Reference transcriptomes of porcine peripheral immune cells created through bulk and single-cell RNA sequencing

Juber Herrera-Uribe[1†], Jayne E. Wiarda[2,3,4†], Sathesh K. Sivasankaran[2,5], Lance Daharsh[1], Haibo Liu[1], Kristen A. Byrne[2], Timothy P.L .Smith[6], Joan K. Lunney[7], Crystal L. Loving[2‡*], Christopher K. Tuggle[1‡*]

[1] Department of Animal Science, Iowa State University, Ames, IA, USA.

[2] Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, Agricultural Research Service, United States Department of Agriculture, Ames, IA, USA

[3] Immunobiology Graduate Program, Iowa State University, Ames, IA, USA

[4] Oak Ridge Institute for Science and Education, Agricultural Research Service Participation Program, Oak Ridge, TN, USA

[5] Genome Informatics Facility, Iowa State University, Ames, IA, USA

[6] USDA, ARS, U.S. Meat Animal Research Center, Clay Center, Nebraska, USA

[7] USDA-ARS, Beltsville Agricultural Research Center, Animal Parasitic Diseases Laboratory, Beltsville, MD, USA.

[†] These authors have contributed equally to this work and share first authorship

[‡] These authors have contributed equally to this work and share senior and last authorship

***Correspondence**:

Corresponding authors: cktuggle@iastate.edu, crystal.loving@usda.gov

Keywords: Pig, immune cells, transcriptome, Single-cell RNA-seq, bulkRNA-seq, FAANG.

ABSTRACT

Pigs are a valuable human biomedical model and an important protein source supporting global food security. The transcriptomes of peripheral blood immune cells in pigs were defined at the bulk cell-type and single cell levels. First, eight cell types were isolated in bulk from peripheral blood mononuclear cells (PBMCs) by cell sorting, representing Myeloid, NK cells and specific populations of T and B cells. Transcriptomes for each bulk population of cells were generated by RNA-seq with 10,974 expressed genes detected. Pairwise comparisons between cell types revealed specific expression, while enrichment analysis identified 1,885 to 3,591 significantly enriched genes across all 8 cell types. Gene Ontology analysis for the top 25% of significantly enriched genes (SEG) showed high enrichment of biological processes related to the nature of each cell type. Comparison of gene expression indicated highly significant correlations between pig cells and corresponding human PBMC bulk RNA-seq data available in Haemopedia. Second, higher resolution of distinct cell populations was obtained by single-cell RNA-sequencing (scRNA-seq) of PBMC. Seven PBMC samples were partitioned and sequenced that produced 28,810 single cell transcriptomes distributed across 36 clusters and classified into 13 general cell types including plasmacytoid dendritic cells (DC), conventional DCs, monocytes, B cell, conventional CD4 and CD8 αβ T cells, NK cells, and γδ T cells. Signature gene sets from the human Haemopedia data were assessed for relative enrichment in genes expressed in pig cells and integration of pig scRNA-seq with a public human scRNA-seq dataset provided further validation for similarity between human and pig data. The sorted porcine bulk RNAseq dataset informed classification of scRNA-seq PBMC populations; specifically, an integration of the datasets showed that the pig bulk RNAseq data helped define the CD4CD8 double-positive T cell populations in the scRNA-seq data. Overall, the data provides deep and well-validated transcriptomic data from sorted PBMC

47 populations and the first single-cell transcriptomic data for porcine PBMCs. This resource will be
48 invaluable for annotation of pig genes controlling immunogenetic traits as part of the porcine
49 Functional Annotation of Animal Genomes (FAANG) project, as well as further study of, and
50 development of new reagents for, porcine immunology.
51
52 **INTRODUCTION**
53 A major goal of biological research is using genomic information to predict complex
54 phenotypes of individuals or individual cells with specific genotypes. Predicting complex
55 phenotypes is an important component of broad Genome-to-Phenome (G2P) understanding
56 (Koltes et al., 2019), and investing in sequencing of multiple animal genomes, including pigs
57 (*Sus scrofa*), for improved genome and cell functional annotation is key in solving the G2P
58 question (Andersson et al., 2015; Giuffra et al., 2019). In addition to their major role in the
59 world supply of dietary protein, pigs have anatomic, physiologic, and genetic similarities to
60 humans and serve as biomedical models for human disease and regenerative medicine (reviewed
61 in (Swindle et al., 2012; Kobayashi et al., 2018). Thus, deep annotation of porcine genome
62 function would be a major milestone for addressing the G2P question. A highly contiguous
63 porcine genome assembly with gene model-level annotation was recently published (Warr et al.,
64 2020). However, this annotation is based primarily on RNA sequencing (RNA-seq) data from
65 solid tissues, with few sample types representative of immune cells, with the exception of
66 alveolar macrophages and dendritic cells (Auray et al 2016). Given the interaction of animal
67 health and growth, any functional annotation of the porcine genome will be incomplete without
68 deep analysis of expression patterns and regulatory elements controlling the immune system.

The transcriptomes of circulating immune cells serve as a window into porcine immune physiology and traits (Chaussabel et al., 2010; Mach et al., 2013; Schroyen and Tuggle, 2015; Auray et al., 2020). Blood RNA profiling has been used to understand variation in porcine immune responses (Huang et al., 2011; Arceo et al., 2013; Knetter et al., 2015; Munyaka et al., 2019) and genetic control of gene expression (Maroilley et al., 2017). One goal of such research is to develop gene signatures predictive of disease states (Berry et al., 2010) and predict responses to immunizations and/or infections (Chaussabel and Baldwin, 2014; Tsang et al., 2014), as has been demonstrated in humans. Whole blood is easily collected from live animals, but represents an extremely complex mixture of cell types. Estimates of gene expression in mixed samples are inherently inaccurate as cell composition differences are difficult to adjust for, complicating the interpretation of RNA differences across samples and treatments. Thus, starting from whole blood transcriptomic data, it is nearly impossible to link gene expression and regulation to a specific cell or cell type. To determine direct regulatory interactions, we must analyze specific cell populations and even individual cells. A cell type-specific understanding of peripheral immune cell gene expression patterns will thus enhance biological understanding of porcine immunity, reveal targets for phenotyping, and provide a comparison to other species.

Predominant immune cell populations in porcine peripheral blood mononuclear cell (PBMC) preparations are comprised mainly of monocytes, B-cells, and T-cells, with minor fractions of dendritic cells (DCs), natural killer (NK) cells, and NKT-cells also present. Porcine peripheral T-cell populations (reviewed in (Gerner et al., 2009a; Gerner et al., 2015) and DCs (Summerfield et al., 2015; Auray et al., 2016) are readily described based on phenotype, though deeper characterization of porcine immune cells could improve identification of valuable reagent targets and biological understanding of porcine immunity. T-cell populations are commonly

92 grouped as αβ or γδ T-cells according to T-cell receptor (TCR) chain expression and further
93 divided based on CD2, CD4, CD8α, and/or CD8β expression. Pigs have a unique $CD2^-$ γδ T-cell
94 lineage contributing to higher percentages of circulating γδ T-cells (Takamatsu et al., 2006) and
95 unique αβ T-cells expressing both CD4 and CD8α (Zuckermann, 1999). Relatively little is
96 known about different circulating B-cell populations in pigs, as reagents for phenotyping are
97 limited.

98 Various technical approaches can be used to enrich or isolate specific cell populations,
99 improving resolution of cell types for deeper interrogation of gene expression. Flow cytometry is
100 used to characterize cells based on expression of cell type-specific protein markers, and live cells
101 can be sorted by magnetic- and/or fluorescence-activated cell sorting (MACS/FACS) for use in
102 subsequent assays. MACS/FACS enrichment followed by transcriptomic analysis can provide
103 additional insight of gene expression in specific cell types, but cells expressing the same
104 combination of markers are often still a heterogeneous mixture (Sutermaster and Darling, 2019).
105 Some major subtypes of porcine immune cell populations can be labeled for cell sorting by
106 existing antibody reagents (Gerner et al., 2009b), but some subtypes such as B-cells lack these
107 resources.

108 An exciting alternative to sorting specific cell types for transcriptomic analysis is single-
109 cell RNA-seq (scRNA-seq). Many scRNA-seq approaches do not require prior
110 phenotypic/functional information or antibody reagents but instead rely on physical partitioning
111 of cells to uniquely tagged transcripts from individual cells and sharpen resolution of subsequent
112 transcriptomic analysis to single cells (Liu and Trapnell, 2016; Vieira Braga et al., 2016; Zheng
113 et al., 2017). scRNA-seq methods have been applied to human PBMCs (Zheng et al., 2017) and
114 provide more accurate and detailed analyses of transcriptional landscapes that can identify new

115 cell types (Villani et al., 2017) when compared to other transcriptomic approaches. There are
116 limitations to scRNA-seq, with tradeoffs in total genes detected per cell versus total cells
117 captured for analysis, depending on the approach used (Wilson and Göttgens, 2018).
118 To deeply annotate the porcine genome for peripheral mononuclear immune cell gene
119 expression and further inform phenotype and function of the heterogenous pool of immune cells
120 in PBMC preparations, two approaches were used to isolate peripheral immune cells for RNA-
121 seq. MACS followed by FACS was used to enrich for eight PBMC populations using
122 population-specific cell surface markers, and RNA isolated from enriched populations was used
123 for bulk RNA-seq (bulkRNA-seq) or a NanoString assay to evaluate gene expression. PBMCs
124 were also subjected to droplet-based partitioning for scRNA-seq. Gene expression patterns of
125 porcine immune cells using different approaches were compared to each other and to multiple
126 human datasets. Complementary methods provided an improved annotation and deeper
127 understanding of porcine PBMCs, as well as explicated datasets for further query by the research
128 community.
129

130 **MATERIAL AND METHODS**

131 **Animals and PBMC isolation**

132 Four separate PBMC isolations were performed, with different animals used in each
133 experiment. Cells were used for bulkRNA-seq, targeted RNA detection (NanoString), or scRNA-
134 seq. PBMCs from experiments were used as follows: Experiment A (ExpA) for bulkRNA-seq of
135 sorted populations from two ~6-month-old pigs (A1, A2); Experiment B (Exp B) for NanoString
136 and scRNA-seq from three ~12-month-old pigs (B1, B2, B3); Experiment C (ExpC) for scRNA-
137 seq from three ~12-month-old pigs (C1, C2, C3); Experiment D (ExpD) for scRNA-seq from

two ~7-week-old pigs (D1, D2). All pigs were crossbred, predominantly Large White and Landrace heritage. All animal procedures were performed in compliance with and approval by NADC Animal Care and Use Committee. PBMCs were isolated, enumerated, and viability assessed as previously described (Byrne et al., 2020).

Enrichment and sorting eight leukocyte populations by MACS/FACS

PBMCs were labeled with biotin labeled anti-porcine CD3ε (PPT3, Washington State University Monoclonal Antibody Center) for 15 min at 4 ºC, mixing continuously. Cells were washed with Hank's Balanced Salt Solution (HBSS), incubated with anti-biotin microbeads (Miltenyi Biotec), placed on LS columns, and separated into $CD3\varepsilon^+$ and $CD3\varepsilon^-$ fractions according to manufacturer's directions (Miltenyi Biotec). $CD3\varepsilon^+$ and $CD3\varepsilon^-$ fractions were each fluorescently-sorted into four subpopulations based on surface marker expression shown in Figure 1 and Table 1. For NanoString assays, B-cells were sorted as $CD3\varepsilon^-CD172\alpha^-CD8\alpha^-$; CD21 was not used for sorting. Each fraction for FACS was confirmed $CD3\varepsilon^+$ or $CD3\varepsilon^-$ by labeling with anti-mouse IgG1-PE-Cy7 to detect anti-CD3ε antibody used for MACS. Cells were sorted into supplemented HBSS using a BD FACSAria II with 70mm nozzle. After sorting, cells were pelleted and enumerated as described above. Sorted cell purity was >85% for each population. Cells were stained, sorted, and further processed within 10h of collection keeping cells on ice between processing steps.

