Supporting information

An analysis of SARS-CoV-2 cell entry genes identifies the intestine and colorectal cancer as susceptible tissues

Mahnaz Darvish-Damavandi, James Laycock, Christopher Ward, Milou S van Driel, Mae A Goldgraben and Simon JA Buczacki

Methods

Genotype-Tissue Expression (GTEx) data were downloaded from the web portal https://gtexportal.org/home/ [1]. TCGA data were downloaded from cBioPortal https://www.cbioportal.org/ and Morpheus https://software.broadinstitute.org/morpheus/ [2]. GEPIA data were acquired direct from the GEPIA webtool http://gepia.cancer-pku.cn/ [3].

Data were statistically analysed and graphs constructed in both RStudio v1.2.5001 and Prism v6.07 (GraphPad). Significance of parametric data were tested using a two-tailed Student's t test or F test. For nonparametric data, the Mann Whitney U test was used. For multiple group comparisons, one-way ANOVA or Kruskal-Wallis testing was performed. K means clustering was carried out in R using the kmeans function. The optimal numbers of clusters for k means was determined using the silhouette method in the ClusterR R package. Principal component plots were generated using the clusplot function from the cluster R package to identify the dominant gene/component in clustering. All clustering and optimal number of clusters were further validated using the mclust R package.

TCGA *ACE2*/*TMPRSS2* correlative analysis was performed using R and cBioPortal. COAD and READ samples were split in to high and low *ACE2* and *TMPRSS2* expressor groups defined by levels greater or less than the median level of each gene in all samples. Comparative analysis was then performed using the 'Compare Groups' function.

Figures were assembled in Adobe Illustrator v 24.1.1. Fig 1a was generated using Biorender https://biorender.com/.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 02/05/2020.

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

1. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.

2. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.

3. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98-w102.

