rectal tumors and corresponding heathy tissue. We visually inspected the differences of the TAA panel list across these datasets in order to confirm that our antigen still accorded with this criterion in a large publicly available dataset (Figure 3C). Indeed, when healthy and tumor tissue expression of each were compared, the differentially expressed nature of several antigens was emphasised. For DNAJB7, ZSWIM1 and CENPQ, the contrast between healthy and tumor datasets was skewed towards tumor expression and suggested these would be better targets than ARSJ or CYP2B6 which did not show visual distinction between the two tissue types. Furthermore, other cancer antigens could be classified as having highly favorable (WT1, LAGE3, MART1, AFP) or hypothetically dangerous (ACRBP, SPA17, KLK3) expression patterns between healthy and tumor tissues, and as such their assessment as *bona fide* TAAs should be reconsidered. Such trends were also present in our data (Figure 3D).

### Analysis of protein expression across multiple healthy tissues highlights DNAJB7 as a cancer-testis antigen and a suitable target for immunotherapy

The protein expression level of each candidate TAA was evaluated using publicly available immunohistochemistry data (12). Whilst each candidate exhibited significant upregulation on tumor tissue over healthy tissue (with the possible exception of ARSJ), DNAJB7, a protein belonging to the evolutionarily conserved DNAJ heat shock family, was unexpectedly identified as a novel cancer-testis antigen given its complete lack of expression on any healthy tissue bar the testis, an immune-privileged site (Supplementary Figure 2).

We sought to corroborate this pattern of staining on paraffin-fixed samples acquired in-house, using the same anti-DNAJB7 antibody (HPA000534). In preliminary experiments, expression was higher in certain tumor samples, although antibody staining was observed in background colon tissue (Supplementary Figure 3). Given this finding and the failure to detect DNAJB7 mRNA in normal colon, we conclude that where DNAJB7 is detected, it is preferentially expressed in cancer tissue.

Furthermore, DNAJB7 was expressed on a very wide range of solid tumors, in particular on tumors of the gastrointestinal tract and accessory organs of digestion, including colorectal cancer and pancreatic ductal adenocarcinoma (Supplementary Figures 2B and 2C).

### DNAJB7 is a superior cancer-testis antigen

The expression profile of DNAJB7 was compared to six other well-defined cancer-testis antigens, including NY-ESO-1, MAGE-A1 and SSX2. High protein expression of all these antigens was confirmed to be confined to the testis (Supplementary Figure 4A). In comparison to the other cancer-testis antigens, DNAJB7 was expressed on the greatest range of tumor types, with more than 67% of all patients tested exhibiting positive (low, medium or high) protein expression on their tumor, except for lymphoma (Supplementary Figure 4B).

### Analysis of candidate TAA $T_H1$ responses reveal DNAJB7 to be immunogenic

Following identification of relevant genes and confirmed protein expression, we assessed their immunogenicity using overlapping peptide pools and culture with PBMC of CRC patients and healthy donors. Analysis of cultured PBMC by IFN-γ ELISpot determined three of the seven proteins to demonstrate immunogenicity in most donors tested (Figure 4A-B). As the size of peptide libraries was highly variable for each protein, we standardized the immunogenicity relevant to the number of peptides in each pool (Figure 4A). This analysis revealed CYP2B6, DNAJB7 and CEACAM3 to be comparably immunogenic across multiple individuals without stratification by HLA-type, and similarly immunogenic to the oncofetal antigen 5T4, a tumor antigen that has successfully been targeted in CRC previously (1,13). Conversely, CENPQ and ARSJ were poorly immunogenic in most donors tested.

Furthermore, our peptide pool design allowed us to interrogate immunogenicity based on a matrix format to determine the peptides responsible for the positive T cell responses (example for DNAJB7, Supplementary Figure 5). This type of analysis may be important for isolation of $T_H1$ stimulating regions of TAA which can be incorporated in vaccines based on immunogenic components of multiple antigens important in CRC, as well as being regions that can be targeted by epitope-based modifications and strategies for enhancement of the immune response (14). An example of one CRC patient revealed positive IFN-$\gamma$ and granzyme B responses to DNAJB7 peptide pools 3, 6 and 10, indicative of T cell responses to epitopes contained within peptides 3 and 23 (Supplementary Figure 5B and C). Indeed, peptide 23 was the most immunogenic region of the DNAJB7 protein, with responses discovered in 39% of CRC patient and healthy control donors tested (Supplementary Figure 5D). DNAJB7 was also found to be immunogenic in patients with other tumor types, including hepatocellular carcinoma, cholangiocarcinoma and the non-gastrointestinal head and neck squamous cell cancer (Figure 4C).

## $CD8^+$ T cell recognition of DNAJB7-expressing colorectal tumor cell lines

$T_H1$ responses described above were dominated by IFN-$\gamma$-secreting $CD4^+$ effector T cells, favoured by the longer 20mer peptides used to stimulate the PBMCs. However, analysis of specific IFN-$\gamma$ and granzyme B production using FluoroSpot assays indicates that granzyme B is abundantly produced in response to the DNAJB7 peptides (9/10 donors tested, Figure 5A and Supplementary Figure 5) suggesting cytotoxic T cell responses are present. In order to ascertain whether these responses were indicative of $CD8^+$ T cells capable of killing DNAJB7-expressing tumor cells, HLA-A*02-restricted $CD8^+$ T cell epitopes derived from DNAJB7 were identified by computer-based epitope prediction algorithms (Figure 5B). The top five scoring peptides stimulated cognate $CD8^+$ T cells derived from HLA-A*$02^+$ healthy donors and a CRC patient. DNAJB7-specific $CD8^+$ T cell lines were successfully enriched in all donors, an example of responses to peptides LTFFLVNSV and GMDNYISVT is shown (Figure 5C).

The HLA-A*02-expressing SW480 and Caco-2 colorectal tumor cell lines were used as targets in a cytotoxicity assay, however there was minimal expression of DNAJB7 in both lines (Figure 5D and Supplementary Figure 6). Cell lines were successfully transduced with an Ad5-DNAJB7 viral vector and found to stably increase DNAJB7 expression over a 5-day period before reducing on day 6 (Figure 5D and Supplementary Figure 6). Upon addition of effector DNAJB7-specific $CD8^+$ T cells to target colorectal tumor cells, real-time impedance traces show a highly significant ($P<0.0001$) reduction in the number, size/shape, and/or attachment quality of DNAJB7-expressing Caco-2 cells (Figure 5E), and a reduction in the growth of DNAJB7-expressing SW480 cells (Figure 5F) over Ad5-EMPTY transduced or non-transduced, untreated (UT) tumor cell lines. Non-specific (DNAJB7-negative) T cell lines did not cause additional killing to DNAJB7-expressing targets, compared to Ad5-EMPTY or UT cell lines (data not shown). Hence, DNAJB7-expressing tumor cell lines present peptides on the cell surface by MHC class I and are selectively eliminated by DNAJB7-specific CD8$^+$ T cells.

