Supplementary Information

Immune Landscape, Evolution, Hypoxia-mediated Viral Mimicry Pathways and Therapeutic Potential in Molecular Subtypes of Pancreatic Neuroendocrine Tumors

Running Title: Immune Landscape of Pancreatic Neuroendocrine Tumors

Supplementary Methods

RNAseq experiments and data analysis (Validation Cohort-1)

Libraries for RNAseq were prepared according to the manufacturer's instructions for the TruSeq Stranded Total RNA Sample Prep Kit with Ribo-Zero Gold rRNA Removal and unique dual indices (Illumina, San Diego, CA) with input of 100ng total RNA. Agilent Bioanalyzer (Santa Clara, CA) and Qubit fluorometry (Invitrogen, Carlsbad, CA) were used to determine the concentration and size distribution of the completed libraries. Libraries were sequenced following Illumina's standard protocol either using a) the Illumina cBot and HiSeq 3000/4000 PE Cluster Kit or b) the Illumina NovaSeq 6000 and an S2 flow cell. For the HiSeq, 100 X 2 paired end reads were sequenced on an Illumina HiSeq 4000 using Hiseq 3000/4000 sequencing kit and HCS v3.3.52 collection software. Base-calling was performed using Illumina's RTA version 2.7.3. For the NovaSeq, 150 X 2 paired end reads were sequenced using the NovaSeq S2 reagent kit and NovaSeq Control Software v1.3.1. Base-calling was performed using Illumina's RTA version 3.3.3. The sequencing for 162 of 173 samples was performed by Medical Genome Facility Genome Analysis Core, Mayo Clinic, Rochester, MN, USA and 11 samples by Aros Applied Biotechnology (which is now part of the Eurofins Genomics, Ebersberg, Germany).

Fastq files were merged before data analysis where samples were sequenced in more than two lanes of NovaSeq platform. The quality control of the sequence reads was performed using FastQC[1]. All the samples had quality score of more than 30 by PHRED quality score metrics. Bowtie (v2.2.6)[2] was used to align the reads to reference transcriptome (GRCh37) as a part of RSEM (v1.2.29)[3]. The quality of mapped reads was checked using RSEM and SAMtools[4,5]. The quality of RNA and RNA species (ribosomal, coding, intronic and intergenic) were further checked using CollectRnaSeqMetrics program from Picard tools (v2.1.0; http://broadinstitute.github.io/picard) by mapping to reference genome (GRCh37; https://www.ncbi.nlm.nih.gov/grc). The read counts were calculated using the RSEM software[3], and they were normalized and converted to trimmed mean of M-values (TMM) and counts per million (CPM; transformed to log values; default parameters) using edgeR package[6].

nCounter immune panel verification

A case set of 38 samples taken from validation cohort-1 was used to perform immune gene profiling using the PanCancer Immune Profiling assay on nCounter platform (NanoString Technologies). Data analysis was carried out as per the manufacturer's instructions and as described previously[7]. The analysis was done for only a subset of SAM significant genes.

Young K, *et al. Gut* 2021; 70:1904–1913. doi: 10.1136/gutjnl-2020-321016

Gene Set Enrichment Analysis (GSEA) and single sample Gene Set Enrichment Analysis (ssGSEA)

GSEA[8] using the gene set "C7: immunologic signatures" from the Molecular Signatures Database (MSigDB)[9] was applied to the training cohort to assess enhancement of immune pathways in specific PanNET subtypes. The analysis was performed using the standalone GSEA package from GenePattern[9]. ssGSEA[8] was performed on the training cohort and both the validation cohorts (1 and 2) to validate the expression of genes associated with immune cell types[10], hypoxia (HALLMARKS_HYPOXIA), necroptosis (GO_NECROPTOTIC_SIGNALING_PATHWAY) and dendritic cell (GSE22282_HYPOXIA_VS_NORMOXIA_MYELOID_DC_UP) gene sets from MSigDB across PanNET subtypes. ssGSEA analysis was conducted using the ssGSEAProjection *R* package[8,11] from GenePattern[9]. ssGSEA calculates a separate enrichment score for each sample and gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular set are coordinately up- or down-regulated within a sample. This provides a gene set enrichment profile for each sample[11].