Fig S1 – Expression of *ACE2* and *TMPRSS2* across normal tissues

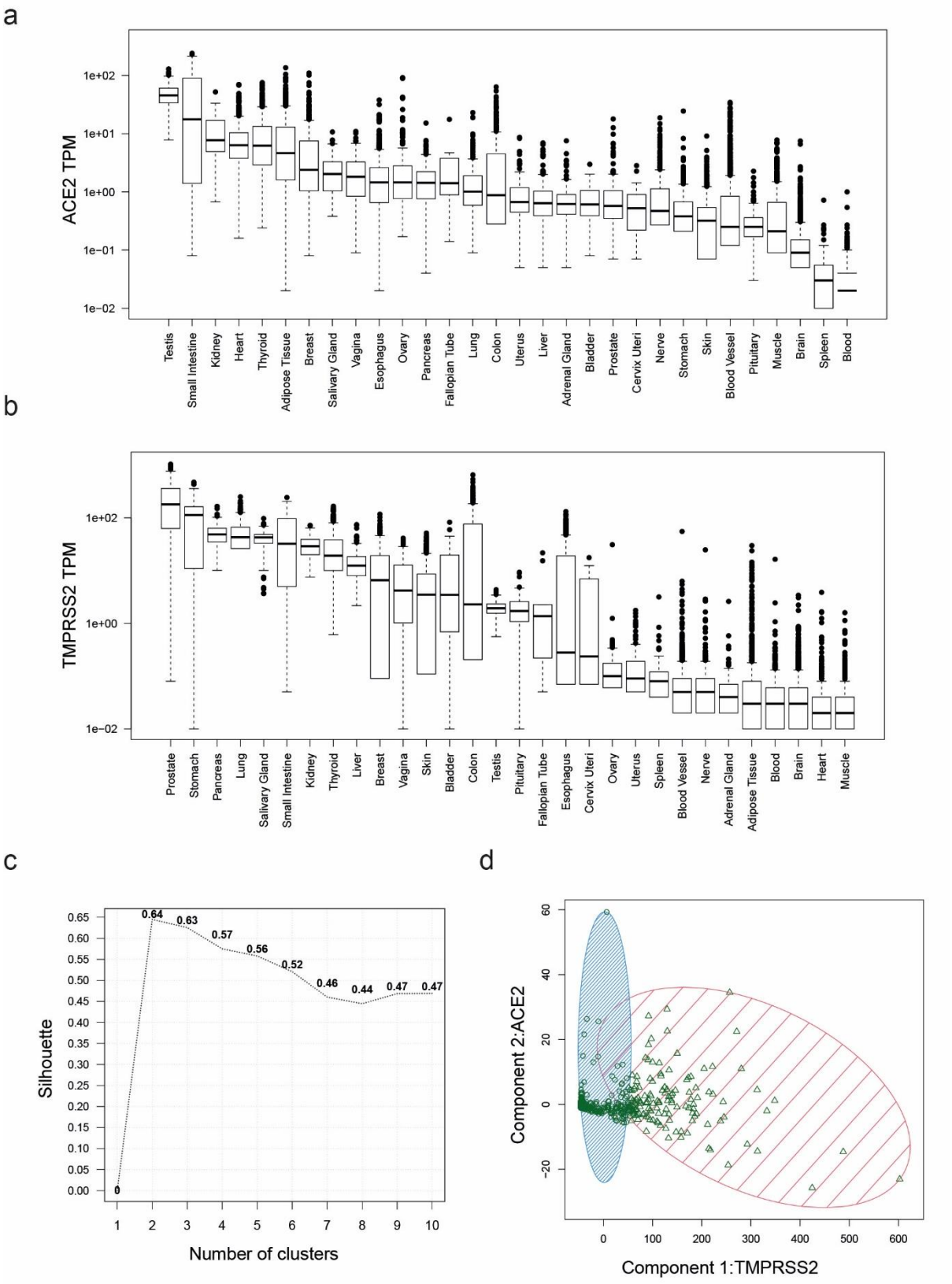


Fig S2 - Colonic *ACE2* and *TMPRSS2* levels by age, sex, location and tissue type