### Anti-DNAJB7 T$_H$1 responses are induced during cyclophosphamide treatment

We have previously demonstrated that anti-tumor T$_H$1 effector responses are controlled by regulatory T cells (Tregs) (15), and that targeting these Tregs either by depletion in vitro, or inhibition/depletion in vivo with low dose cyclophosphamide, increases the anti-tumor (5T4) immune response (1,13). We sought to assess whether T cell responses were induced to the novel tumor antigens in a CRC patient (Figure 6) and an HCC patient (Supplementary Figure 7) receiving short-term metronomic cyclophosphamide. Anti-5T4 T$_H$1 responses increased by >4-fold in both patients, an effect previously identified as associating with improved survival outcomes (13); intriguingly, anti-DNAJB7 T$_H$1 responses also mirrored this treatment response profile in both instances, whereas no responses were induced to ARSJ, CENPQ, ZSWIM1 and CYP2B6 (Figure 6B and C, and Supplementary Figure 7A and B). This could suggest that responses to DNAJB7 and CEACAM3 are suppressed in CRC and HCC, given that responses were unmasked by efficient regulatory T cell depletion (Figure 6D and Supplementary Figure 7C).

## Discussion

The pursuit of new cancer vaccines that can be administered regardless of patient HLA-type or neoantigen load relies on the investigation and discovery of novel TAAs. Paired RNA-sequencing of tumor and healthy tissue facilitates TAA identification but is limited by the diversity of cellular input in each sequencing sample. Here we purified EpCAM$^+$ cellular populations from healthy colon and primary colorectal tumors; RNA-sequencing data from purified samples revealed multiple genes that showed significant differential expression across three donors.

In a comparison with non-purified tumor tissue, no genes were classified as significant according to the same criteria, demonstrating the power of using purified tissues in antigen discovery. Further analyses of differentially expressed gene lists in tissues near and far from the tumor helped identify 18 genes which showed differential expression patterns suitable for therapeutics. Four protein products of the identified genes exhibited significant immunogenicity in healthy donors and cancer patients; of these DNAJB7 also demonstrated the most favorable expression profile based on immunohistochemistry data of healthy and cancerous tissue. DNAJB7 belongs to the evolutionarily conserved DNAJ/Heat Shock Protein (HSP)40 family of proteins and is a molecular chaperone to HSP40. It is likely that its upregulation in tumors is in response to increased expression of many heat shock proteins, aiding tumor cell proliferation in hostile environments (16). Indeed, other HSP40 family members DNAJB6 and DNAJB8 have both been previously shown to be upregulated in cancer, contributing to cancer-initiating cell maintenance (17-19). Therapies targeting heat shock proteins and their molecular chaperones are already showing promise in cancer treatment (20). Immunotherapeutic targeting of these proteins may also yield further anti-cancer benefits, as implicated by this study.

Robust $T_H1$ responses to DNAJB7 and several other novel tumor antigens were found in healthy controls, in keeping with previous findings for tumor-associated antigens from our laboratory (21,22) and others (23,24). It is possible these responses are indicative of a normal functioning process of immunosurveillance to remove aberrant epithelial cells. The presence or absence of such responses are now beginning to be exploited for cancer diagnostics and prognostication ((24) and NCT02840058). How and why these T cells exist and are maintained at such a frequency in the memory pool remains unknown: possibilities range from transient upregulation of TAAs during periods of inflammation, e.g. of the colon (22), incomplete thymic selection or antigenic cross-reactivity / mimicry to microbial proteins (25).

There are limitations to our study, including the selection of luminal tumor sites for cell enrichment as opposed to the invasive margin (required for histopathological assessment of the tumor), the use of purification procedures, i.e. fluorescence activated cell sorting, that may influence mRNA expression prior to RNA isolation, and low initial sample size. However, despite the relatively small scale, the approach described here has successfully identified novel, highly antigenic proteins expressed in cancers. These antigens could be incorporated into vaccines for both therapeutic and prophylactic use. One goal of cancer vaccination in the context of CRC immunotherapy is to reduce relapse rates following surgical intervention. Curative rates following resection of primary colorectal tumors are ~60-70% but could be improved if relapse was prevented by safely boosting immunity to the proteins with differential expression patterns as a form of prophylactic immunotherapy (26). Indeed, loss of anti-tumor immune responses associates with advancing tumor stage (21,22), and these patients can benefit from anti-cancer vaccination strategies. At the moment, although there is some success using a single TAA in CRC (1), better therapeutic strategies are necessary with superior vaccine targets combined with manipulations of immune regulation. These approaches necessitate discovery of more TAAs and greater investigation of the negative impacts of T cell cross-reactivity and off-target immune effects. Questions over the ideal differential expression pattern, specifically the extent to which some expression in healthy tissue can be tolerated relative to tumor expression are highly relevant. In addition, the targeting of multiple tumor antigens is more likely to overcome inherent tumor immune evasion and evolution. However, cancer-testis antigens allay some of these concerns and represent an ideal tumor target for immunotherapy (27). Indeed, here we identified that most donors have the capability to mount anti-DNAJB7 T cell responses, and these responses can be significantly boosted in cancer patients receiving cyclophosphamide. Enhancing this anti-tumor immune response could hold significant potential in future therapeutic and prophylactic treatment strategies.

## Acknowledgments

This study was supported by a Cancer Research Wales programme grant (to A. Godkin and A. Gallimore), a Wellcome Trust Collaborator Award Grant (209213/Z/17/Z) (to A. Godkin, A. Gallimore, R. Stanton), a Cancer Research UK programme grant (C16731/A21200) (to A. Gallimore and A. Godkin) and funding from the Wales Cancer Research Centre (to S. Burnell). We are grateful to Prof Richard Houlston for useful preliminary discussion on experimental design. We thank Dr Catherine Naseriyan for assistance with flow cytometry and cell sorting, Anna Fuller for provision of peptides, Dr Mat Clement for assistance with western blots and Dr Ruban Rex Peter Durairaj & Owen Moon for assistance with xCelligence.

## Authors' Contributions

**Conception and design:** M. Scurr, A. Gallimore and A. Godkin
**Development of methodology:** M. Scurr, A. Greenshields-Watson, K. Ashelford, R. Andrews, A. Parker, R. Stanton and A. Godkin
**Acquisition of data:** M. Scurr, A. Greenshields-Watson, E. Campbell, M. Somerville and Y. Chen.
**Analysis and interpretation of data:** M. Scurr, A. Greenshields-Watson, E. Campbell, M. Somerville, Y. Chen, K. Ashelford, R. Andrews, A. Gallimore and A. Godkin
**Writing, review, and/or revision of the manuscript:** M. Scurr, A. Greenshields-Watson, M. Somerville, Y. Chen, S. Burnell and A. Godkin
**Administrative, technical, or material support:** M. Somerville, S. Hulin-Curtis, S. Burnell, J. Davies, M. Davies, R. Hargest, S. Phillips, A. Christian, K. Ashelford, R. Andrews, A. Parker and R. Stanton.