RNA isolation for bulkRNA-seq/NanoString

BulkRNA-seq: after FACS, cells were pelleted, enumerated, and immediately lysed in RLT Plus buffer. RNA extractions were performed using the AllPrep DNA/RNA MiniKit

(QIAGEN) following manufacturer's instructions. Eluted RNA was treated with RNase-free DNase (QIAGEN). RNA quantity/integrity were assessed with an Agilent 2200 TapeStation system (Agilent Technologies). Samples used had RNA integrity numbers (RINs) ≥7.9. From ExpA, only one RNA sample for NK cells was used.

For NanoString assay: after FACS, cells were pelleted, enumerated, and immediately stored in Trizol. RNA extraction was performed using the Direct-zol RNA MicroPrep Kit (Zymo) with on-column DNase treatment following manufacturer's instructions. RNA quantity and integrity were assessed as described above, with RINs ≥6.9. RNA was preserved at -80 °C until further use.

BulkRNA-seq library preparation and data analysis

RNA was fragmented and 15 libraries prepared using the TruSeq Stranded Total RNA Sample Preparation Kit (Illumina). Libraries were diluted and pooled in approximately equimolar amounts. Pooled libraries were sequenced in paired-end mode (2x150-bp reads) using an Illumina NextSeq 500 (150 cycle kit).

Preprocessing, mapping, alignment, quality control

Data processing was performed as previously reported (Herrera-Uribe et al., 2020) using Sscrofa 11.1 genome and annotation v11.1.97 were used. Counts per gene of each sample in the two count tables were added together to get the final count table. Given that different types of immune cells have different transcriptome profiles (Hicks and Irizarry, 2015), YARN (Paulson et al., 2017), a tissue type-aware RNA-seq data normalization tool, was used to filter and normalize the count table. Genes with extremely low expression levels (<4 counts in at least one cell type)

were filtered out using filterLowGenes(). The final count table contained 12,261 genes across 15 samples, which was then normalized using normalizeTissueAware(), which leverages the smooth quantile normalization method (Hicks et al., 2018).

Data quality control was performed using DESeq2 (v1.24.0) (Love et al., 2014) within RStudio s (v1.2.1335). Regularized log-transformation was applied to the normalized count table with the rld function. Then principal component analysis (PCA) and sample similarity analyses were carried out and visualized using plotPCA() and distancePlot(), respectively. Heatmaps to display enriched genes were created using pheatmap (v1.0.12) within RStudio.

Cell type-enriched and cell type-specific gene identification

The normalized count table was used for differential gene expression (DGE) analysis with DESeq2 by setting the size factor for each sample to 1. A generalized linear model was fitted for each gene in the count table, with negative binomial response and log link function of the effect of cell types and pig subjects. nbinomWaldTest() was used to estimate and test the significance of regression coefficients with the following explicit parameter settings: betaPrior=FALSE,maxit=50000,useOptim=TRUE,useT=FALSE,useQR=TRUE. Cell type-enriched genes and cell type-specific genes were identified using the results function separately. A gene was labeled as cell-type enriched if the expression level (averaged across replicates) in one cell type was at least 2x higher than the average across all remaining cell types and adjusted p-value <0.05. A gene was labeled as cell type-specific if the averaged expression level in one cell type was at least 2x higher in pairwise comparison to the average in each other cell type and adjusted p-value <0.05 (Benjamini and Hochberg, 1995). Heatmaps to display specific genes were created as mentioned above.

For cross-species comparison, human hematopoietic cell (Haemopedia) RNA-seq expression data (Hilton Laboratory at the Walter and Eliza Hall Institute[1]) was used. Only orthologous genes with one-to-one matches between human and pig (orthologous gene list obtained from BioMart (Durinck et al., 2009) were compared. Orthologous gene transcript per million (TPM) values from naive and memory B-cells, myeloid dendritic cells (myDC), myeloid dendritic cells CD123+ (CD123PmDC), plasmacytoid DC (pDC), monocytes, NK cells, CD4T and CD8T cells from healthy donors were used (Choi et al., 2019). Spearman rank correlation analyses was performed to identify correlation between orthologous gene expression levels (absolute TPM) in pig and human sorted populations. Significance level was set at $P<0.05$ and level of Spearman's rank correlation coefficient (rho) was defined as low (<0.29), moderate (0.3-0.49), and strong (0.5-1) correlation.

Gene Ontology (GO) enrichment analysis

Metascape analysis (Zhou et al., 2019) was performed for GO analysis of the top 25% enriched genes and specific genes identified as described above, with threshold p-value <0.01. Several terms were clustered into the most enriched GO term. Term pairs with Kappa similarity score >0.3 were displayed as a network to show relationship among enriched terms. Terms associated with more genes tended to have lower P-values. All networks displayed were visualized using Cytoscape. All Ensembl Gene IDs with detectable expression level in each cell type were used as the background reference.

NanoString assay and data analysis

[1] https://www.haemosphere.org/datasets/show

A total of 230 test genes with nine housekeeping genes, eight positive and nine negative control genes were chosen for gene expression quantification on the NanoString nCounter analysis system (NanoString Technologies) using custom-made probes. The custom designed CodeSet was selected from genes and pathways associated with porcine blood, lung, lymph node, endometrium, placenta or macrophage response to infection with a porcine virus (Van Goor et al., 2020). RNA samples were diluted to 25-100 ng/ul in RNase-free water, and 5 ul of each sample was used in the assay using manufacturer's instructions with the nCounter Master kit.

The nCounter analysis system produces discrete count data for each gene assayed within each sample. We used the NanoString software nSolver Analysis Software (v3.0, NanoString Technologies), following manufacturer's instructions. The nSolver corrected for background based on negative control samples, performed within-sample normalization based on positive control probes, and performed normalization across samples using the median expression values of housekeeping genes (*GAPDH, HMBS, HPRT1, RPL32, RPL4, SDHA, TBP, TOP2B, YWHAZ*), providing confidence in our normalization method.

All statistical analyses were performed using the statistical programming language R v3.5. Raw count data were normalized using normalizationFactors() and NanoStringDataNormalization() from NanoStringDiff (v1.1.2.0) (Wang et al., 2016). One gene (*ISG20*) without detected expression in any samples was removed. Hierarchical clustering and PCA suggested there were substantial hidden variations among the expression data. Surrogate variable analysis has been shown to be a powerful method to detect and adjust for hidden variations in high throughput gene expression data (Li et al., 2014; Qian Liu, 2016), so surrogate variable analysis was applied to remove further hidden variations in the gene expression data

using svaseq() from sva (v3.30.1) (Leek et al., 2012). A full model with cell subpopulations and RINs as independent variables, and a reduced model with RINs as the only independent variable were used. Three surrogate variables were estimated and used to adjust for the hidden variations.

Gene expression values were transformed to $\log_2$(TPM) using voom() from limma (Law et al., 2014). Linear mixed effect models were used to fit the transformed gene expression data by using lmer()in lme4 (Bates et al., 2015). The model included fixed effect for cell subpopulation, RIN, the three surrogate variables, and random effect for each animal. One minus Spearman correlation coefficient was used as distance measure for gene clustering, and Euclidian distances was used for sample clustering.

Additionally, Spearman correlation analysis was performed to assess the correlation between bulkRNA-seq and NanoString results. The significant level was set at $P < 0.05$, and the level of Spearman's rank correlation coefficient (rho) was defined as described above.

scRNA-seq library preparation

PBMC isolation experiments were performed at different times and samples sequenced in different runs. For ExpB, $1x10^7$ viable PBMCs per animal were cryopreserved according to 10X Genomics Sample Preparation Demonstrated Protocol, shipped on dry ice to University of Minnesota's Core Sequencing Facility, and thawed, partitioned, and scRNA-seq libraries prepared. For ExpC/ExpD, freshly isolated PBMCs were transported on ice to Iowa State University Core Sequencing Facility for partitioning and library preparation. Partitioning and library preparation were performed according to Chromium Single Cell 3' Reagent Kits v2 User Guide (10X Genomics). For all experiments, 100 base paired-end reads were sequenced on an

Illumina HiSeq3000 at ISU Core Sequencing Facility. One sample from ExpB was omitted from further analyses due to poor sequence performance.

scRNA-seq data analysis

Read alignment/gene quantification

Raw read quality was checked with FASTQC[1]. Reads 2 (R2) were corrected for errors using Rcorrector (Song and Florea, 2015), and 3' polyA tails >10 bases were trimmed. After trimming, R2 >25 bases were re-paired using BBMap[2]. *Sus scrofa* genome Sscrofa 11.1 and annotation GTF (v11.1.97) from Ensembl were used to build the reference genome index (Yates et al., 2020). The annotation file was modified to include both gene symbol (if available) and Ensembl ID as gene reference (e.g. GZMA_ENSSSCG00000016903) using custom Perl scripts. Processed paired-end reads were aligned and gene expression count matrices generated using CellRanger (v4.0; 10X Genomics) with default parameters. Only reads that were confidently mapped (MAPQ=255), non-PCR duplicates with valid barcodes, and unique molecular identifiers (UMIs) were used to generate gene expression count matrices. Reads with same cell barcodes, same UMIs, and/or mapped to the same gene feature were collapsed into a single read.

Quality control/filtering

CellRanger output files were used to remove ambient RNA from each sample with SoupX (Young and Behjati, 2020) function autoEstCont(). Corrected non-integer gene count matrices were outputted in CellRanger file format using DropletUtils (Lun et al., 2019) function write10xCounts() and used for further analyses. Non-expressed genes (sum zero across all

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc
[2] https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/

296 samples) and poor quality cells (>10% mitochondrial genes, <500 genes, or <1,000 UMIs per
297 cell) were removed using custom R scripts and Seurat (Stuart et al., 2019). Filtered count
298 matrices were generated using write10xCounts() and used for further analyses. High probability
299 doublets were removed using Scrublet (Wolock et al., 2019), specifying 0.07 expected doublet
300 rate and doublet score threshold of 0.25.
301
302 ***Integration, visualization, and clustering***
303 Post-quality control/filtering gene counts/cells from each sample were loaded into a
304 Seurat object and transformed individually using SCTransform(). Data were integrated with
305 SelectIntegrationFeatures(), PrepSCTIntegration(), FindIntegrationAnchors(), and
306 IntegrateData() with default parameters. PCA was conducted with RunPCA(), and the first 14
307 principal componenets (PCs) were selected as significant based on <0.1% variation of successive
308 PCs. Significant PCs were used to generate two-dimensional t-distributed stochastic neighbor
309 embedding (t-SNE) and uniform manifold approximation and projection (UMAP) coordinates
310 for visualization with RunTSNE() and RunUMAP(), respectively, identify nearest neighbors and
311 clusters with FindNeighbors() and FindClusters() (clustering resolution = 1.85), respectively, and
312 perform hierarchical clustering with BuildClusterTree(). Counts in the RNA assay were further
313 normalized and scaled using NormalizeData() and ScaleData().
314
315 ***Differential Gene Expression (DGE) analyses***
316 Normalized counts from the RNA assay were used for DGE analyses. Differentially-
317 expressed genes (DEGs) between pairwise cluster combinations were calculated using
318 FindMarkers(). DEGs in one cluster relative to the average of all other cells in the dataset were

319 calculated using FindAllMarkers(). The default Wilcoxon Rank Sum test was used for DGE
320 analyses. Genes expressed in >20% of cells within one of the cell populations being compared,
321 with |logFC|>0.25, and adjusted p-value <0.05 were considered DEGs.
322
323 ***Gene set enrichment analyses (GSEA)***
324 Enrichment of gene sets within our porcine scRNA-seq dataset were performed using
325 AUCell (v1.10.0) (Aibar et al., 2017). Enriched genes in sorted porcine bulkRNA-seq
326 populations were identified as described in preceding methods. Log2FC values were used to
327 curate gene sets of genes enriched in the top 25%, 20%, 15%, 10%, 5%, or 1% of bulkRNA-seq
328 populations. Gene sets from human bulkRNA-seq cell populations (Choi et al., 2019) were
329 recovered by performing a High Expression Search on the Haemosphere website[1], setting
330 Dataset=Haemopedia-Human-RNASeq and Sample group=celltype. Gene sets for CD4:+ T-cell;
331 CD8:+ T-cell; Memory B-cell; Monocyte; Myeloid Dendritic Cell; Myeloid Dendritic Cell
332 CD123+; Naïve B-cell; Natural Killer Cell; and Plasmacytoid Dendritic Cell options
333 corresponded to CD4T, CD8T, MemoryB, Monocyte, mDC, CD123PmDC, NaïveB, NK, and
334 pDC designations, respectively. Genes with high expression scores >0.5 (lower enrichment level)
335 or >1.0 (higher enrichment level) were selected and filtered to include only one-to-one gene
336 orthologs as described in preceding methods. Human gene identifiers were converted to
337 corresponding porcine gene identifiers or gene names used for scRNA-seq analyses .
338 Within each cell of the finalized scRNA-seq dataset, gene expression was ranked from
339 raw gene counts. Area under the curve (AUC) scores were calculated from the top 5% of
340 expressed genes in a cell and the generated gene sets. Higher AUC scores indicated a higher