Probabilistic principal component analysis with covariates (PPCCA)

PPCCA, which is implemented as a part of our previously published exploBATCH machine-learning tool[12,13], can be applied to simultaneously assess the association of two clinical or biological factors (covariates) in a subset of gene expression data from patient samples. This is carried out using the statistical test $\Delta_{bk} = \beta_{bk} / SE(\beta_{bk})$ to determine if samples and their gene expression profile are distributed according to the two covariates in the PPCCA principal subspace. The same tool can assess if a particular (secondary) factor of interest retains or alters its association with the subset of gene expression data by statistically normalizing the main (primary) factor. The regression coefficient, β_{bk}, measures the effect of covariate(s) b on the k^{th} probabilistic principal component (pPC), and SE is the corresponding standard error. A factor can be normalized as follows:

$$\underline{u}_{ck} = \underline{u}_{ak} - \underline{x}_b \beta_{bk}$$

where $\underline{u}_{ak}$ and $\underline{u}_{ck}$ are scores of the k^{th} pPC before and after correcting for the covariate, respectively, whilst $\underline{x}_b$ is the variable defining covariate information.

Here, the PPCCA method[12,13] was used to confirm the association between PanNET subtypes (primary factor), hypoxia (secondary factor) and DAMP gene expression data (subset of genes) suggested by enrichment and gene expression analysis. Since methods like correlation or other association studies (most of the time) perform only pairwise comparisons, we applied the PPCCA method to perform simultaneous comparison of the three factors listed above. Here, the tool was applied to assess whether statistically normalizing for the primary factor (MLP-1) would affect the secondary factor (hypoxia) changing its correlation with the gene expression profile. The tool is publicly available - https://github.com/syspremed/exploBATCH.

Multiplex immunofluorescence

The Opal™ 7 Solid Tumor Immunology Kit was used to detect 6 markers including 4 lymphocyte markers (CD4, CD8, CD20, FOXP3), a macrophage marker (CD68), and an epithelial marker (Pan-cytokeratin) on formalin-fixed paraffin-embedded (FFPE) tissues. The Vectra® fluorescent multispectral imaging system was used for the analysis, which was carried out on 7 samples from the training cohort and 21 from the RNAseq validation cohort, including 12 MLP1/2 and 16 Intermediate. Using a median score of cell type/megapixel we considered differences in protein expression across the PanNET subtypes. Insulinoma-like samples were excluded as FFPE was only available for 2 cases.

Visualization of gene expression data

For the heatmap, genes were clustered (hierarchical clustering) by Cluster 3.0[14] using the default settings, followed by visualization of the clusters using *GENE-E* from GenePattern[9].

Statistics

Survival analysis was performed using Kaplan-Meier curve and log-rank test using ***R**-package survival* from GenePattern platform[9]. Kruskal-Wallis statistical test was performed. Graphical plots were done using ***R**-packages ggplot2 and survminer*. Student t test and fold change analysis were performed.

Supplementary Figures

Supplementary Figure 1. Analysis of differentially expressed immune related genes in PanNET according to subtype, clinicopathological characteristics and mutations. A. Heatmap of the 132 differentially expressed immune related genes (derived from the training cohort (n=72)) in validation cohort-1 (n=109). Immune gene expression landscape overlaps with that of the training cohort**. B-C.** B) Barplot and C) cross-table showing concordance of PanNET subtype classification of 10 samples from the training cohort that were also profiled by RNAseq. The classification was based on original 221 subtype-specific genes published in Sadanandam *et al*. Cancer Discovery 2015[15]. **D.** Heatmap of the 132 differentially expressed immune-related genes (derived from the training cohort (n=72)) in validation cohort-2 (n=26). Immune gene expression landscape overlaps with that of the training cohort**. E.** Heatmap showing the expression of the 132 immune-related genes according to PanNET subtype as detected by nCounter PanCancer Immune Profiling panel of genes from NanoString Technologies in 38 samples selected from training cohort and validation cohort-1. **F.** Heatmap of the insulinoma specific genes from 132 differentially expressed immune-related genes, derived from the training cohort (n=72). Genes specific to normal pancreas and β cells are highlighted. **G-K.** Differentially expressed immune related genes in the training cohort according to clinicopathological and mutational data. F) Heatmap of 12 differentially expressed immune-related genes according to tumour grade. G) Heatmap of 6 differentially expressed immune-related genes according to *MEN1* pathway mutations. H) Heatmap of 9 differentially expressed immune-related genes according to *DAXX/ATRX* pathway mutations. I) Heat map of 3 differentially expressed immune-related genes according to *mTOR* pathway mutations. J) Heat map of 4 differentially expressed immune-related genes according to T-stages. None

of the genes from F-J are specific to any particular phenotype (for example, specific grade, stage or mutation).