## References

1. Scurr M, Pembroke T, Bloom A, Roberts D, Thomson A, Smart K*, et al.* Effect of Modified Vaccinia Ankara-5T4 and Low-Dose Cyclophosphamide on Antitumor Immunity in Metastatic Colorectal Cancer: A Randomized Clinical Trial. JAMA Oncol **2017**;3(10):e172579 doi 10.1001/jamaoncol.2017.2579.
2. Parkhurst MR, Yang JC, Langan RC, Dudley ME, Nathan DA, Feldman SA*, et al.* T cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. Mol Ther **2011**;19(3):620-6 doi 10.1038/mt.2010.272.
3. Scurr MJ, Brown CM, Costa Bento DF, Betts GJ, Rees BI, Hills RK*, et al.* Assessing the prognostic value of preoperative carcinoembryonic antigen-specific T-cell responses in colorectal cancer. J Natl Cancer Inst **2015**;107(4) doi 10.1093/jnci/djv001.
4. Raman MC, Rizkallah PJ, Simmons R, Donnellan Z, Dukes J, Bossi G*, et al.* Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. Sci Rep **2016**;6:18851 doi 10.1038/srep18851.
5. Hilf N, Kuttruff-Coqui S, Frenzel K, Bukur V, Stevanović S, Gouttefangeas C*, et al.* Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature **2019**;565(7738):240-5 doi 10.1038/s41586-018-0810-y.
6. Martowicz A, Seeber A, Untergasser G. The role of EpCAM in physiology and pathology of the epithelium. Histol Histopathol **2016**;31(4):349-55 doi 10.14670/HH-11-678.
7. Schnell U, Cirulli V, Giepmans BN. EpCAM: structure and function in health and disease. Biochim Biophys Acta **2013**;1828(8):1989-2001 doi 10.1016/j.bbamem.2013.04.018.
8. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics **2003**;19(3):368-75.
9. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G*, et al.* Integrative genomics viewer. Nat Biotechnol **2011**;29(1):24-6 doi 10.1038/nbt.1754.
10. Butterfield LH. Cancer vaccines. BMJ **2015**;350:h988 doi 10.1136/bmj.h988.
11. Garcia-Soto AE, Schreiber T, Strbo N, Ganjei-Azar P, Miao F, Koru-Sengul T*, et al.* Cancer-testis antigen expression is shared between epithelial ovarian cancer tumors. Gynecol Oncol **2017**;145(3):413-9 doi 10.1016/j.ygyno.2017.03.512.
12. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A*, et al.* Proteomics. Tissue-based map of the human proteome. Science **2015**;347(6220):1260419 doi 10.1126/science.1260419.
13. Scurr M, Pembroke T, Bloom A, Roberts D, Thomson A, Smart K*, et al.* Low-Dose Cyclophosphamide Induces Antitumor T-Cell Responses, which Associate with Survival in Metastatic Colorectal Cancer. Clin Cancer Res **2017**;23(22):6771-80 doi 10.1158/1078-0432.CCR-17-0895.
14. Cole DK, Gallagher K, Lemercier B, Holland CJ, Junaid S, Hindley JP*, et al.* Modification of the carboxy-terminal flanking region of a universal influenza epitope alters CD4$^+$ T-cell repertoire selection. Nat Commun **2012**;3:665 doi 10.1038/ncomms1665.
15. Betts G, Jones E, Junaid S, El-Shanawany T, Scurr M, Mizen P*, et al.* Suppression of tumour-specific CD4$^+$ T cells by regulatory T cells is associated with progression of human colorectal cancer. Gut **2012**;61(8):1163-71 doi 10.1136/gutjnl-2011-300970.
16. Mitra A, Shevde LA, Samant RS. Multi-faceted role of HSP40 in cancer. Clin Exp Metastasis **2009**;26(6):559-67 doi 10.1007/s10585-009-9255-x.
17. Kusumoto H, Hirohashi Y, Nishizawa S, Yamashita M, Yasuda K, Murai A*, et al.* Cellular stress induces cancer stem-like cells through expression of DNAJB8 by activation of heat shock factor 1. Cancer Sci **2018**;109(3):741-50 doi 10.1111/cas.13501.

18. Morita R, Nishizawa S, Torigoe T, Takahashi A, Tamura Y, Tsukahara T*, et al.* Heat shock protein DNAJB8 is a novel target for immunotherapy of colon cancer-initiating cells. Cancer Sci **2014**;105(4):389-95 doi 10.1111/cas.12362.
19. Meng E, Shevde LA, Samant RS. Emerging roles and underlying molecular mechanisms of DNAJB6 in cancer. Oncotarget **2016**;7(33):53984-96 doi 10.18632/oncotarget.9803.
20. Chatterjee S, Burns TF. Targeting Heat Shock Proteins in Cancer: A Promising Therapeutic Approach. Int J Mol Sci **2017**;18(9) doi 10.3390/ijms18091978.
21. Besneux M, Greenshields-Watson A, Scurr MJ, MacLachlan BJ, Christian A, Davies MM*, et al.* The nature of the human T cell response to the cancer antigen 5T4 is determined by the balance of regulatory and inflammatory T cells of the same antigen-specificity: implications for vaccine design. Cancer Immunol Immunother **2018** doi 10.1007/s00262-018-2266-1.
22. Scurr M, Bloom A, Pembroke T, Srinivasan R, Brown C, Smart K*, et al.* Escalating regulation of 5T4-specific IFN-γ(+) CD4(+) T cells distinguishes colorectal cancer patients from healthy controls and provides a target for in vivo therapy. Cancer Immunol Res **2013**;1(6) doi 10.1158/2326-6066.CIR-13-0035.
23. Costa-Nunes C, Cachot A, Bobisse S, Arnaud M, Genolet R, Baumgaertner P*, et al.* High-throughput Screening of Human Tumor Antigen-specific CD4 T Cells, Including Neoantigen-reactive T Cells. Clin Cancer Res **2019**;25(14):4320-31 doi 10.1158/1078-0432.CCR-18-1356.
24. Laheurte C, Dosset M, Vernerey D, Boullerot L, Gaugler B, Gravelin E*, et al.* Distinct prognostic value of circulating anti-telomerase CD4+ Th1 immunity and exhausted PD-1+/TIM-3+ T cells in lung cancer. Br J Cancer **2019** doi 10.1038/s41416-019-0531-5.
25. Zitvogel L, Ayyoub M, Routy B, Kroemer G. Microbiome and Anticancer Immunosurveillance. Cell **2016**;165(2):276-87 doi 10.1016/j.cell.2016.03.001.
26. Finn OJ. The dawn of vaccines for cancer prevention. Nat Rev Immunol **2018**;18(3):183-94 doi 10.1038/nri.2017.140.
27. Gjerstorff MF, Andersen MH, Ditzel HJ. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. Oncotarget **2015**;6(18):15772-87 doi 10.18632/oncotarget.4694.
28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **2014**;30(15):2114-20 doi 10.1093/bioinformatics/btu170.
29. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**;29(1):15-21 doi 10.1093/bioinformatics/bts635.
30. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **2014**;30(7):923-30 doi 10.1093/bioinformatics/btt656.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **2014**;15(12):550 doi 10.1186/s13059-014-0550-8.
32. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics **2016**;32(4):511-7 doi 10.1093/bioinformatics/btv639.
33. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics **1999**;50(3-4):213-9 doi 10.1007/s002510050595.
34. Stanton RJ, McSharry BP, Armstrong M, Tomasec P, Wilkinson GW. Re-engineering adenovirus vector systems to enable high-throughput analyses of gene function. Biotechniques **2008**;45(6):659-62, 64-8 doi 10.2144/000112993.
35. Uusi-Kerttula H, Davies J, Coughlan L, Hulin-Curtis S, Jones R, Hanna L*, et al.* Pseudotyped αvβ6 integrin-targeted adenovirus vectors for ovarian cancer therapies. Oncotarget **2016**;7(19):27926-37 doi 10.18632/oncotarget.8545.