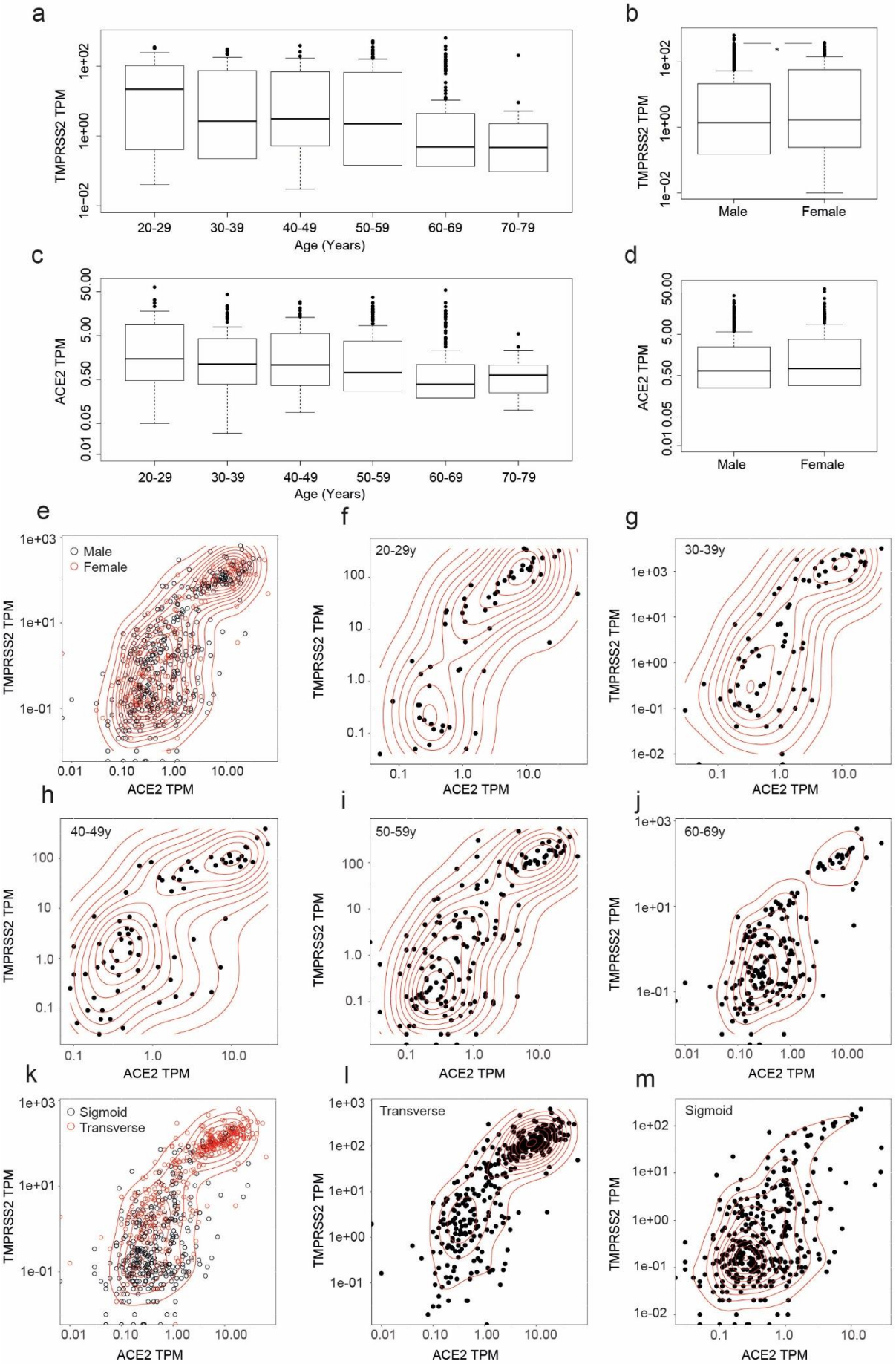


Fig S3 – Tumours demonstrating high relative changes in *ACE2* and *TMPRSS2* levels between normal and cancer