[1] https://www.haemosphere.org/searches

341 percentage of genes from a gene set were found amongst the top expressed genes for a cell. For
342 overlay of AUC scores onto UMAP coordinates of the scRNA-seq dataset, a threshold value was
343 manually set for each gene set based on AUC score distributions. For visualization by heatmap,
344 AUC scores were calculated for each cell, scaled relative to all other cells in the dataset, and
345 average scaled AUC scores were calculated for each cluster.
346
347 ***Deconvolution analysis (CIBERSORTx)***
348 To deconvolve cluster-specific cell subsets from bulkRNA-seq of sorted populations,
349 CIBERSORTx (Newman et al., 2019) was used to derive a signature matrix from scRNA-seq
350 data . 114 cells were taken from each cluster using the Seurat subset() function and labelled with
351 corresponding cluster identities. Cluster-labeled cells were used to obtain a single-cell reference
352 matrix (scREF-matrix) that was used as input and run on CIBERSORTx online server using
353 "Custom" option. Default values for replicates (5), sampling (0.5), and fraction (0.0) were used.
354 Additional options for kappa (999), q-value (0.01), and No. Barcode Genes (300-500) were kept
355 at default values. CIBERSORTx scREF-matrix was used to impute cell fractions from the
356 bulkRNA-seq of sorted cell population "mixtures". The mixture file (TPM values) was used as
357 an input and run on CIBERSORTx online server using the "Impute Cell Fractions" analysis with
358 the "Custom" option selected, and S-mode batch-correction was applied. Cell fractions were run
359 in relative mode to normalize results to 100%. The number of permutations to test for
360 significance were kept at default (100). Resulting output provided estimated percentages of what
361 scRNA-seq clusters defined each bulkRNA-seq sorted cell population.
362
363 ***Reference-based label transfer/mapping and de novo integration/visualization***

A CITE-seq dataset of human PBMCs (Hao et al., 2020) was used to transfer cell type annotations onto our porcine scRNA-seq dataset. Due to the cross-species comparison, we distilled human reference and pig query datasets to only include 1:1 orthologous gene, and human reference dataset was re-normalized and integrated mirroring previous methods (Hao et al., 2020). Each sample of the porcine query dataset was separately normalized using SCTransform. Anchors were found between the human reference and each pig query sample using FindTransferAnchors. Identified anchors were used to calculate mapping scores for each cell using MappingScore. The mapping scores provided a 0-1 confidence value of how well a porcine cell was represented by the human reference dataset. Prediction scores were calculated using available level 2 cell types from the human reference dataset. Prediction scores provided a 0-1 percentage value for an individual cell type prediction, based on how many nearby human cells shared the same cell type annotation that was predicted. Predicted cell annotations were projected back onto original UMAP of the porcine dataset. Cluster-averaged prediction and mapping scores were also calculated.

In order to identify cells from the porcine dataset that were not well represented by the human reference dataset the two datasets were integrated to perform *de novo* visualization by merging the two datasets and their respective sPCAs to create a new UMAP. From two-dimensional *de novo* UMAP, porcine cells that did not overlap with human cells were identified.

Cluster subsetting

For deeper analyses of only subsets of clusters in the scRNA-seq dataset, cells belonging to only selected clusters were place in a new Seurat object using subset(). Genes with zero overall expression in the new data subset were removed using DietSeurat(), and counts were re-

scaled with ScaleData(). Original cluster designations and PCs were left intact. UMAP/t-SNE visualization, hierarchical clustering, and DGE analyses were re-performed as described in the original analyses. Pairwise DGE analyses were not re-performed.

Random Forest (RF) Modeling

The RF models provided an estimate of cluster similarity based on error rates. The R packages caret[1] and ranger[2] were used to create RF models trained on cluster identities of cells. A normalized count matrix was used as input data for RF models. Each cell was labeled by its previously defined cluster. Two different types of models were created: (1) pairwise models where training data included only cells from two different clusters (ex. Clusters 0 & 3); (2) models where training data included cells from all clusters of a specified dataset (ex. all γδ T cell clusters). Each model was trained on the cluster identity of each cell, with trees created=500, target node size = 1, variables=14,386, variables to sample at each split (Mtry)=119. Each tree in the model is grown from a bootstrap resampling process that calculates an out-of-bag (OOB) error that provides an efficient and reasonable approximation of the test error. Variable importance was used to find genes or sets of genes that can be used to identify certain types of cells or discriminate groups of cells from one another. RF models are advantageous because they can provide ranked lists of genes most important for discriminating cells between different clusters. This method was used to identify groups of important genes to supplement single DGE analyses. Variable importance was assigned by measuring node impurity (Impurity) and using permutations (Permutation). Features that reduced error in predictive accuracy are ranked as

[1] https://cran.r-project.org/web/packages/caret/caret.pdf
[2] https://cran.r-project.org/web/packages/ranger/index.html

more important. High error rate in the model suggests cells from the groups being compared are more similar to each other, whereas low error rate suggests cells from each cluster are unique.

Gene name replacement

Several gene names/Ensembl IDs used for data analysis were replaced in main text/figures for the following reasons: gene symbol was not available in the annotation file but was available under the gene description on Ensembl, gene symbol was updated in future Ensembl releases, or multiple Ensembl IDs corresponded to a single gene symbol. Affected genes included: ABI3=ENSSSCG0000003522, ABRACL=ENSSSCG00000004145, AP3S1=ENSSSCG00000037595, CCDC12=ENSSSCG00000011329, CCL23=ENSSSCG00000033457, CD163L1=ENSSSCG00000034914, CDNF=ENSSSCG00000039658, CR2=ENSSSCG00000028674, CRIP1=ENSSSCG00000037142, CRK=ENSSSCG00000038989, EEF1A1=ENSSSCG00000004489, FCGR3A=ENSSSCG00000036618, GBP1=ENSSSCG00000024973, GBP7=ENSSSCG00000006919, GIMAP4=ENSSSCG00000027826, GZMA=ENSSSCG00000016903, HMGB1=ENSSSCG00000009327, HOPX=ENSSSCG00000008898, IFITM1=ENSSSCG00000014565, IGLL5=ENSSSCG00000010077, KLRB1B=ENSSSCG00000034555, KLRC1=ENSSSCG00000000640, KLRD1=ENSSSCG00000026217, MAGOHB=ENSSSCG00000000635, MAL=ENSSSCG00000040098, MAN2B1=ENSSSCG00000013720, MDK=ENSSSCG00000013260, MYL12A=ENSSSCG00000003691, NT5C3A=ENSSSCG00000022912, PRKCH=ENSSSCG00000005095,

PTTG1=ENSSSCG00000017032, RPL14=ENSSSCG00000011272, RPL22L1=ENSSSCG00000036114, RPL23A=ENSSSCG00000035080, RPL35A=ENSSSCG00000040273, RPS15A=ENSSSCG00000035768, RPS19=ENSSSCG00000003042, RPS27A=ENSSSCG00000034617, RPS3=ENSSSCG00000014855, RPS8=ENSSSCG00000003930, S100B=ENSSSCG00000026140, SIRPA=ENSSSCG00000028461, SLA-DQA1=ENSSSCG00000001456, SLA-DRA=ENSSSCG00000001453 (listed as HLA-DRA in the gene annotation used), SLA-DRB1=ENSSSCG00000001455, SLPI=ENSSSCG00000022258, SPIB=ENSSSCG00000034211, TMSB4X=ENSSSCG00000012119, TXN=ENSSSCG00000005453, WIPF1=ENSSSCG00000027348.

RESULTS

BulkRNA-seq revealed common and distinct transcriptomes in circulating immune cells

Eight immune cell populations (Table 1) were sorted by cell-surface marker phenotypes for transcriptomic profiling by bulkRNA-seq (Figure 1) using primarily criteria previously outlined (Gerner et al., 2009b), with some modifications. Our protocol utilized an antibody reactive to swine workshop cluster 6 (SWC6) protein to identify γδ T-cells, but the antibody only labels $CD2^-$ γδ T-cells (Yang and Parkhouse, 1996; Davis et al., 1998; Stepanova and Sinkora, 2013; Sedlak et al., 2014). $CD2^+$ γδ T-cells were likely sorted into the $CD3\varepsilon^+CD4^-CD8\alpha^-$ fraction that was not retained or the CD8T ($CD3\varepsilon^+CD4^-CD8\alpha^+$) population (Davis et al., 1998; Stepanova and Sinkora, 2013; Sedlak et al., 2014). A pan-B-cell marker for pigs is not currently available, so B-cells are often characterized through a series of negative gates. Cells in the $CD3\varepsilon^-$

fraction were considered B-cells if they also lacked expression of CD172α and CD8α. B-cells characterized in this manner were further terminally sorted into B-cell populations with or without CD21 (complement receptor 2) expression (CD21pB and CD21nB, respectively; Figure 1, gates 7 and 8 respectively). We acknowledge that the CD21nB gate likely contained other circulating cell types that were not sorted through positive gating approaches.

Transcriptomic profiles of sorted cell populations were constructed by bulk RNA-seq, and relationships among porcine immune cell transcriptomes were assessed and visualized through dimensionality reduction and hierarchical clustering (Figure 2A and 2B and Supplementary File 1). Specifically, T-cell populations (SWC6gdT, CD4T, CD4CD8T, CD8T), B-cell populations (CD21pB, CD21nB), myeloid leukocyte populations (Myeloid), and a single NK cell population (NK) were well separated from each other (Figure 2A) by PCA. Replicates of specific sorted cell populations clustered most closely together, while within T-cell populations or B-cell populations, considerable transcriptional similarity was observed (Figure 2B).