36. Hulin-Curtis SL, Davies JA, Nestić D, Bates EA, Baker AT, Cunliffe TG*, et al.* Identification of folate receptor α (FRα) binding oligopeptides and their evaluation for targeted virotherapy applications. Cancer Gene Ther **2020** doi 10.1038/s41417-019-0156-0.

## Figure Legends

**Figure 1. Isolation of epithelial and tumor cells by EpCAM-sorting prior to RNA-seq.** (A) Schematic of tumor and healthy tissue resection taken at two distances from the tumor site. Samples were taken from rectal tumor of three patients. (B) Sample processing and purification flow chart. (C) Flow cytometry gating for EpCAM$^+$ and CD3$^-$ purification, pre and post cell sorting.

**Figure 2. Identification of candidates for further investigation based on differential expression analysis.** (A) Workflow for obtaining gene lists of differentially expressed genes based on two comparisons (purified tumor versus purified healthy colon "far," and purified tumor versus purified healthy colon "near," left-hand side and right-hand side, respectively). Gene lists were obtained from significantly differentially expressed genes across all three patients and separately in two of three patients. These were aligned and cross referenced between "far" and "near" comparisons to give a smaller gene list which was further reduced based on expression in healthy tissue, and finally suitability for further investigation. (B) Normalized counts for each of the seven genes selected for further analysis. Counts are shown for each of the four conditions as box plots representing all three patients. (C) Heatmap showing 317 genes that were differentially expressed between EpCAM purified and bulk tumor samples and were expressed in at least two of three EpCAM purified samples. Normalized counts were scaled by gene to show relative expression between each sample. (D) Heatmap of novel tumor antigen gene expression (in addition to 5T4), showing differences in normalized counts between EpCAM purified and bulk tumor, scaled as part (C).

**Figure 3.** (A) Heatmap showing relative expression levels of known and novel TAAs using normalized counts scaled by sample across EpCAM purified tumor. (B) Corresponding analysis using TCGA data for rectal (green bar) and colon (purple bar) tumors scaled by sample. (C) Comparison of expression levels between tumor and solid tissue normal in available TCGA data, scaled by gene. Sample type and tissue/tumor location (colon or rectal) are indicated in the top two bars. (D) Corresponding analysis performed on our EpCAM purified tumor and healthy data.

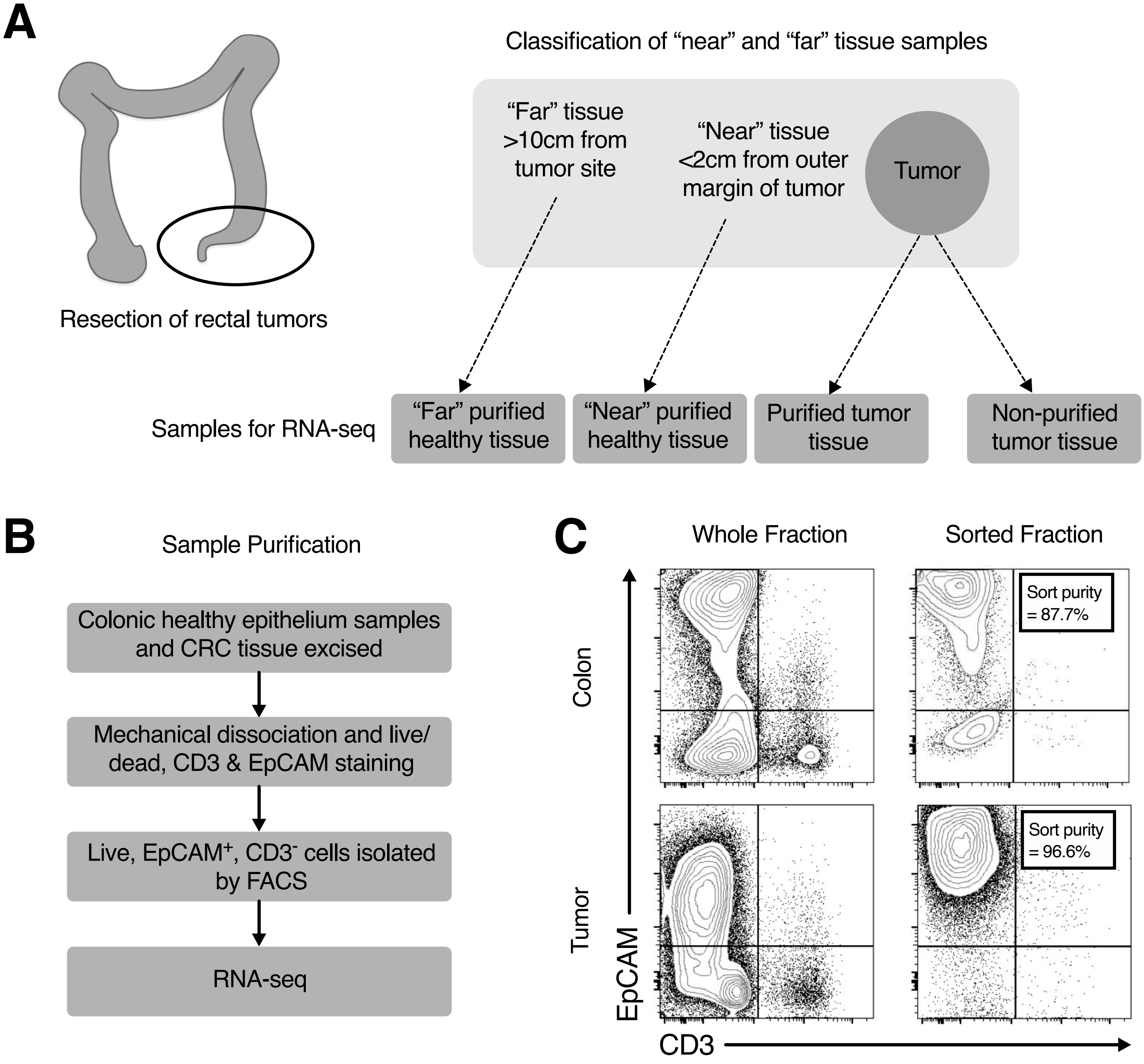
**Figure 4. Immunogenicity of candidate TAAs.** T cell responses to peptide pools spanning the entire protein sequence of each candidate TAA were assessed by cultured IFN-$\gamma$ ELISpot (see Supplementary Tables 2-8 for peptide sequences). The total number of IFN-$\gamma^+$ spot-forming cells (SFC) per $10^5$ cultured PBMC relative to the number of peptides spanning the protein was assessed and ranked by mean response (grey bars) amongst all donors tested (A) and then subdivided by CRC patients (blue circles TNM Stage 1/2; n=6, black circles TNM Stage 3; n=8) and healthy donors (‘HD’, white circles; n=10) (B). (C) Patients with other gastrointestinal cancers were tested for their ability to mount anti-DNAJB7 T cell responses (CC – cholangiocarcinoma; HCC – hepatocellular carcinoma; HNC – head and neck squamous cell carcinoma).