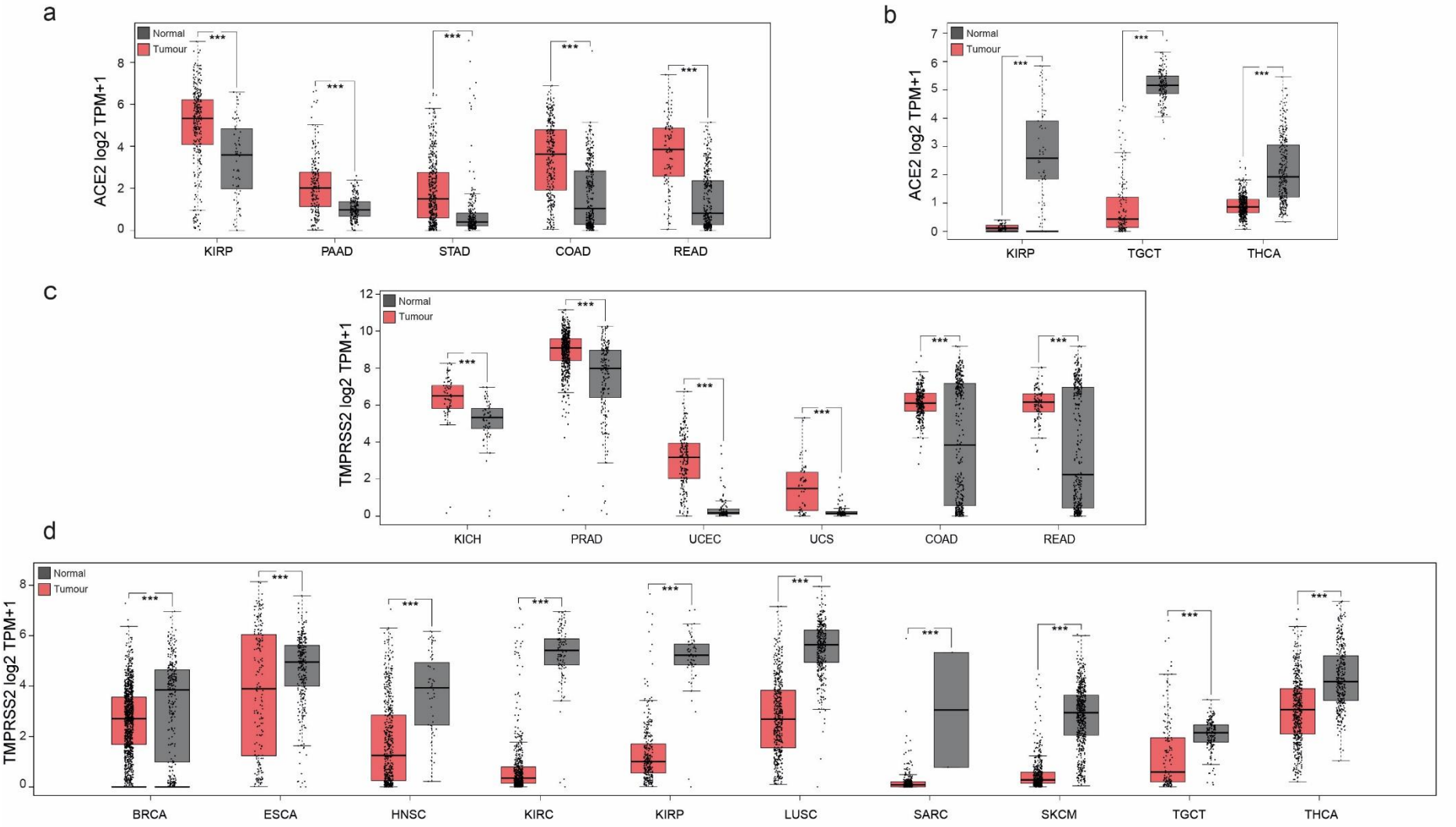
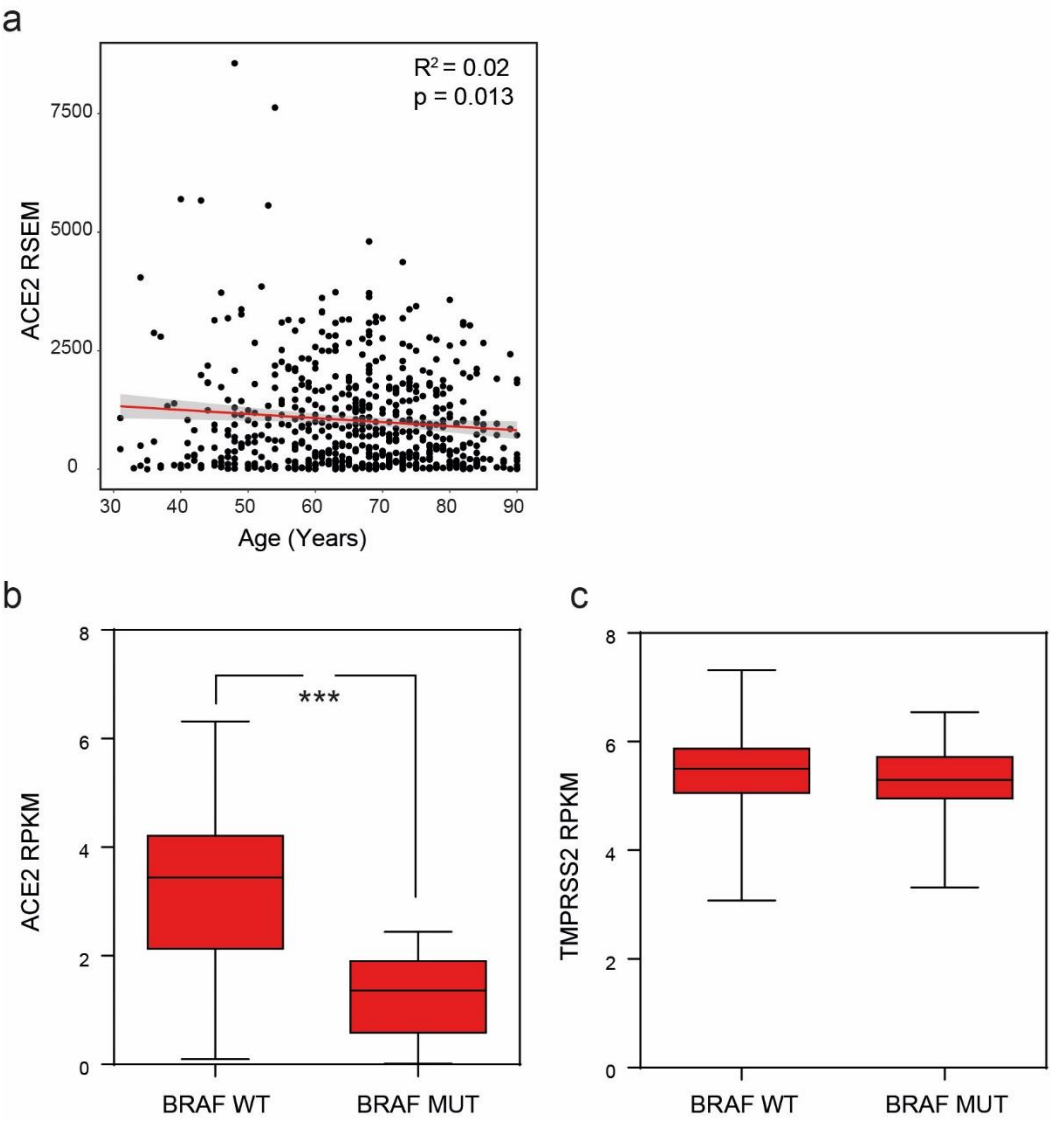