The total number of expressed genes in each sorted population was similar (Supplementary File 1). Significantly enriched genes (SEGs) with expression significantly different and at least 2x greater than the average of all other cell populations (see Methods) were identified for each sorted population (Supplementary File 2). Notably, around 12-18% of SEGs are not fully annotated (no symbol/gene name) in the Sscrofa 11.1 genome and annotation v11.1.97. The SWC6gdT population had the highest number of SEGs (3,591), while the NK population had the fewest (1,885) (Table 2). SEG lists were queried for corresponding protein targets used to sort cells, if known, to confirm enrichment of expression of genes corresponding to protein phenotypes (Figure 2C). Expression of *SIRPA*[*] (encoding CD172α) had the highest

[*] Refer to gene name replacement in Methods section

476 fold-change in the Myeloid population, and *CR2 (*encoding CD21*, ENSSSCG00000028674*), was
477 highest in the CD21pB population, as would be predicted based on protein phenotypes. The two
478 $CD4^+$ T-cell populations (CD4T and CD4CD8T) had the highest fold-change for *CD4.* The
479 CD8T population had the highest fold-change for *CD8A*, with CD4CD8T and NK populations
480 also having near a log_2FC enrichment value of 5, in line with these populations also expressing
481 CD8α. The SWC6gdT population had the highest fold-change for *TRDC*, though CD8T and
482 CD21nB populations also had enrichment for *TRDC*. As noted previously, it's unlikely our
483 sorting for γδ T-cells based on SWC6 captured all γδ T-cells, thus some γδ T-cells may be
484 represented in other sorted populations. Thus, the CD8T population is likely comprised not only
485 $CD8\alpha^+$ αβ T-cells, but also potentially $SWC6^-$ γδ T-cells expressing CD8α.
486 A subset of SEGs (25% highest log2FC values) for each sorted population, referred to as
487 highly enriched genes (HEGs) that distinguish different circulating pig immune cell populations,
488 were used for data visualization and GO analysis. The log_2FC values for HEGs were clustered
489 and visualized in Figure 3 (four $CD3\varepsilon^-$ populations) and Supplementary Figure 2 (four $CD3\varepsilon^+$
490 populations). GO analyses using HEG lists for each cell population indicated enrichment for
491 biological processes characteristic of each respective cell population, depicted as networks of
492 similar terms (Figure 3E-3H, Supplementary File 3, Supplementary Figure. 2E-2H). Terms for
493 Myeloid HEGs included Myeloid leukocyte activation and response to bacterium (Figure 3E),
494 and terms for NK HEGs included positive regulation of cell killing and natural killer cell
495 mediated cytotoxicity (Figure 3F). Many terms enriched for CD21pB HEGs overlapped with
496 those for CD21nB HEGs, as 38% of HEGs were shared between these populations (Figure 3C,
497 3D). Thus, top GO terms for B-cells, including adaptive immune response and B-cell
498 proliferation were present in both populations (Figure 3G and 3D). However, some GO terms

were unique to either B-cell population. GO related to B-cell activation, such as positive regulation of B-cell activation/proliferation processes associated with B-cell receptor signaling, were identified exclusively for CD21pB HEGs. For CD21nB HEGs, processes associated with humoral immunity and red blood cell processes such as coagulation or platelet activation were noted, which could indicate contamination of different cell types given the non-specific cell sorting approach used for CD21nB cells (Figure 1). For all sorted T-cell populations (CD8T, CD4T, CD4CD8T and SWC6gdT), HEG lists showed overlap (Supplementary Figure 2A-D). GO terms included T-cell activation, T-cell receptor signaling pathway, cytokine-cytokine receptor interaction and biological processes related to cytotoxicity activity (Supplementary Figure 2E-2F). Overall, GO exploration of HEGs for sorted populations provided evidence that sorted immune cells represented expected immune cell functions.

The TPM values of expressed genes in sorted porcine cells were compared with orthologous human genes expressed in sorted human naïve hematopoietic cells from the Haemopedia (Choi et al., 2019) in order to identify cell-specific transcriptome similarities across species. Gene expression correlations assessed by Spearman's rank correlation indicated highly significant and moderately strong correlations (rho=0.30-0.43, P<2.2e-16) between porcine and anticipated corresponding human immune cell populations (Supplementary Figure 3, Supplementary File 4). A closer evaluation of genes reported as canonical cell markers for different mouse and human peripheral immune cell populations[1] and expression of those genes in each of the sorted porcine populations revealed several commonalities. Specifically, genes such as *EBF1*, *CD19*, *MS4A1, CD79B*, *PAX5*, *HLA-DOB* (in CD21nB, CD21pB); *CD28* (in CD8T, CD4T, CD4CD8T); *CD5* (in CD8T, CD4T, CD4CD8T, SWC6gdT); *GZMA*, *GNLY*, *CCL5*,

[1] http://biocc.hrbmu.edu.cn/CellMarker/#

521 *KLRK1*, *KLRB1*, *CD244* (in NK, CD8T); and *VLDLR*, *NLRP3*, *CD14*, *STEAP4*, *CD163*, *DEFB1*
522 (in Myeloid) for human cells showed specific enrichment in respective porcine populations
523 (Supplementary Figure 4). Thus, additional query confirmed sorted porcine immune cell
524 populations were equivalent to human counterparts in many ways.
525
526 **High homogeneity amongst sorted T-cell and B-cell populations and transcriptomic**
527 **distinctions in Myeloid and NK populations**
528 Pairwise DGE analyses between the cell populations identified genes with transcript
529 abundance at least 2x higher in one population than in all other populations (adjusted p-value
530 <0.05, see Methods) which we define as cell type-specific. Consistent with PCA (Figure 2A),
531 more cell type-specific genes were identified in the Myeloid population than in NK, T or B-cells.
532 In total, we identified 2, 5, 8, 29, and 397 cell type-specific genes for CD8T, CD21pB,
533 SWC6gdT, NK, and Myeloid populations, respectively (Table 2, Supplementary Figure 5). GO
534 analyses using cell type-specific genes for the Myeloid population resulted in enrichment of
535 terms such as Myeloid leukocyte, cytokine-cytokine receptor interaction, and pattern recognition
536 receptor activity (Supplementary Figure 5, Supplementary file 3). Next, we determined if the cell
537 type-specific genes identified were present in the list of HEGs for each population. In total, 2, 2,
538 5, 14, and 271 cell type-specific genes were identified in respective HEG lists for CD8T,
539 CD21pB, SWC6gdT, NK, and Myeloid populations, respectively (Table 3), indicating the most
540 highly-enriched cell type-specific genes were present in NK and Myeloid populations. Cell type-
541 specific genes could not be identified for the remaining three sorted populations (CD4T,
542 CD4CD8T, and CD21nB) using the criteria described above, indicating between-population
543 transcriptional heterogeneity even for these enriched populations.

We then explored immune cell transcriptomic patterns to identify genes that could expand our knowledge of pathways active in specific cell populations, as well as predict new genes suitable to use for molecular analyses in immunology studies. Of interest, we found a remarkably high number of HEGs in our Myeloid population (Table 3), including immune-related genes involved in TLR signaling (*CD14*, *CD36*, *TLR2/3/4/8/9*, *NOD2*) and cytokine activity (*CSF1R*, *CSF2RA*, *CSF3R*, *IFNGR1*, *IL1B*, *IL1RAP*, *CXCR2*, *CCL21*, *CCL23*, *TNFRSF1B*, *IL1R2*, *TNFSF13*, *TNFSF13B*, *TNFRSF21*, *CXCL16*, *CCR2*). In NK cells fewer specific genes were detected than the Myeloid population (Table 3), with genes such as *OTOP2*, *OTOP3*, *OSPBL3*, *LY6D*, *RET* related to cytotoxic activity, a typical characteristic of NK cells (Rusmini et al., 2013; Rusmini et al., 2014; Belizário et al., 2018; Costanzo et al., 2018; Tu et al., 2018; Upadhyay, 2019) , although their function in porcine NK cells is unexplored. In CD21pB cells, the gene for CD21 (*CR2*) used for sorting the B-cell populations was predicted to be a HEG. The SWC6gdT population showed specific expression of *AVCR2A*, which is a Th17 cell specific gene in mice (Ihn et al., 2011) and regulates the proliferation of γδ T-cells in murine skin (Antsiferova et al., 2011). The CD8T population specifically expressed *TMIGD2* (a CD28 family member) and *JAML*, which encode T-cell transmembrane proteins (Zhu et al., 2013; Alvarez et al., 2015; Krueger et al., 2017).

Finally, we compared pair-wise transcriptome differences between our porcine sorted CD4T and CD8T populations (Supplementary File 2) with the comparable populations from a previous study (Foissac et al., 2019). Even though the sorting approaches were different, 85% of the genes more highly expressed in in CD4T compared to CD8T, respectively, were detected by Foissac and colleagues in their respective $CD4^+$ to $CD8^+$ comparison. Similar overlap was found (87%) for the genes more abundant in the “CD8+ high” list, while little overlap was found in the

inverse comparisons (2.5% and 1%), strongly indicating these cell type gene expression patterns were similar between studies. However, given the lack of identification of cell-type specific genes for CD4T and CD8T populations, shared gene expression patterns may not be surprising.

NanoString assay validated bulkRNA-seq

RNA abundance of each gene target (Supplementary File 5) in each sample was used to perform a hierarchical clustering analysis (Supplementary Figure 6). Similar to relationships observed in the bulkRNA-seq dataset, biological replicates clustered most closely together. T-cell populations (SWC6gdT, CD4T, CD4CD8T, CD8T) were more similar to each other than to other populations, with the exception of NK cells. RNA abundance for the genes encoding the marker proteins used for sorting cell populations confirmed cell identity in NanoString assays (Supplementary Figure 7). RNA abundance for each tested gene and cell population is included in Supplementary File 5. To validate gene expression levels calculated by bulkRNA-seq, a Spearman rank correlation analysis was performed between expression values determined by bulkRNA-seq and NanoString (Supplementary Figure 8). Highly significant and strong correlation (rho=0.62-0.88, p-value<2.2e-16) was observed for all sorted cell types (Supplementary File 4). Overall, gene expression estimates in the bulkRNA-seq dataset were confirmed by using the NanoString assay.

Defining the transcriptomic landscape of porcine PBMCs at single-cell resolution

Single-cells from PBMCs of seven conventional pigs were partitioned, sequenced, clustered, and visualized (Supplementary File 6). In total, the final dataset included 28,810 cells expressing from 9,176-12,683 genes, and each cell was assigned to one of 36 transcriptionally-

590 distinct clusters (Figure 4A, Supplementary Figure 9A-C; Supplementary File 6). For
591 identification of general cell types in each cluster, expression levels of genes known to be active
592 in distinct porcine immune cell populations were mapped across single-cell clusters (Figure 4B-
593 C). The 36 clusters were deduced to 13 general cell types (Figure 4D) as described below.
594 Monocyte clusters (13, 19, 20, 25, 27) expressed *CSF1R* and genes associated with
595 microbial recognition (*CD14, CD163, NLRP3, TLR4*), reported as highly expressed by porcine
596 monocytes (Auray et al., 2016). DC clusters (30, 32) expressed porcine pan-DC marker *FLT3*
597 and were further classified as conventional DCs (cDCs; cluster 30) by elevated expression of
598 *FCER1A* and MHCII-encoding genes (*SLA-DRB1*[*], *SLA-DRA*[*]) and pDCs (cluster 32) by
599 elevated expression of *TCF4, XBP1, CLEC12A, CD93, IRF8, CD4,* and *CD8B* (Auray et al.,
600 2016). Co-stimulatory gene *CD86* was expressed by all monocyte and DC clusters as reported
601 (Auray et al., 2016). *SIRPA*[*], encoding CD172α is expressed by porcine monocytes/DCs (Piriou-
602 Guzylack and Salmon, 2008; Auray et al., 2016) and used to sort myeloid leukocytes for
603 bulkRNA-seq above, was minimally expressed in DC clusters.
604 B-cell clusters (2, 7, 8, 10, 11, 15, 16, 23, 26, 33) expressed *CD79A, CD19,* and *PAX5*
605 (Faldyna et al., 2007; Piriou-Guzylack and Salmon, 2008; Bordet et al., 2019). Antibody-
606 secreting cells (ASCs; cluster 29) expressed *IRF4* and *PRDM,* genes ascribed to immunoglobulin
607 secretion (Shi et al., 2015; Liu et al., 2020). Detection of *CR2*[*], the gene encoding CD21 protein,
608 was very low in any cluster.
609 Expression of *CD3E*, which encodes pan-T-cell CD3ε protein, identified T-cell clusters
610 (0, 3, 4, 5, 6, 9, 12, 14, 17, 18, 21, 22, 24, 28, 31) (Gerner et al., 2009a). Cluster 1 cells largely
611 lacked *CD3E, CD5,* and *CD6* expression, while expressing *CD2, CD8A*, *PRF1,* NK receptor-

[*] Refer to gene name replacement in Methods section

612 encoding genes *KLRB1* (CD161) and *KLRK1* (NKG2D), and NK receptor signaling adaptor
613 molecules *HCST* (DAP10) and *TYROBP* (DAP12), corresponding to a NK cell designation
614 (Denyer et al., 2006; Piriou-Guzylack and Salmon, 2008; Gerner et al., 2009a; Toka et al., 2009).
615 γδ T-cells were identified by *TRDC* expression, encoding the γδTCR δ chain, and were
616 subdivided into two major subtypes based on presence/absence of *CD2* expression (Stepanova
617 and Sinkora, 2013; Sedlak et al., 2014) (Piriou-Guzylack and Salmon, 2008; Gerner et al.,
618 2009a). Clusters 6 and 21 were identified as $CD2^-$ γδ T-cells and clusters 24 and 31 as $CD2^+$ γδ
619 T-cells. Clusters expressing *CD3E* but not *TRDC* were considered αβ T-cells and were further
620 subdivided based on *CD4* expression (0, 3, 4, 28 classified as $CD4^+$ αβ T-cells) or *CD8A* and
621 *CD8B* expression (9, 12, 14, 18, 22 classified as $CD8\alpha\beta^+$ αβ T-cells) (Piriou-Guzylack and
622 Salmon, 2008; Gerner et al., 2009a). Clusters 5 and 17 were more difficult to fully classify and
623 likely represented a mixture of cells, with some but not all cells expressing *CD3E*. Cells in
624 clusters 5 and 17 largely lacked expression of *CD5*, *CD6, TRDC, CD4,* and *CD8B* but did largely
625 express *CD2, CD8A*, *KLRB1*, and *KLRK1* and were therefore characterized as a mixture of
626 $CD8\alpha^+$ αβ T- and NK cells.