**Figure 5. Enriched DNAJB7-specific CD8$^+$ T cells target DNAJB7-expressing colorectal tumor cell lines.** T cell responses to two peptide pools spanning the entire DNAJB7 protein sequence were assessed by cultured Granzyme B FluoroSpot (see Supplementary Table 2 for peptide sequences). The total number of Granzyme B$^+$ spot-forming cells (SFC) per $10^5$ cultured PBMC relative to the 30 peptides spanning the DNAJB7 protein was assessed in 9 CRC patients (A). (B) HLA class I epitope prediction algorithms were used to identify HLA-A*02-restricted DNAJB7 9mers predicted to bind with the highest affinity; top 5 across the algorithms are indicated. (C) DNAJB7-specific CD8$^+$ T cells were enriched in multiple donors, a representative example of the IFN-$\gamma$ response in one T cell line to DNAJB7 epitopes LTFFLVNSV and GMDNYISVT is shown. (D) The SW480 CRC cell line was transduced with Ad5-DNAJB7 or an Ad5-

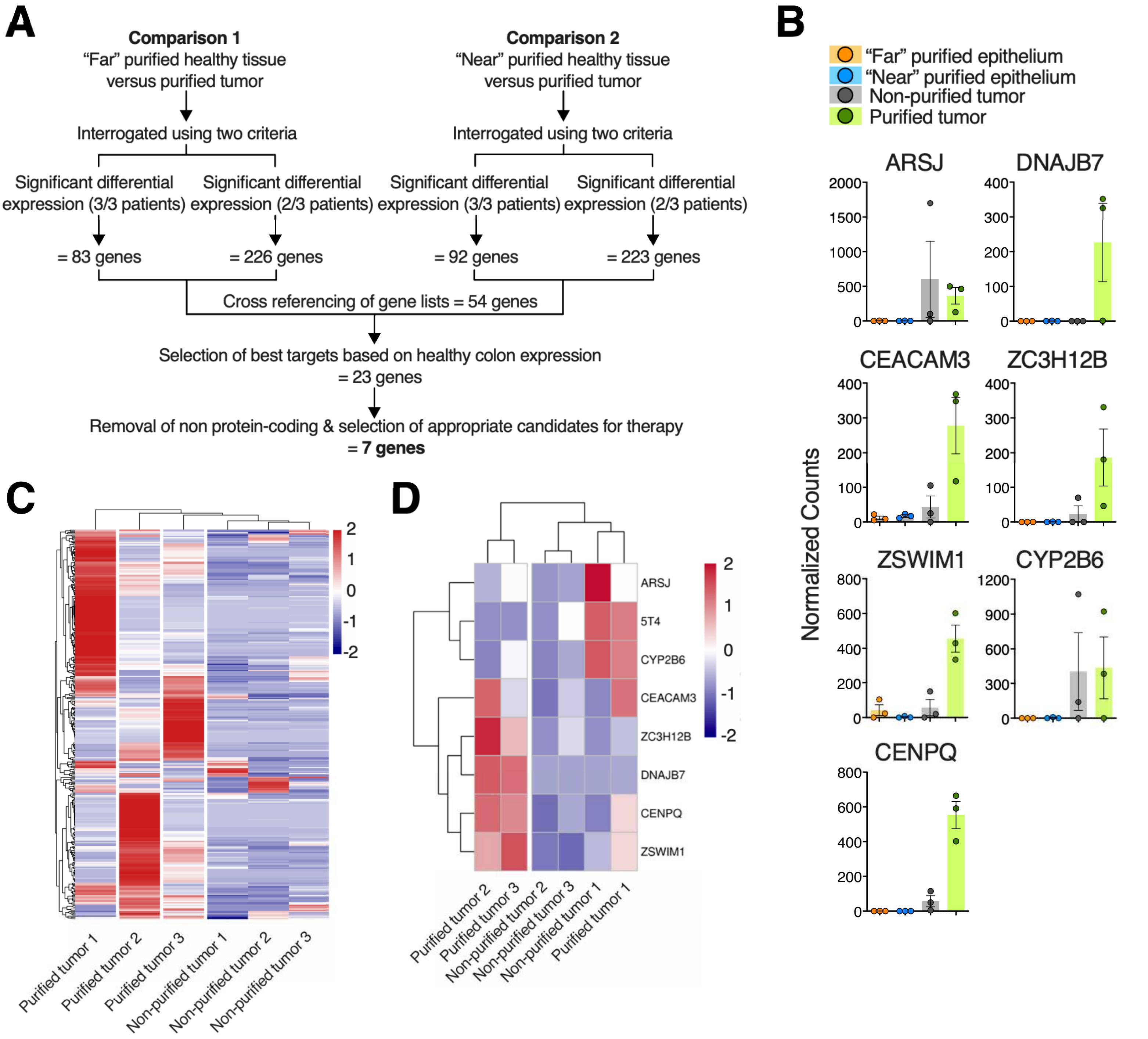
EMPTY vector, with the expression of DNAJB7 protein indicated by the band at 35kDa and actin control at 45kDa. UT = untreated (non-transduced) SW480 cells. DNAJB7-specific T cell lines from a healthy donor ('Donor 1') and a CRC patient ('Donor 2') were seeded into 96-well E-plates, co-incubated with the indicated transduced / non-transduced Caco-2 (E) or SW480 (F) cell lines at an effector to target ratio of 5:1. Changes in impedance over a 24-hour period, normalized at the timepoint immediately preceding the addition of effector T cells, are given as a dimensionless normalized cell index. Experiments were performed in duplicates. Statistical results of two-way ANOVA are indicated (*** P<0.0001).