Supplementary Figure 2. ssGSEA analysis of hypoxia and necroptosis related gene sets in two PanNET cohorts according to molecular subtype. A. Number of cases with high or low ssGSEA hypoxia score in each PanNET subtype using training cohort (n=72). **B.** Kaplan-Meier survival plot and number of patients at risk according to low and high hypoxia score from training cohort (n=72). **C.** Number of cases with high or low ssGSEA hypoxia score in each PanNET subtype using validation cohort-1 (RNAseq, n=109). **D.** Percentage of cases of validation cohort-1 with high or low ssGSEA hypoxia score in MLP-1 subtype from validation cohort-1 (RNAseq, n=109). **E.** Number of cases with high or low ssGSEA necroptosis score in each PanNET subtype using the training cohort (n=72). **F.** Kaplan-Meier survival plot and number of patients at risk according to necroptosis score from training cohort (n=72). **G.** Number of cases with high or low ssGSEA necroptosis score in each PanNET subtype using validation cohort-1 (RNAseq, n=109). **H.** Percentage of cases of validation cohort-1 with high or low ssGSEA necroptosis score in MLP-1 subtype. Median hypoxia and necroptosis ssGSEA scores were used as a cut-off to dichotomize hypoxia and necroptosis scores into high or low. Kruskal-Wallis test was used for barplots and log-rank test for survival analyses.

Supplementary Figure 3. DAMP pathway and association analysis with hypoxia and PanNET subtypes. A. Heatmap of the 12 DAMP pathway genes in validation cohort-1, demonstrating enrichment in the MLP-1 subtype. Red indicates elevated expression, blue decreased, and white no change. **B-C.** 9 of 14 DAMP genes are B) statistically significant (FDR<0.2; as shown in the bar plot) and C) differential expression of 9 genes between MLP-1 vs. other subtypes. Majority of the genes are highly expressed in MLP-1 than in the other subtypes. Kruskal-Wallis statistical test was used. **D.** Pearson correlation of DAMP pathway genes and necroptosis score in MLP-1 samples of the training cohort, demonstrating a positive correlation in 11 genes and a negative correlation in 1 gene. * represents significant $p<0.05$ from correlation test. **E.** Significant ($p<0.0001$) association of 12 DAMP genes with subtypes (MLP-1 vs. others) and hypoxia (high vs. low) as assessed by PPCCA method (see Methods section). **F.** Number of samples of the training cohort (n=72) with high or low *TLR3* expression in each PanNET subtype from training cohort (n=72). **G.** Number of samples of the training cohort with high or low ssGSEA dendritic cell (DC) score in each PanNET subtype from training cohort (n=72). Kruskal-Wallis test was used for barplots.

Supplementary Figure 4. Validation of ssGSEA analysis for immune cell type specific gene sets and expanded immune score in PanNET subtypes. A-F. ssGSEA score (FDR ≤ 0.2) vs. PanNET subtype in A-C) validation cohort-1 (n=109) and D-F) validation cohort-2 (n=26). Based on the results of training cohort's analysis, immune cell type gene sets chosen for validation were macrophages, co-inhibition of T Cells and MHC Class I. **G-H.** T cell inflamed GEP[16] in validation cohorts 1 and 2.

Supplementary Figure 5. Comparison of immune gene expression and T cell inflamed GEP between PanNET subtypes and melanoma disease stages (primary vs. metastases vs. normal) samples. **A-B.** Expression of A) PDL1 and B) PDL2 in PanNET subtypes and melanoma disease stages. **C.** T cell inflamed GEP scores in PanNET subtypes and melanoma disease stages. In A-C, T cell inflamed GEP genes including PDL1 and PDL2 were merged after batch correction between

PanNET training cohort and melanoma dataset (GSE15605[17]). **D.** *FOXP3* gene expression across PanNET subtypes from training cohort (n=72).

Supplementary Tables

Supplementary Table 1. Cohort and subtype information. A-C. Summary characteristics of the A) training cohort (n=72), B) validation cohort-1 (n=119) and C) validation cohort-2 (n=26) classified into 4 PanNET subtypes. **D.** Classification of validation (RNASeq) cohort-1 (n=119) into 4 PanNET subtypes using NMF analysis of 221 PanNET subtype genes.