627 Cells in cluster 34 could not be characterized well enough to broadly classify as myeloid,
628 B, T, or NK lineage leukocytes based on the porcine cell markers described and remained
629 unclassified. Cluster 35 expressed *HBM* and *AHSP*, indicating erythrocytes. Clusters 34 and 35
630 were still included in further scRNA-seq analyses; however, results pertaining to these clusters
631 were not discussed.

632

633 **Gene signatures of bulkRNA-seq populations had limitations in resolving single-cell**
634 **identities**

Gene set enrichment analyses (GSEA) using SEG lists defined at different levels of enrichment for each sorted bulkRNA-seq population (Supplementary File 3, see Methods) was performed to identify scRNA-seq clusters were likely represented (Figure 5A-B, Supplementary Figure 10A, Supplementary File 8). Some gene sets had high relative enrichment in anticipated corresponding scRNA-seq clusters, such as Myeloid gene sets to monocyte/DC clusters, CD21nB/CD21pB gene sets to B-cell clusters, and SWC6gdT gene sets to $CD2^-$ γδ T-cell clusters. Interestingly, highest relative enrichment (2.51) for the top 1% of CD21nB SEGs was noted for ASCs in cluster 29, followed by erythrocytes in cluster 35 (1.68). Within sorted NK and T-cell populations, some gene sets showed high relative enrichment for their anticipated corresponding clusters in the scRNA-seq dataset. We also noted off-target relative enrichment for gene sets in clusters not anticipated to be included in specific sorted cell populations. Cluster 28 had lower relative enrichment for CD4T and CD4CD8T SEG lists at top 5-25% SEG levels (-0.02 to 0.73) than did several non-$CD4^+$ αβ T-cell clusters. Similar phenomena were observed for CD8T top 5-25% SEG lists, whereby clusters 1, 24, and 31 had higher relative enrichment for CD8T SEG lists (0.69 to 1.56) than did clusters 14 or 18 (-0.04 to 0.95 relative enrichment) that were anticipated to be included in the CD8T population. Clusters 24 and/or 31 showed off-target relative enrichment for all T/NK gene sets to various degrees, though these cells would not be expected to make up a sizeable portion of any of those sorted cell populations.

Further comparison of porcine bulk and scRNA-seq data by CIBERSORTx deconvolution analysis largely supported our single-cell cluster designations by predicting which clusters proportionally represented the bulk RNA-seq data. (Supplemental Figure 10 B, Supplemental File 7). Several clusters with poor AUCell enrichment for anticipated bulkRNA-seq gene sets in Figure 5A-B, such as cluster 28, were predicted to constitute considerable

proportions of their anticipated cell populations by CIBERSORTx deconvolution analysis. Additionally, clusters that demonstrated off-target enrichment by AUCell analysis, such as clusters 1, 9, 22, 24, and 31, were not predicted to be largely present in those off-target populations using CIBERSORTx. However, CIBERSORTx failed to predict many single-cell clusters to have notable abundances in any bulkRNA-seq populations, such as clusters 8, 19, 26, 32, and 34 having < 3.33% predicted abundance for any one bulkRNA-seq sample.

Additional GSEA comparing gene sets derived from public bulkRNA-seq data of sorted human PBMC populations with porcine single-cell gene expression profiles informed cluster identity as it relates to human immune cells (Figure 5C, Supplementary Figure 10C-D, Supplementary File 9). High relative enrichment for human monocyte gene sets in porcine monocyte populations, human CD123PmDC gene sets in porcine cDCs, and human pDC gene sets in porcine pDCs was observed, in general consensus with gene expression profiles of anticipated corresponding porcine single-cell clusters. NaiveB gene signatures had positive relative enrichment in all porcine B-cell clusters except cluster 33 at both the 0.5 and 1.0 resolution level, while the MemoryB signature had highest relative enrichment scores for B and ASC clusters at the 0.5 level, with little relative enrichment at the 1.0 level (likely due to a limited number of genes in the gene set). Human T/NK gene sets had off-target enrichment very similar to patterns observed in GSEA with porcine gene sets. Overall, GSEA between human bulkRNA-seq gene signatures and gene expression profiles of porcine scRNA-seq data supported many of the same findings when comparing between porcine bulkRNA-seq gene sets and gene expression profiles of porcine scRNA-seq data. Results indicated limitations of gene profiles obtained from sorted bulkRNA-seq populations in accurately describing/accounting for transcriptional heterogeneity resolved by scRNA-seq.

Integration of porcine and human scRNA-seq datasets to further annotate porcine cells

We examined porcine single-cell identities by comparing the porcine scRNA-seq data to a highly annotated scRNA-seq dataset of human PBMCs, providing a higher level of resolution than available with bulkRNA-seq. Transfer of more highly-specified human cell type labels onto porcine cells could reveal the most likely human counterparts for these porcine populations. Mapping scores were further calculated to determine how well porcine cells were truly represented by the human dataset (Figure 6A, Supplementary Figure 11A-B, Supplementary File 10.

Many porcine clusters had >95% of cells mapping to a specific human cell type, with average mapping scores >0.9, including monocyte, pDC, cDC, and ASC clusters, suggesting high congruency between pig and human for these cell types (Figure 6B). All porcine B-cell clusters, omitting cluster 33, mapped primarily to human B-cell clusters, but average mapping scores were slightly lower (0.80-0.87), indicating less ideal representation in the human data. In addition, every porcine B-cell cluster had overlap with all three human B-cell types (Figure 6A). Of the porcine $CD4^+$ αβ T-cells, most cluster 0 cells were predicted as human CD4 naïve cells, clusters 3 and 4 cells as human CD4 T central memory (TCM) cells, and cluster 28 cells as human CD4 proliferating cells. From porcine $CD8\alpha\beta^+$ αβ T-cells, clusters 14 and 18 were largely assigned as human replicating cell types, while 90% of cluster 9 cells were predicted as human CD8 T effector memory (TEM) cells. Highest cluster 12 predictions were mainly to human CD4/CD8 naïve T-cells, and cluster 22 cells predicted to match a range of human cell populations, with the largest percentage predicted as human CD8 TEMs. Porcine $CD8\alpha^+$ αβ T/NK and NK clusters had predictions split primarily across human CD8 TEM and NK

704 designations. Porcine $CD2^+$ γδ T-cell clusters 24 and 31 had 74% and 98%, respectively, of cells
705 predicted as human CD8 TEM, NK, or γδ T-cells. Porcine $CD2^-$ γδ T-cell clusters 6 and 21 had
706 the majority of cells predicted as human CD4 TCM, innate lymphoid cell (ILC), or γδ T-cells,
707 though the average mapping scores were lower for those assigned as CD4 TCM (0.73-0.74) or
708 gdT (0.74-0.78) than those assigned as ILCs (0.82-0.83) (Supplementary File 10). Overall, cross-
709 species comparison to a well-annotated human scRNA-seq dataset helped elucidate porcine cell
710 type identities at a higher resolution than porcine or human bulkRNA-seq datasets (Figure 5),
711 though some discordance was clearly still present.

712 Several porcine clusters had low mapping scores to a human cell type, indicating the
713 porcine cells may not be well represented by the human reference dataset (Figure 6B and
714 Supplementary File 10). Therefore, *de novo* visualization was performed on the combined human
715 and porcine data, to identify cells in the pig dataset not well represented in the human data
716 (Figure 6C-D). Porcine clusters could be identified that had low similarity to human cells, and
717 vice versa (Figure 6C). Specifically, porcine clusters 6, 16, 21, and 33 weakly overlapped human
718 cells in the two-dimensional *de novo* visualization (compare 6C and 6D). Furthermore, clusters
719 6, 16, 21, and 33 had lower average mapping scores to any human cell type (Figure 6B).
720

721 **Different activation states of porcine $CD4^+$ αβ T-cells based on CD8α expression**

722 We further compared scRNA-seq gene expression profiles amongst only $CD4^+$ αβ T-cell
723 clusters to gain functional inferences and correspondence to $CD8\alpha^-$ versus $CD8\alpha^+$ phenotypes
724 that were used to sort $CD4^+$ αβ T-cells for bulkRNA-seq. $CD4^+$ αβ T-cell clusters (0, 3, 4, 28)
725 were comprised of 5,082 total cells (Figure 7A). Hierarchical clustering and pairwise DGE
726 (Supplementary File 7), as well as random forest (RF) analyses, a deep-learning classification

727 method, (see Methods; Supplementary File 11), cumulatively revealed clusters 3 and 4 to be the
728 most transcriptionally similar to each other. Clusters 3 and 4 had the smallest hierarchical
729 distance, fewest DEGs (67), and largest RF error rate (19.5) between them, while cluster 28 was
730 the most distantly related to the other 3 clusters (Figure 7B).

731 *CD8A* gene expression was detected in a subset of cells in the $CD4^+$ αβ T-cell clusters
732 (3.5%, 13.1%, 20.9%, 39.7% of cells in clusters 0, 3, 4, 28, respectively; Figure 7C). *CD8A*
733 expression was significantly greater in clusters 4 and 28 compared to cluster 0 by pairwise DGE
734 analyses (Supplementary File 7) but not in cluster 3 compared to 0, due to not meeting a
735 minimum threshold of cells (20%) expressing the gene in either cluster implemented for DGE
736 analysis. However, cluster 3 had significantly greater expression of *CD8A* compared to cluster 0
737 when removing the minimum cell expression threshold (average log2FC =0.37, adjusted p-value
738 =5.52x10-21). GSEA of DEGs identified by pairwise DGE analysis of CD4T and CD4CD8T
739 populations recovered from bulkRNA-seq (Supplementary File 2) revealed genes significantly
740 enriched in CD4T compared to CD4CD8T populations were relatively enriched in cluster 0,
741 while genes significantly enriched in CD4CD8T compared to CD4T populations showed greater
742 relative enrichment in clusters 4 and 28 and to a lesser extent in cluster 3 (Figure 7D,
743 Supplementary File 12).

744 The top genes contributing to overall transcriptional heterogeneity amongst four clusters
745 of $CD4^+$ αβ T-cells, as determined by RF analysis (Figure 7 E-F, Supplementary File 13), highly
746 overlapped with genes identified in overall DGE analysis (Figure 7G, Supplementary File 13).
747 Of eight genes with mutually highest permutation and impurity scores from overall RF analysis
748 (Figure 7E-F), one gene had significantly greater expression in cluster 0 compared to all other
749 clusters (*RPS3A*), while the other seven genes had significantly greater expression in clusters 3,

4, and 28 compared to cluster 0 (*FCGR3A*[*]*,TMSB10, COX1, S100A6, GPX1, CRIP1*[*]*, S100A11*), as determined by pairwise DGE analyses (Supplementary File 7).