**Figure 6. Regulatory T cell depletion unmasks $T_H1$ responses to novel TAAs.** (A) A post-colectomy CRC patient received low-dose, metronomic cyclophosphamide on treatment days 1-8 and 15-22, with blood samples collected weekly throughout treatment. T cell responses to peptide pools spanning the entire protein sequence of each candidate TAA were assessed by cultured IFN-$\gamma$ ELISpot at each timepoint; example images of IFN-$\gamma$ ELISpot wells are shown (B). (C) The total number of IFN-$\gamma^+$ spot-forming cells (SFC) per $10^5$ cultured PBMC (mean of duplicate wells) were calculated for each TAA. (D) $CD3^+CD4^+CD25^{hi}Foxp3^+$ regulatory T cell numbers and $\%Ki67^+$ Tregs were measured by flow cytometry during cyclophosphamide treatment.

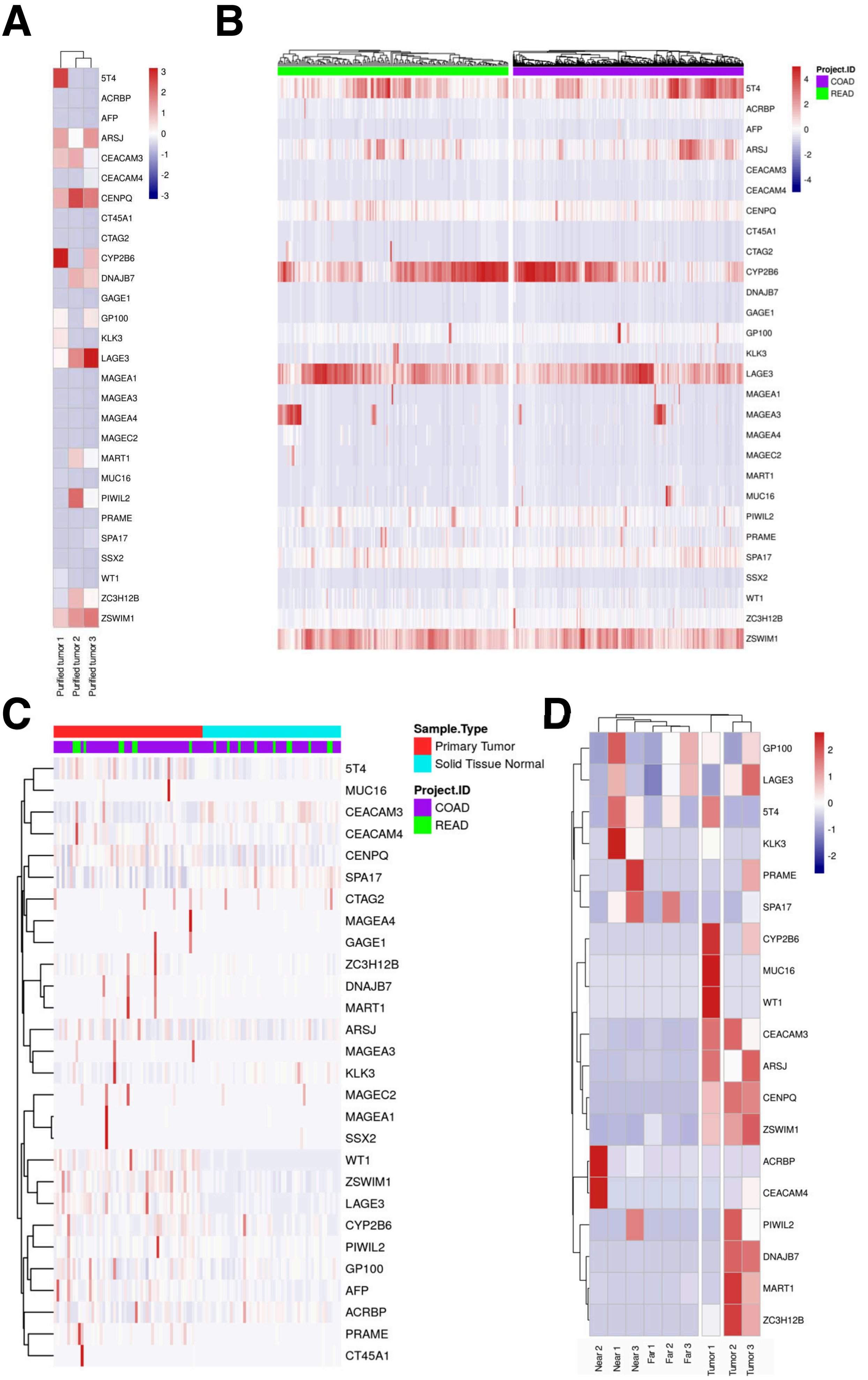
Figure 1


A
Resection of rectal tumors
Classification of “near” and “far” tissue samples
“Far” tissue >10cm from tumor site
“Near” tissue <2cm from outer margin of tumor
Tumor
Samples for RNA-seq
“Far” purified healthy tissue
“Near” purified healthy tissue
Purified tumor tissue
Non-purified tumor tissue
B
Sample Purification
Colonic healthy epithelium samples and CRC tissue excised
Mechanical dissociation and live/dead, CD3 & EpCAM staining
Live, EpCAM+, CD3- cells isolated by FACS
RNA-seq
C
Whole Fraction
Sorted Fraction
Colon
Tumor
Sort purity = 87.7%
Sort purity = 96.6%
EpCAM
CD3