Supplementary Table 2. SAM analysis for different cohorts and different phenotypes including PanNET subtypes, somatic mutations and clinico-pathological characteristics. A. SAM analysis between PanNET subtypes and 600 immune-related genes**. B.** 132 differentially expressed genes from SAM analysis between PanNET subtypes and 600 immune-related genes, classified according to the subtype with highest expression. **C.** Matched genes across validation cohorts with 132 gene from training cohort SAM analysis and across two validation cohorts. **D.** SAM analysis between tumour grades using 600 immune-related genes in PanNETs. **E.** SAM analysis between presence/absence of *MEN1* mutation and 600 immune-related genes in PanNETs. **F.** SAM analysis between presence/absence of *DAXX/ATRX* mutation using 600 immune-related genes in PanNETs. **G.** SAM analysis between presence/absence of MTOR pathway (*PTEN* and *TSC2*) mutations using 600 immune-related genes in PanNETs. **H.** SAM analysis between different T (tumor)-stages of samples using 600 immune-related genes in PanNETs. **I.** SAM analysis between samples with tumour size greater than 2 cm and lesser than/equal to 2 cm using 600 immune-related genes in PanNETs.

Supplementary Table 3. ssGSEA analysis for hypoxia and necroptosis and characteristics of tumour size. A. ssGSEA enrichment scores for the training cohort using hypoxia gene set (hypoxia hallmarks.gmt,), necroptosis gene sets (GO necroptotic process msigdb.gmt) and dendritic cells gene set (dendritic cells Bosco msigdb.gmt); performed using GenePattern tool**. B.** Proportion of samples with high or low hypoxia score for each PanNET subtype**. C.** Tumour size data for the PanNET samples**. D.** Proportion of samples with high or low necroptosis score for each PanNET subtype.

Supplementary Table 4. Analysis of the relationship between DAMP pathway, PanNET subtype, and necroptosis in the training cohort. A. Enrichment analysis of 14 DAMP genes in the MLP-1 subtype; genes enriched at FDR < 0.2 are in blue. **B.** Pearson correlation coefficients and p values of 12 significant DAMP genes with ssGSEA score of necroptosis from 72 PanNET samples representing all the four subtypes from training cohort. **C.** Proportion of samples with high and low TLR3 expression for each PanNET subtype**. D.** Gene Card results for the 74 Immune-related genes showing the highest expression in MLP-1 subtype**. E.** Enrichment analysis for genes highly expressed in MLP-1 subtype from the training cohort using MSigDB's "Investigational analysis" tool and C7 gene sets.

Supplementary Table 5. immune cell type analysis. A. Immune cell type-related genes significantly overexpressed in MLP-1 compared to the other PanNET subtypes in the training cohort. **B.** ssGSEA scores of the training cohort using Immune cell type-

related gene sets**. C.** M1 and M2 macrophage genes that are significantly differential expressed between subtypes. **D-E.** ssGSEA scores using immune cell type-related gene sets from D) validation cohort-1 and E) validation cohort-2.

Supplementary Table 6. Multiplex immunofluorescence samples and details.

References

1 Andrews SF. A Quality Control Tool for High Throughput Sequence Data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 2010.
2 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;**10**:R25.
3 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;**12**:1-16.
4 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N*, et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;**25**:2078-9.
5 Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;**27**:2987-93.
6 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;**26**:139-40.
7 Ragulan C, Eason K, Fontana E, Nyamundanda G, Tarazona N, Patil Y*, et al.* Analytical Validation of Multiplex Biomarker Assay to Stratify Colorectal Cancer into Molecular Subtypes. Sci Rep 2019;**9**:7665.
8 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 2005;**102**:15545-50.
9 Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006;**38**:500-1.
10 Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;**160**:48-61.
11 Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF*, et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;**462**:108-12.
12 Nyamundanda G, Brennan L, Gormley IC. Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 2010;**11**:571.
13 Nyamundanda G, Poudel P, Patil Y, Sadanandam A. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. Sci Rep 2017;**7**:10849.
14 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 1998;**95**:14863-8.
15 Sadanandam A, Wullschleger S, Lyssiotis CA, Grotzinger C, Barbi S, Bersani S*, et al.* A cross-species analysis in pancreatic neuroendocrine tumors reveals

molecular subtypes with distinctive clinical, metastatic, developmental, and metabolic characteristics. Cancer Discovery 2015;**5**:1296-313.
16 Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR*, et al.* IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest 2017;**127**:2930-40.
17 Raskin L, Fullen DR, Giordano TJ, Thomas DG, Frohm ML, Cha KB*, et al.* Transcriptome profiling identifies HMGA2 as a biomarker of melanoma progression and prognosis. J Invest Dermatol 2013;**133**:2585-92.