Genes associated with a naïve phenotype, including *CCR7, SELL, LEF1,* and *TCF7* (Szabo et al., 2019; Kim et al., 2020) had significantly increased expression in cluster 0 (Figure 7G, Supplementary File 9 and 13), in line with the result obtained by comparing to human scRNA-seq data that indicated a good alignment of cluster 0 with human naïve CD4 T-cells (Figure 6A). From Figure 6A, clusters 3 and 4 aligned with human CD4 Tcm (central memory) cells, and cluster 28 aligned with human CD4 proliferating cells. Correspondingly, genes associated with activation, such as *ITGB1, CD40LG, IL6R,* and MHC II-associated genes (*CD74, SLA-DRA, SLA-DQB1, SLA-DRB1*[*]*, SLA-DQA1*[*]) (Grewal and Flavell, 1996; Gerner et al., 2009b; Zemmour et al., 2018; Zhu et al., 2020) had significantly greater expression in clusters 3, 4, and/or 28, and cluster 28 expressed many genes specific for cellular replication and division (*PCLAF, BIRC5, TK1, PCNA*) (Dabydeen et al., 2019; Giotti et al., 2019) (Figure 7G, Supplementary File 9 and 13). Overall, we leveraged single-cell gene expression profiles to confirm likely identity of cluster 0 as naïve $CD4^{+}CD8\alpha^{-}$ αβ T-cells and clusters 3, 4, and 28 as potentially previously activated $CD4^{+}CD8\alpha^{+}$ αβ T-cells.[1]

Heterogeneity between/amongst $CD2^{+}$ and $CD2^{-}$ γδ T-cells

Clusters predicted to be porcine γδ T-cells were examined to reveal transcriptional distinctions within this cell type. Four clusters containing 2,652 cells were previously identified as $CD2^{-}$ γδ T-cells (clusters 6, 21) or $CD2^{+}$ γδ T-cells (clusters 24, 31) (Figure 8A). We could

[*] Refer to gene name replacement in methods

771 further segregate these clusters by *CD2* and *CD8A* expression into $CD2^{-}CD8\alpha^{-}$ (clusters 6, 21),
772 $CD2^{+}CD8\alpha^{-}$ (cluster 24), and $CD2^{+}CD8\alpha^{+}$ (cluster 31) designations used to functionally define
773 porcine γδ T-cells previously (Stepanova and Sinkora, 2013; Sedlak et al., 2014) (Figure 8B).
774 $CD2^{-}$ γδ T-cell clusters 6 and 21 were most closely related to one another by hierarchical
775 clustering, had the fewest pairwise DEGs (30), and had the highest pairwise RF analysis error
776 rate (23.5), indicating clusters 6 and 21 to be the most transcriptionally similar γδ T-cell clusters
777 of the four clusters (Figure 8C, Supplementary File 7 and 14). $CD2^{+}$ γδ T-cell clusters 24 and 31
778 were most similar to each other by hierarchical clustering, had the second fewest pairwise DEGs
779 (236), and had the second highest pairwise RF error rate (5.12), indicating clusters 24 and 31 to
780 be most similar to each other. When performing pairwise comparison between any $CD2^{-}$ and
781 $CD2^{+}$ γδ T-cell clusters, the number of DEGs increased and RF error rates decreased, indicating
782 greater transcriptional differences between cells of the $CD2^{-}$ and $CD2^{+}$ γδ T-cell lineages than
783 amongst them (Figure 8C, Supplementary File 7 and 14).
784 The top genes contributing to overall transcriptional heterogeneity amongst γδ T-cell
785 clusters, as determined by RF analysis (Figure 8D-E, Supplementary File 14), overlapped with
786 genes identified with significant and highest logFC expression in overall DGE analysis (Figure
787 8F, Supplementary File 14). Six of the top seven genes with mutual highest impurity (the best
788 features that correctly split the data) and permutation scores from RF analysis (Figure 8D-E)
789 were also DEGs between both $CD2^{-}$ compared to both $CD2^{+}$ γδ T-cell clusters by pairwise DGE
790 analysis (Supplementary File 7), again indicating large transcriptional differences between $CD2^{-}$
791 and $CD2^{+}$ γδ T-cells. In total, 31 genes had significantly greater expression in both $CD2^{-}$
792 γδ T cell clusters compared to both $CD2^{+}$ γδ T cell clusters, and 49 genes had significantly

793 greater expression in both $CD2^+$ γδ T cell clusters compared to both $CD2^-$ γδ T cell clusters
794 (Table 4), as determined using the pairwise DGE analyses (Supplementary File 7).
795 Intra-lineage heterogeneity of $CD2^-$ γδ T-cells (between clusters 6 and 21) and $CD2^+$ γδ
796 T-cells (between clusters 24 and 31) demonstrated additional complexity beyond the inter-
797 lineage heterogeneity between $CD2^-$ and $CD2^+$ γδ T-cells. Pairwise comparison between clusters
798 24 and 31 (Supplementary Data 8) revealed 80 genes with significantly greater expression in
799 cluster 24 ($CD2^+CD8\alpha^-$ γδ T-cells) and 156 genes with significantly greater expression in cluster
800 31 ($CD2^+CD8\alpha^+$ γδ T-cells). Genes with the greatest logFC expression (logFC > 1.5) in cluster
801 31 compared to cluster 24 were related to cellular activation and/or effector functions (*CCL5,*
802 *GNLY, FCGR3A*[*], KLRK1, GZMA*[*], NKG7, FCER1G, GZMB*) (Rincon-Orozco et al., 2005;
803 Pizzolato et al., 2019; Szabo et al., 2019). Of the 30 DEGs between clusters 6 and 21
804 (Supplementary Data 8), three genes had significantly greater expression in cluster 6, while 27
805 genes had significantly greater expression in cluster 21. Several genes with greater expression in
806 cluster 21 encoded for activation- or stress-induced molecules, including *GPX1, LGALS1,*
807 *ITGB1, LTB,* several genes encoding for S100 proteins (*S100A4, S100A6, S100A10, S100A11*),
808 and genes related to MHCII presentation (*CD74, SLA-DRA*[*]) (Blaser et al., 1998; Ware, 2005;
809 Gerner et al., 2009b; Steiner et al., 2011; Kesarwani et al., 2013; Siegers, 2018). Genes encoding
810 transcriptional regulators playing important roles in cell fate determination, including *ID3* and
811 *GATA3,* had greater expression in cluster 6, while *ID2* expression was significantly greater in
812 cluster 21 (Blom et al., 1999; Zhang et al., 2014; Rodríguez-Gómez et al., 2019).[1]
813

[*] Refer to gene name replacement in methods

DISCUSSION

We present the first comprehensive annotation of the global transcriptome of all major circulating porcine blood mononuclear cells. We applied bulkRNA-seq to determine transcriptomes of eight sorted PBMC populations and scRNA-seq to annotate transcriptomic diversity of PBMCs into transcriptionally-distinct clusters. Deep RNA sequencing detected significant heterogeneity between sorted populations except for T cell populations, while further heterogeneity was unmasked by scRNA-seq. Collectively, the data sets revealed specific immune functional expression patterns and highlighted substantial diversity in some subsets, such as T-cells. The combined approach helps to unite porcine transcriptomics and cellular immunology, as transcriptional differences and functional relationships of porcine immune cells have remained unclear due to lack of sufficient reagents to label distinct porcine immune cell populations. While cross-species comparisons have been done with many RNA-seq datasets of partially purified cell populations (Kapetanovic et al., 2013; Herrera-Uribe et al., 2020), our new porcine data demonstrates global similarity to human bulkRNA-seq and scRNA-seq transcriptomes that can be used to further unravel porcine cell function and extend comparative immune investigation.

Gene expression patterns from the bulkRNA-seq datasets revealed distinct transcript profiles enriched in biological pathways characteristic of each respective cell population, based on previous findings in pig and other species (Alter et al., 2004; Palmer et al., 2006; Wang et al., 2008; Foissac et al., 2019; Monaco et al., 2019; Summers et al., 2020). However, bulkRNA-seq data from the porcine sorted populations had limited ability to identify genes with specific transcriptional patterns for some sorted lymphocyte populations. The transcriptomes of eight different cell types we provide include three types of transcriptomes that have not reported

837 before in pig, including NK, CD21pB and CD21nB. Lists of SEGs, pairwise DGE between all
838 populations and cell type-specific genes data sets presented here, could be used for further
839 analysis in other pig or even in cross-species comparisons. Notably, we were able to identify a
840 large number of HEGs in the Myeloid population. Some HEGs in Myeloid cells were reported as
841 a Myeloid cell markers in pig (e.g. *CD14* and *CD36*) (Fairbairn et al., 2013) and other HEGs
842 may be considered as new potential cell markers. Also, in comparison to sorted CD4T and CD8T
843 populations reported in a previous porcine RNA-seq study (Foissac et al., 2019), we observed
844 concordant transcriptional patterns in essentially equivalent populations. However, we extended
845 transcriptional annotation to two additional T-cell populations (CD4CD8T, SWC6gdT), thus
846 identifying transcriptional differences across more T-cell populations. We demonstrated the
847 utility of an established NanoString CodeSet (Van Goor et al., 2020; Dong et al., 2021) to
848 validate RNA-seq results and further profile porcine sorted PBMC populations. At the bulk
849 RNAseq level, we concluded substantial transcriptional heterogeneity was present across sorted
850 T-cell and B-cell populations, as fewer enriched or cell type-specific genes were detected. As
851 described below, the lack of identification of cell type-specific genes was likely caused by the
852 lack of further sub-setting during sorting to separate functionally distinct cells. However, we
853 were able to find several specific transcriptional patterns in B- and T-cells using bulkRNA-seq,
854 and some of the identified genes encode for transmembrane proteins. Beyond further description
855 of well-annotated genes, we also demonstrated that up to 18% of our predicted cell-type specific
856 and enriched genes are currently poorly annotated, i.e., genes with no recognized human
857 ortholog. These data thus increase the functional annotation of these genes, as co-expression
858 patterns linking such genes with known genes can be an important component for Gene

Ontology classifications and disease-association gene prediction (van Dam et al., 2018), and is an important proposed outcome of the FAANG project (Giuffra et al., 2019).

Comparison of our sorted population expression patterns to a similar human RNA-seq dataset revealed both similarities and differences between species. While we compared the transcriptomes of the sorted cells with human populations that were isolated using similar cell markers, we cannot exclude that we are biasing this comparison due to different immunoreagent markers used across species. However, we did find similar transcriptional patterns across immune cell populations that are intrinsic to a lineage, such as the porcine Myeloid population correlating with the human myDC123 population, in agreement with other studies (Auray et al., 2016).

Previous global gene expression studies using either porcine whole blood or specific immune cell types have failed to thoroughly describe all major PBMC populations (Freeman et al., 2012; Dawson et al., 2013; Mach et al., 2013; Auray et al., 2016; Foissac et al., 2019). Providing the transcriptomes of bulk sorted cell populations will be readily useful to the majority of porcine immunology research labs that use sorting techniques to analyze porcine immune cell function and RNA expression patterns, as new lists of co-expressed genes in these cell populations are now available. However, our combined analysis of such bulkRNAseq data with the scRNAseq data demonstrated that the former approach has significant heterogeneity, limiting the ability to resolve specific cell types for deeper transcriptional interrogation. A combined analysis provided evidence confirming our hypothesis that scRNA-seq would lead to identification of more specific and novel transcriptional signatures to improve annotation and understanding of circulating porcine immune cells.