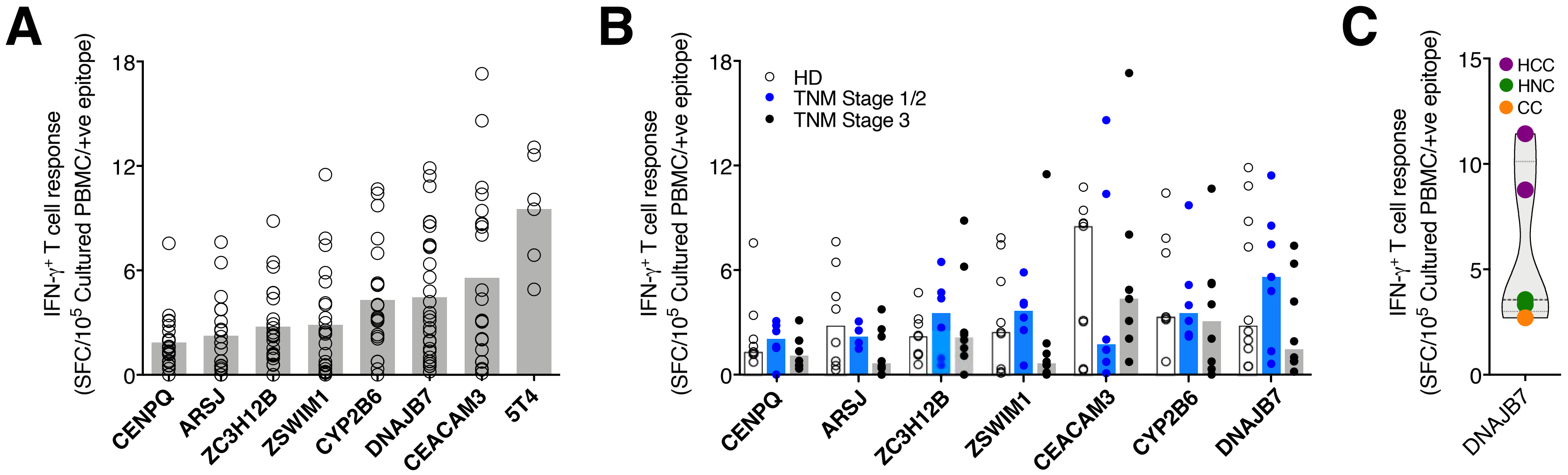
# Figure 2


A
Comparison 1
"Far" purified healthy tissue versus purified tumor
Comparison 2
"Near" purified healthy tissue versus purified tumor
Interrogated using two criteria
Interrogated using two criteria
Significant differential expression (3/3 patients)
Significant differential expression (2/3 patients)
Significant differential expression (3/3 patients)
Significant differential expression (2/3 patients)
= 83 genes
= 226 genes
= 92 genes
= 223 genes
Cross referencing of gene lists = 54 genes
Selection of best targets based on healthy colon expression = 23 genes
Removal of non protein-coding & selection of appropriate candidates for therapy = 7 genes
B
"Far" purified epithelium
"Near" purified epithelium
Non-purified tumor
Purified tumor
Normalized Counts
ARSJ
0
500
1000
1500
2000
DNAJB7
0
100
200
300
400
CEACAM3
0
100
200
300
400
ZC3H12B
0
100
200
300
400
ZSWIM1
0
200
400
600
800
CYP2B6
0
300
600
900
1200
CENPQ
0
200
400
600
800
C
2
1
0
-1
-2
Purified tumor 1
Purified tumor 2
Purified tumor 3
Non-purified tumor 1
Non-purified tumor 2
Non-purified tumor 3
D
ARSJ
5T4
CYP2B6
CEACAM3
ZC3H12B
DNAJB7
CENPQ
ZSWIM1
2
1
0
-1
-2
Purified tumor 2
Purified tumor 3
Non-purified tumor 2
Non-purified tumor 3
Non-purified tumor 1
Purified tumor 1

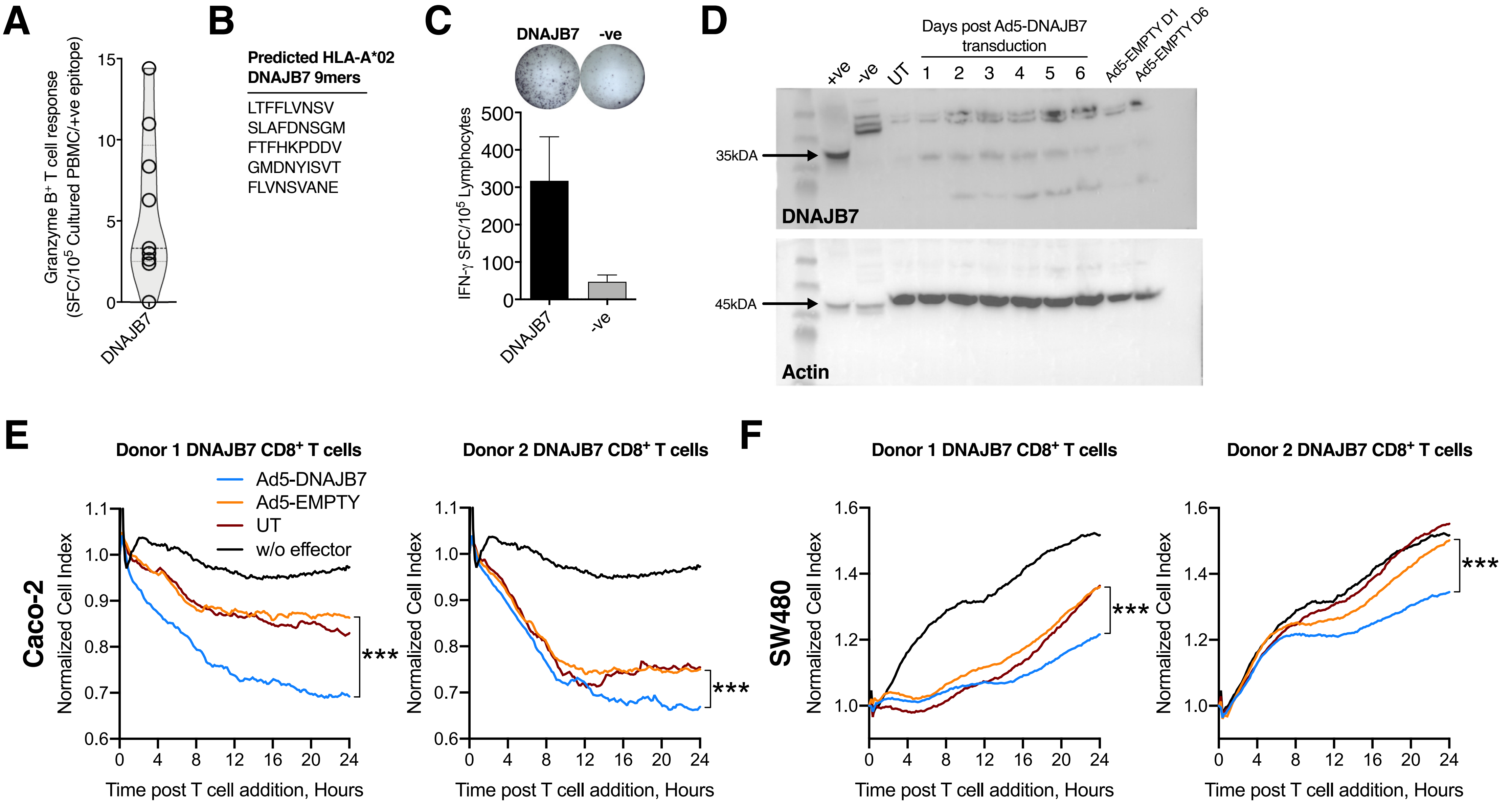
# Figure 3


A
B
C
D
5T4
ACRBP
AFP
ARSJ
CEACAM3
CEACAM4
CENPQ
CT45A1
CTAG2
CYP2B6
DNAJB7
GAGE1
GP100
KLK3
LAGE3
MAGEA1
MAGEA3
MAGEA4
MAGEC2
MART1
MUC16
PIWIL2
PRAME
SPA17
SSX2
WT1
ZC3H12B
ZSWIM1
3
2
1
0
-1
-2
-3
Purified tumor 1
Purified tumor 2
Purified tumor 3
Project.ID
COAD
READ
4
2
0
-2
-4
Sample.Type
Primary Tumor
Solid Tissue Normal
5T4
MUC16
CEACAM3
CEACAM4
CENPQ
SPA17
CTAG2
MAGEA4
GAGE1
ZC3H12B
DNAJB7
MART1
ARSJ
MAGEA3
KLK3
MAGEC2
MAGEA1
SSX2
WT1
ZSWIM1
LAGE3
CYP2B6
PIWIL2
GP100
AFP
ACRBP
PRAME
CT45A1
GP100
LAGE3
5T4
KLK3
PRAME
SPA17
CYP2B6
MUC16
WT1
CEACAM3
ARSJ
CENPQ
ZSWIM1
ACRBP
CEACAM4
PIWIL2
DNAJB7
MART1
ZC3H12B
2
1
0
-1
-2
Near 2
Near 1
Near 3
Far 1
Far 2
Far 3
Tumor 1
Tumor 2
Tumor 3