Fig S4 – Significant correlations between colorectal cancer *ACE2* and *TMPRSS2* levels and clinical parameters



Supplementary figure legends

SF1 – **(a)** Box and whisker plot of the range of expression of *ACE2* across normal tissue types. Median +/- IQR. TPM=Transcripts per million. **(b)** Box and whisker plot of the range of expression of *TMPRSS2* across normal tissue types. Median +/- IQR. **(c)** Plot of calculated silhouette value against number of clusters for colon GTEx samples. The higher the silhouette value equates to the most appropriate number of clusters. **(e)** Scatter cluster plot of the principal components from colon GTEx samples confirming two clusters ($ACE2^{High}$ $TMPRSS2^{High}$/ $ACE2^{Low}$ $TMPRSS2^{Low}$) and the relative contribution to clustering from *ACE2* and *TMPRSS2*.

SF2 – **(a)** Box and whisker plot of expression of *TMPRSS2* by age across GTEx colon samples. Median+/- IQR. TPM=Transcripts per million. **(b)** Box and whisker plot of expression of *TMPRSS2* by sex across GTEx colon samples. Median+/-IQR. *, $p<0.05$. **(c)** Box and whisker plot of expression of *ACE2* by age across GTEx colon samples. Median+/-IQR. **(d)** Box and whisker plot of expression of *ACE2* by sex across GTEx colon samples. Median+/-IQR. **(e)** Scatter plot of expression of *ACE2* and *TMPRSS2* from colon GTEx samples and coloured by sex (Red=Female, Black=Male). **(f-j)** Scatter plots of *ACE2* and *TMPRSS2* expression separated by age group from colon GTEx colon samples **(f)** 20-29y **(g)** 30-39y **(h)** 40-49y **(i)** 50-59y **(j)** 60-69y. **(k)** Scatter plot of *ACE2* and *TMPRSS2* expression from colon GTEx samples and coloured by location/tissue (Red=transverse colon/mucosa, Black=sigmoid colon/mucosa depleted). **(l,m)** Scatter plots of expression of *ACE2* and *TMPRSS2* from **(l)** transverse colon (mucosa) and **(m)** sigmoid colon (mucosa depleted).

SF3 – **(a)** Box and whisker plot of expression of *ACE2* between normal (Grey, TCGA and GTEx) and tumour (Red, TCGA) samples where expression is higher in tumour. One-way ANOVA. ***, $p<0.001$. TPM=Transcripts per million. **(b)** Box and whisker plot of expression of *ACE2* between normal (Grey, TCGA and GTEx) and tumour (Red, TCGA) samples where expression is higher in normal. One-way

ANOVA. ***, P<0.001. TPM=Transcripts per million. **(c)** Box and whisker plot of expression of *TMPRSS2* between normal (Grey, TCGA and GTEx) and tumour (Red, TCGA) samples where expression is higher in tumour. One-way ANOVA. ***, P<0.001. TPM=Transcripts per million. **(d)** Box and whisker plot of expression of *TMPRSS2* between normal (Grey, TCGA and GTEx) and tumour (Red, TCGA) samples where expression is higher in normal. One-way ANOVA. ***, P<0.001. TPM=Transcripts per million.

TCGA data set abbreviations

BRCA Breast invasive carcinoma

COAD Colon adenocarcinoma

ESCA Esophageal carcinoma

HNSC Head and neck squamous cell carcinoma

KICH Kidney chromophobe

KIRC Kidney renal clear cell carcinoma

KIRP Kidney renal papillary cell carcinoma

LUSC Lung squamous cell carcinoma

PAAD Pancreatic adenocarcinoma

PRAD Prostate adenocarcinoma

READ Rectum adenocarcinoma

SARC Sarcoma

SKCM Skin cutaneous melanoma

STAD Stomach adenocarcinoma

TGCT Testicular germ cell tumours

THCA Thyroid carcinoma

UCS Uterine carcinosarcoma

UCEC Uterine corpus endometrial carcinoma

SF4 – **(a)** (A) Scatter plot of the expression of *ACE2* by age from TCGA COAD/READ data sets. R^2 = 0.0172, p=0.013. RSEM = RNA-Seq by Expectation-Maximization. **(b,c)** Box and whisker plots of expression of *ACE2* **(b)** and *TMPRSS2* **(c)** between BRAF wildtype (wt) and BRAF mutant (mut) TCGA COAD and READ samples. Median +/- IQR. ***, p<0.001. RPKM = Reads Per Kilobase of transcript, per Million mapped reads.