881 scRNA-seq provides many noted benefits in transcriptomic analysis, however there are
882 limitations to the approach. Of benefit, scRNA-seq captured transcriptomes of cells excluded
883 from our bulkRNA-seq analysis, as scRNA-seq approach did not rely on protein marker
884 expression and selection of sorting criteria based on specific marker phenotypes. As mentioned
885 above, scRNA-seq also established that greater levels of cellular heterogeneity exist, since
886 sequencing was resolved to the level of individual cells rather than a sorted population. We
887 recognize the scRNAseq-predicted clusters may contain transitory cell states that may be very
888 challenging to further study for the relationship between cellular function and transcriptional
889 patterns (Bassler et al., 2019). Further, we assumed single-cell gene expression profiles would be
890 indicative of protein expression for cell type-specific markers; however, gene expression for
891 many such markers, including *SIRPA*[*] and *CR2*[*] that encode proteins used for bulk RNA-seq cell
892 sorting, was sparse. Sparsity of data is a known limitation of the scRNA-seq approach utilized
893 herein, while methods such as imputation have been proposed to improve sensitivity (Andrews et
894 al., 2021). We chose not to use imputation due to our current inability to estimate effects on cell
895 patterns through comparison to an external reference (Andrews et al., 2021). Thus, these
896 limitations made it difficult to decipher between low- and non-expression for some genes of
897 interest, including canonical markers used for identifying cell types in the immunology literature.
898 Instead, reliance on gene expression profiles of multiple markers was used. For example, *SIRPA*[*]
899 expression was observed at low levels in monocyte clusters but was virtually absent in DC
900 clusters, though both porcine monocytes and DCs express CD172A protein. Because DCs
901 express CD172A at lower levels than monocytes (Piriou-Guzylack and Salmon, 2008; Auray et
902 al., 2016), *SIRPA*[*] expression in DCs may have been below our limit of detection using scRNA-

[*] Refer to gene name replacement in Methods section

seq, as it was insufficiently expressed in DCs but not in monocytes. We utilized a droplet-based partitioning method for scRNA-seq that can detect a large number of cells but a lower number of transcripts per cell. By this method, we could retain a large number of cells (>25,000 cells from seven samples) at the expense of limited sequencing depth per cell (minimum of 500 unique genes and 1,000 unique transcripts per cell). Utilizing higher sequencing depth per cell or different partitioning platforms for scRNA-seq that have more efficient transcript capture per cell will be beneficial for deeper analysis of specific cells/genes of interest. It is likely some gene expression profiles are not predictive of protein expression, due to post-transcriptional regulation mechanisms. Using newly available co-expression lists to formulate more refined cell sorting regimens and scRNAseq analysis of such sorted populations will also increase the ability to define transcriptomes of such cell types (Nestorowa et al., 2016). It was notable that the lists of genes predicted to be significantly enriched in the 36 scRNAseq clusters had overall a very similar fraction of poorly annotated genes (average of 18%; cite in Supplementary File 6) to those predicted for bulkRNAseq, indicating that even the genes with expression patterns predicted to be more discriminatory contribute a similar level of genome annotation improvement.

We used multiple methods to compare these high-dimensional expression datasets to further interpret genes predicted to be different between sorted cell populations, between clusters, or between human and pig. GSEA and/or deconvolution analyses of bulkRNA-seq to scRNA-seq datasets was only partially effective in correlating sorted populations with assumed corresponding clusters in the scRNA-seq dataset (regardless of inter-species or intra-species comparison). At a higher level of resolution, both methods were able to assign most corresponding cell-type designations between scRNA-seq and bulkRNA-seq data. However,

several different scRNA-seq clusters were not predicted to make up a large portion of any bulkRNA-seq sample. While methodology could account for these differences, it is more likely that CIBERSORTx was unable to discriminate between certain clusters due to their high similarity. For example, cells that could have been predicted to be assigned to cluster 8, which makes up a large proportion of the scRNA-seq data, may have been assigned to other similar B cell clusters. The ability to discriminate between similar clusters may have been impacted by down sampling each cluster to include the same number of cells for the analysis. Overall, deconvolution was useful in assigning cell type level data but in some instances, it could not fully deconvolute bulk RNAseq to the cluster specific level.

Integration of porcine PBMC scRNA-seq with a human PBMC scRNA-seq dataset did allow further resolution of porcine cluster annotations and yielded high confidence of homology between many porcine and human single cell populations. While we cannot completely discount the potential for recognized cell types in our scRNA-seq dataset are not present in sorted populations used for bulkRNA-seq (or vice-versa), it seems more likely this is similar evidence to that described above indicating that the same level of resolution simply was not captured by bulkRNA-seq and could not well represent all cell types found in the scRNA-seq data. Integration with another scRNA-seq dataset, even when accounting for cross-species comparison, was in many ways more informative for further annotating porcine single cells, highlighting the enhanced ability of scRNA-seq to define cellular landscapes. Moreover, cross-species integration extended our knowledge of comparative immunology between humans and pigs, as we could identify most similar human counterparts by reference-based prediction. Conversely, we could also identify clusters of $CD2^-$ γδ T-cells (clusters 6 and 21) and B-cells (clusters 16 and 33) that were largely specific to the porcine dataset by *de novo* visualization of

949 clustered cells using the combined human and pig data. The clusters of cells in pig samples either
950 represent porcine cells either lacking close human cellular counterparts or the equivalent human
951 counterparts were not well-captured in the human PBMC scRNA-seq dataset.

952 While we did not perform deeper biological query of all cell types identified in our
953 scRNA-seq dataset, we did attempt to deduce biological significance for the different $CD4^+$
954 αβ T-cell populations that have unique aspects in pigs. Deeper query of $CD4^+$ αβ T-cells was
955 performed, as there is functional interest in determining activation states of porcine $CD4^+$ αβ T^-
956 cells based on CD8α expression, which may be gained upon activation and retained in a memory
957 state (Summerfield et al., 1996; Zuckermann, 1999; Saalmüller et al., 2002; Gerner et al.,
958 2009b). We found it difficult to identify $CD4^+$ αβ T-cell clusters as $CD8\alpha^+$ or $CD8\alpha^-$ due to
959 sparsity in *CD8A* expression but could leverage comparison of CD4T and CD4CD8T
960 populations from bulkRNA-seq to formulate gene sets enriched in each CD4 expressing T-cell
961 population. GSEA helped identify one cluster of $CD4^+CD8\alpha^-$ αβ T-cells that corresponded
962 mostly to human naïve CD4 T-cells, while three clusters of $CD4^+CD8\alpha^+$ αβ T-cells
963 corresponded to human memory or proliferating CD4 T-cells. Collectively, these data reinforce
964 previous porcine literature, elucidate parallels to human cells, and provide greater insight into the
965 spectrum of activation states present in $CD4^+CD8\alpha^+$ αβ T-cells. Future analysis of activated T-
966 cells or trajectory analysis may provide even further insight on the transition of activation states
967 in porcine peripheral T-cells.

968 Pigs are a 'γδ high' species, named as such because they have a higher proportions of γδ
969 T-cells in circulation, largely attributed to the presence of $CD2^-$ γδ T-cells that are absent in
970 humans and mice (Stepanova and Sinkora, 2013). Three major γδ T-cell populations are
971 characterized in pigs: $CD2^-CD8\alpha^-$ γδ T-cells that express SWC6 and $CD2^+CD8\alpha^{-/+}$ γδ T-cells

that do not express SWC6, where $CD2^{-}CD8\alpha^{-}$ γδ T-cells become $CD2^{+}CD8\alpha^{+}$ upon activation (Stepanova and Sinkora, 2013; Sedlak et al., 2014). As our sorting strategy for bulkRNA-seq utilized an anti-SWC6 antibody rather than a pan-γδ T-cell-specific antibody; thus, γδ T-cells for bulk RNA-seq included $CD2^{-}CD8\alpha^{-}$ γδ T-cells in the SWC6gdT population or $CD2^{+}CD8\alpha^{+}$ γδ T-cells found in combination with $CD4^{-}CD8\alpha^{+}$ αβ T-cells in the CD8T population. $CD2^{+}CD8\alpha^{-}$ γδ T-cells were expected to be excluded in cell sorting. In future sorting strategies, it may be beneficial to utilize a pan-γδ T-cell reactive antibody and/or identify $CD4^{-}CD8^{+}$ αβ T-cells with anti-CD8β antibody, which should not label with $CD2^{+}CD8\alpha^{+}$ γδ T-cells (Gerner et al., 2009b). though this may still exclude potential $CD4^{-}CD8\alpha^{+}CD8\beta^{-}$ αβ T cells, such as we observed in clusters 5 and 17. Despite limitations in sorting, the bulkRNA-seq profiles were still informative when comparing to scRNA-seq data. The highest relative enrichment of SWC6gdT gene signatures was detected in $CD2^{-}$ γδ T-cell clusters, while $CD2^{+}$ γδ T-cell clusters showed relative enrichment to a lesser level, indicating some conserved gene expression between $CD2^{-}$and $CD2^{+}$ γδ T-cells. Comparison between $CD2^{+}$ γδ T-cell clusters further supported previous biological understanding, where $CD2^{+}CD8\alpha^{+}$ γδ T-cells had greater expression of genes related to cellular activation and cytotoxicity relative to $CD2^{+}CD8\alpha^{-}$ γδ T-cells (Yang and Parkhouse, 1997; Stepanova and Sinkora, 2013; Sedlak et al., 2014). On the other hand, $CD2^{-}$ γδ T-cells are less well described than $CD2^{+}$ γδ T-cells, largely due to lack of comparable populations in humans or mice that may be used for biological inference. Integration with human scRNA-seq data supported previous observations of the absence of $CD2^{-}$ γδ T-cells in humans, as close counterparts for $CD2^{-}$ γδ T-cell clusters could not be found by *de novo* visualization, and reference-based integration indicated closest human counterparts to be a mixture of primarily γδ T-cells, ILCs, and CD4 T_{CM}s, and mapping scores were highest for human ILCs rather than γδ T-

995 cells, indicating human ILCs to be the closest, albeit still poor, human match. Nonetheless, we
996 were able to highlight transcriptional distinctions that better annotate $CD2^-$ γδ T-cells, including
997 DEGs between $CD2^-$ and $CD2^+$ γδ T-cells that defined the two γδ T cell lineages and between
998 two clusters of $CD2^-$ γδ T-cells that have not yet been described.
999

1000 **Conclusion**

1001 This study provides a first-generation atlas annotating circulating porcine immune cell
1002 transcriptomes at both the cell surface marker-sorted population and single-cell levels. These
1003 findings illuminate the landscape of immune cell molecular signatures useful for porcine
1004 immunology and a deeper annotation of the genome, a goal of the FAANG project. These results
1005 also provide useful resources to identify new porcine cell biomarkers for discrimination and
1006 isolation of specific cell types, urgently needed in the field.
1007

1008 **Abbreviations**

1009 AUC: area under the curve
1010 ASC: antibody-secreting cell
1011 B: B-cell
1012 bulkRNA-seq: bulk RNA sequencing
1013 cDC: conventional dendritic cell
1014 DC: dendritic cell
1015 DEGs: differentially expressed genes
1016 DGE: differential gene expression
1017 Exp: experiment

FAANG: Functional Annotation of Animal Genomes

FACS: Fluorescent activated cell sorting

G2P: Genome-to-Phenome

GO: gene ontology

GSEA: gene set enrichment analysis/analyses

HBSS: Hank's balanced salt solution

HEGs: highly enriched genes

MACS: Magnetic activated cell sorting

mDC/myDC: myeloid dendritic cell

n: negative

NK: natural killer

p: positive

PBMC: peripheral blood mononuclear cell

PC: principal component

PCA: principal component analysis

pDC: plasmacytoid dendritic cell

RF: random forest

RIN: RNA integrity number

RNA-seq: RNA sequencing

scRNA-seq: single-cell RNA sequencing

scREF-matrix: single-cell reference matrix

SEG: significantly enriched genes

sPCA: supervised principal component analysis

SWC6: swine workshop cluster 6.

T: T-cell.

TCR: T-cell receptor

TPM: transcripts per million

t-SNE: t-distributed stochastic neighbor embedding

UMAP: uniform manifold approximation and projection

UMI: unique molecular identifier

γδ: Gamma-delta

αβ: alpha beta

Acknowledgments

The authors acknowledge the DNA facility of the Iowa State University for provision of technical support and sequencing platforms utilized in this study. We are grateful to the NADC animal care staff for their efforts. We thank Dr. Catherine Ernst at Michigan State University for providing the Yorkshire pigs used in the bulkRNA-seq study. We thank Samuel Humphrey for cell sorting technical assistance. We thank Kristen Walker for performing the NanoString data collection and initial analyses. We thank Zahra Bond for technical assistance with cell isolation and FACS, and Dr. Julian Trachsel for early data visualization assistance.

Author Contributions Statements

JH, JEW, KAB, and CLL collected samples and isolated cells. KAB performed cell staining and FACS. JH performed bulk RNA isolations. JL supervised the NanoString assay data collection. JH, LD, and HL performed bulkRNA-seq analyses. HL performed NanoString analyses. JEW,

LD, SKS, and HL performed scRNA-seq analyses. JH, JEW, CLL, and CKT interpreted the data and drafted the manuscript. All authors contributed to the writing of the materials and methods, edited the manuscript, and approved the final version.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as potential conflict of interest.