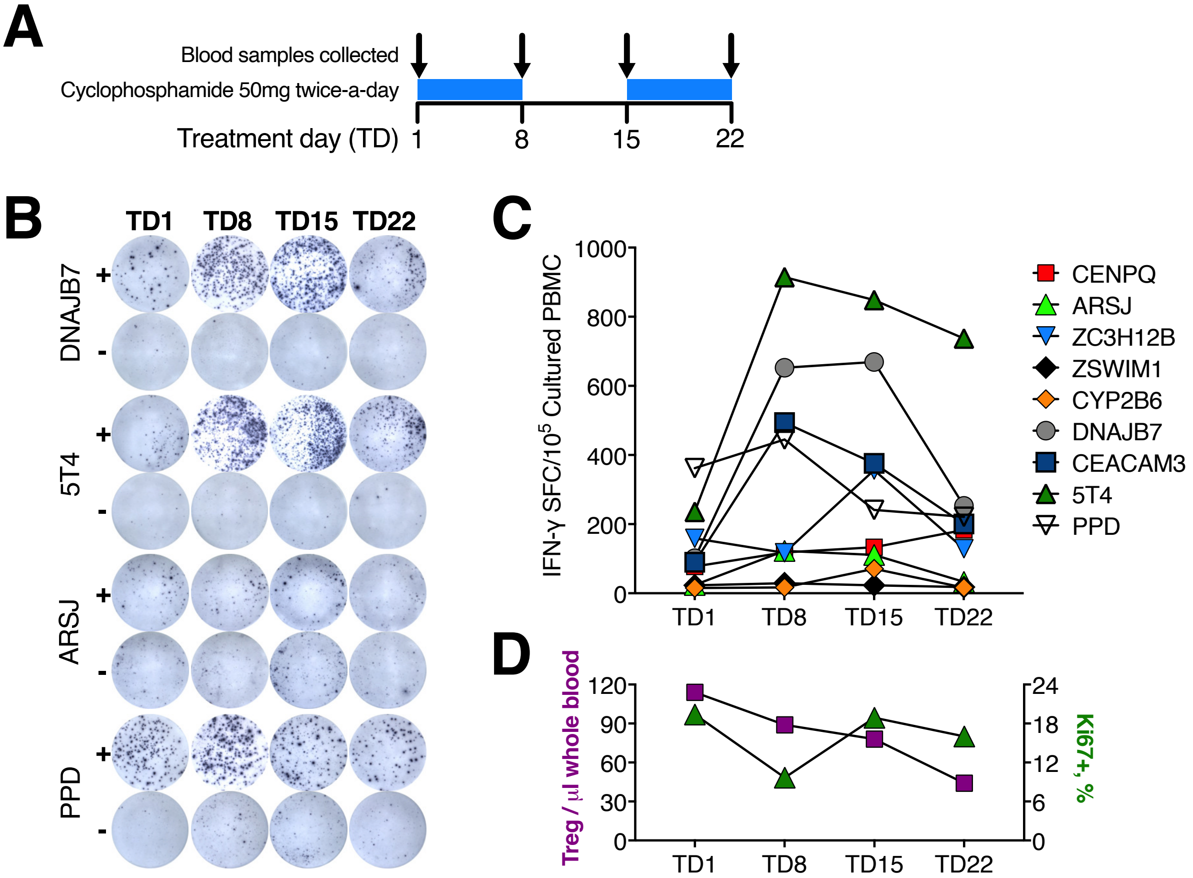
# Figure 4


A
IFN-γ⁺ T cell response
(SFC/10⁵ Cultured PBMC/+ve epitope)
0
6
12
18
CENPQ
ARSJ
ZC3H12B
ZSWIM1
CYP2B6
DNAJB7
CEACAM3
5T4
B
IFN-γ⁺ T cell response
(SFC/10⁵ Cultured PBMC/+ve epitope)
0
6
12
18
○ HD
● TNM Stage 1/2
● TNM Stage 3
CENPQ
ARSJ
ZC3H12B
ZSWIM1
CEACAM3
CYP2B6
DNAJB7
C
IFN-γ⁺ T cell response
(SFC/10⁵ Cultured PBMC/+ve epitope)
0
5
10
15
● HCC
● HNC
● CC
DNAJB7

# Figure 5


A
Granzyme B$^+$ T cell response (SFC/10$^5$ Cultured PBMC/+ve epitope)
15
10
5
0
DNAJB7
B
Predicted HLA-A*02 DNAJB7 9mers
LTFFLVNSV
SLAFDNSGM
FTFHKPDDV
GMDNYISVT
FLVNSVANE
C
DNAJB7
-ve
IFN-γ SFC/10$^5$ Lymphocytes
500
400
300
200
100
0
DNAJB7
-ve
D
Days post Ad5-DNAJB7 transduction
+ve
-ve
UT
1
2
3
4
5
6
Ad5-EMPTY D1
Ad5-EMPTY D6
35kDA
DNAJB7
45kDA
Actin
E
Caco-2
Donor 1 DNAJB7 CD8$^+$ T cells
Ad5-DNAJB7
Ad5-EMPTY
UT
w/o effector
Normalized Cell Index
1.1
1.0
0.9
0.8
0.7
0.6
***
0
4
8
12
16
20
24
Time post T cell addition, Hours
Donor 2 DNAJB7 CD8$^+$ T cells
Normalized Cell Index
***
Time post T cell addition, Hours
F
SW480
Donor 1 DNAJB7 CD8$^+$ T cells
Normalized Cell Index
1.6
1.4
1.2
1.0
***
Time post T cell addition, Hours
Donor 2 DNAJB7 CD8$^+$ T cells
Normalized Cell Index
***
Time post T cell addition, Hours

# Figure 6


A
Blood samples collected
Cyclophosphamide 50mg twice-a-day
Treatment day (TD)
1
8
15
22
B
TD1
TD8
TD15
TD22
DNAJB7
+
-
5T4
+
-
ARSJ
+
-
PPD
+
-
C
IFN-γ SFC/10^5 Cultured PBMC
0
200
400
600
800
1000
TD1
TD8
TD15
TD22
CENPQ
ARSJ
ZC3H12B
ZSWIM1
CYP2B6
DNAJB7
CEACAM3
5T4
PPD
D
Treg / μl whole blood
0
30
60
90
120
Ki67+, %
0
6
12
18
24
TD1
TD8
TD15
TD22