Funding

This work was supported by (1) the National Institute of Food and Agriculture (NIFA) Project 2018-67015-2701, (2) the NRSP-9 Swine Genome Coordination project, (3) appropriated funds from USDA-ARS CRIS project 5030-31320-004-00D, and (4) an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE- SC0014664, (5) USDA ARS CRIS Project 8042-32000-102. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of USDA, ARS, DOE, or ORAU/ORISE.

Data Availability Statement

Raw sequencing data from bulkRNA-seq and scRNA-seq are available through the European Nucleotide Archive (project: PRJEB43826), https://www.ebi.ac.uk/ena/browser/view/PRJEB43826.

REFERENCES

Alter, G., Malenfant, J.M., and Altfeld, M. (2004). CD107a as a functional marker for the identification of natural killer cell activity. *J Immunol Methods* 294(1-2), 15-22. doi: 10.1016/j.jim.2004.08.008.

Alvarez, J.I., Kébir, H., Cheslow, L., Charabati, Chabarati, M., Larochelle, C., et al. (2015). JAML mediates monocyte and CD8 T cell migration across the brain endothelium. *Ann Clin Transl Neurol* 2(11), 1032-1037. doi: 10.1002/acn3.255.

Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* 16, 57. doi: 10.1186/s13059-015-0622-4.

Andrews, T.S., Kiselev, V.Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc* 16(1), 1-9. doi: 10.1038/s41596-020-00409-w.

Antsiferova, M., Huber, M., Meyer, M., Piwko-Czuchra, A., Ramadan, T., MacLeod, A.S., et al. (2011). Activin enhances skin tumourigenesis and malignant progression by inducing a pro-tumourigenic immune cell response. *Nat Commun* 2, 576. doi: 10.1038/ncomms1585.

Arceo, M., Ernst, C.W., Lunney, J.K., Choi, I., Raney, N.E., Huang, T., et al. (2013). Characterizing differential individual response to Porcine Reproductive and Respiratory Syndrome Virus infection through statistical and functional analysis of gene expression. *Frontiers in Livestock Genomics (Manuscript in preparation)*.

Auray, G., Keller, I., Python, S., Gerber, M., Bruggmann, R., Ruggli, N., et al. (2016). Characterization and Transcriptomic Analysis of Porcine Blood Conventional and Plasmacytoid Dendritic Cells Reveals Striking Species-Specific Differences. *J Immunol* 197(12), 4791-4806. doi: 10.4049/jimmunol.1600672.

Auray, G., Talker, S.C., Keller, I., Python, S., Gerber, M., Liniger, M., et al. (2020). High-Resolution Profiling of Innate Immune Responses by Porcine Dendritic Cell Subsets. *Front Immunol* 11, 1429. doi: 10.3389/fimmu.2020.01429.

Bassler, K., Schulte-Schrepping, J., Warnat-Herresthal, S., Aschenbrenner, A.C., and Schultze, J.L. (2019). The Myeloid Cell Compartment-Cell by Cell. *Annu Rev Immunol* 37, 269-293. doi: 10.1146/annurev-immunol-042718-041728.

Bates, D., M\"a, c., Martin, Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1-48.

Belizário, J.E., Neyra, J.M., and Setúbal Destro Rodrigues, M.F. (2018). When and how NK cell-induced programmed cell death benefits immunological protection against intracellular pathogen infection. *Innate Immun* 24(8), 452-465. doi: 10.1177/1753425918800200.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57(1), 289-300. doi: 10.2307/2346101.

Berry, M.P., Graham, C.M., McNab, F.W., Xu, Z., Bloch, S.A., Oni, T., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466(7309), 973-977. doi: nature09247 [pii]
10.1038/nature09247.

1130 Blaser, C., Kaufmann, M., Müller, C., Zimmermann, C., Wells, V., Mallucci, L., et al. (1998).
1131 Beta-galactoside-binding protein secreted by activated T cells inhibits antigen-induced
1132 proliferation of T cells. *Eur J Immunol* 28(8)**,** 2311-2319. doi: 10.1002/(SICI)1521-
1133 4141(199808)28:08<2311::AID-IMMU2311>3.0.CO;2-G.
1134 Blom, B., Heemskerk, M.H., Verschuren, M.C., van Dongen, J.J., Stegmann, A.P., Bakker, A.Q.,
1135 et al. (1999). Disruption of alpha beta but not of gamma delta T cell development by
1136 overexpression of the helix-loop-helix protein Id3 in committed T cell progenitors.
1137 *EMBO J* 18(10)**,** 2793-2802. doi: 10.1093/emboj/18.10.2793.
1138 Bordet, E., Frétaud, M., Crisci, E., Bouguyon, E., Rault, S., Pezant, J., et al. (2019).
1139 Macrophage-B Cell Interactions in the Inverted Porcine Lymph Node and Their
1140 Response to Porcine Reproductive and Respiratory Syndrome Virus. *Front Immunol* 10**,**
1141 953. doi: 10.3389/fimmu.2019.00953.
1142 Byrne, K.A., Tuggle, C.K., and Loving, C.L. (2020). Differential induction of innate memory in
1143 porcine monocytes by. *Innate Immun***,** 1753425920951607. doi:
1144 10.1177/1753425920951607.
1145 Chaussabel, D., and Baldwin, N. (2014). Democratizing systems immunology with modular
1146 transcriptional repertoire analyses. *Nat Rev Immunol* 14(4)**,** 271-280. doi: nri3642 [pii]
1147 10.1038/nri3642.
1148 Chaussabel, D., Pascual, V., and Banchereau, J. (2010). Assessing the human immune system
1149 through blood transcriptomics. *BMC Biol* 8**,** 84. doi: 10.1186/1741-7007-8-84.
1150 Choi, J., Baldwin, T.M., Wong, M., Bolden, J.E., Fairfax, K.A., Lucas, E.C., et al. (2019).
1151 Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and
1152 humans. *Nucleic Acids Res* 47(D1)**,** D780-D785. doi: 10.1093/nar/gky1020.
1153 Costanzo, M.C., Kim, D., Creegan, M., Lal, K.G., Ake, J.A., Currier, J.R., et al. (2018).
1154 Transcriptomic signatures of NK cells suggest impaired responsiveness in HIV-1
1155 infection and increased activity post-vaccination. *Nat Commun* 9(1)**,** 1212. doi:
1156 10.1038/s41467-018-03618-w.
1157 Dabydeen, S.A., Desai, A., and Sahoo, D. (2019). Unbiased Boolean analysis of public gene
1158 expression data for cell cycle gene identification. *Mol Biol Cell* 30(14)**,** 1770-1779. doi:
1159 10.1091/mbc.E19-01-0013.
1160 Davis, W.C., Zuckermann, F.A., Hamilton, M.J., Barbosa, J.I., Saalmüller, A., Binns, R.M., et al.
1161 (1998). Analysis of monoclonal antibodies that recognize gamma delta T/null cells. *Vet*
1162 *Immunol Immunopathol* 60(3-4)**,** 305-316. doi: 10.1016/s0165-2427(97)00107-4.
1163 Dawson, H.D., Loveland, J.E., Pascal, G., Gilbert, J.G., Uenishi, H., Mann, K.M., et al. (2013).
1164 Structural and functional annotation of the porcine immunome. *BMC Genomics* 14**,** 332.
1165 doi: 10.1186/1471-2164-14-332.
1166 Denyer, M.S., Wileman, T.E., Stirling, C.M., Zuber, B., and Takamatsu, H.H. (2006). Perforin
1167 expression can define CD8 positive lymphocyte subsets in pigs allowing phenotypic and
1168 functional analysis of natural killer, cytotoxic T, natural killer T and MHC un-restricted
1169 cytotoxic T-cells. *Vet Immunol Immunopathol* 110(3-4)**,** 279-292. doi:
1170 10.1016/j.vetimm.2005.10.005.
1171 Dong, Q., Lunney, J.K., Lim, K.S., Nguyen, Y., Hess, A.S., Beiki, H., et al. (2021). Gene
1172 expression in tonsils in swine following infection with porcine reproductive and
1173 respiratory syndrome virus. *BMC Vet Res* 17(1)**,** 88. doi: 10.1186/s12917-021-02785-1.

1174 Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the
1175 integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*
1176 4(8), 1184-1191. doi: 10.1038/nprot.2009.97.
1177 Fairbairn, L., Kapetanovic, R., Beraldi, D., Sester, D.P., Tuggle, C.K., Archibald, A.L., et al.
1178 (2013). Comparative analysis of monocyte subsets in the pig. *J Immunol* 190(12), 6389-
1179 6396. doi: 10.4049/jimmunol.1300365.
1180 Faldyna, M., Samankova, P., Leva, L., Cerny, J., Oujezdska, J., Rehakova, Z., et al. (2007).
1181 Cross-reactive anti-human monoclonal antibodies as a tool for B-cell identification in
1182 dogs and pigs. *Vet Immunol Immunopathol* 119(1-2), 56-62. doi:
1183 10.1016/j.vetimm.2007.06.022.
1184 Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-
1185 species annotation of transcriptome and chromatin structure in domesticated animals.
1186 *BMC Biol* 17(1), 108. doi: 10.1186/s12915-019-0726-5.
1187 Freeman, T.C., Ivens, A., Baillie, J.K., Beraldi, D., Barnett, M.W., Dorward, D., et al. (2012). A
1188 gene expression atlas of the domestic pig. *BMC Biol* 10, 90. doi: 10.1186/1741-7007-10-
1189 90.
1190 Gerner, W., Kaser, T., and Saalmuller, A. (2009a). Porcine T lymphocytes and NK cells--an
1191 update. *Developmental and comparative immunology* 33(3), 310-320. doi:
1192 10.1016/j.dci.2008.06.003.
1193 Gerner, W., Käser, T., and Saalmüller, A. (2009b). Porcine T lymphocytes and NK cells--an
1194 update. *Dev Comp Immunol* 33(3), 310-320. doi: 10.1016/j.dci.2008.06.003.
1195 Gerner, W., Talker, S.C., Koinig, H.C., Sedlak, C., Mair, K.H., and Saalmuller, A. (2015).
1196 Phenotypic and functional differentiation of porcine alphabeta T cells: current knowledge
1197 and available tools. *Mol Immunol* 66(1), 3-13. doi: 10.1016/j.molimm.2014.10.025.
1198 Giotti, B., Chen, S.H., Barnett, M.W., Regan, T., Ly, T., Wiemann, S., et al. (2019). Assembly of
1199 a parts list of the human mitotic cell cycle machinery. *J Mol Cell Biol* 11(8), 703-718.
1200 doi: 10.1093/jmcb/mjy063.
1201 Giuffra, E., Tuggle, C.K., and Consortium, F. (2019). Functional Annotation of Animal
1202 Genomes (FAANG): Current Achievements and Roadmap. *Annu Rev Anim Biosci* 7, 65-
1203 88. doi: 10.1146/annurev-animal-020518-114913.
1204 Grewal, I.S., and Flavell, R.A. (1996). The role of CD40 ligand in costimulation and T-cell
1205 activation. *Immunol Rev* 153, 85-106. doi: 10.1111/j.1600-065x.1996.tb00921.x.
1206 Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W., Zheng, S., Butler, A., et al. (2020).
1207 "Integrated analysis of multimodal single-cell data". (bioRxiv).
1208 Herrera-Uribe, J., Liu, H., Byrne, K.A., Bond, Z.F., Loving, C.L., and Tuggle, C.K. (2020).
1209 Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages
1210 Following LPS or PolyIC Exposure. *Front Genet* 11, 817. doi:
1211 10.3389/fgene.2020.00817.
1212 Hicks, S.C., and Irizarry, R.A. (2015). quantro: a data-driven approach to guide the choice of an
1213 appropriate normalization method. *Genome biology* 16(1), 117-117. doi: 10.1186/s13059-
1214 015-0679-0.
1215 Hicks, S.C., Okrah, K., Paulson, J.N., Quackenbush, J., Irizarry, R.A., and Bravo, H.C. (2018).
1216 Smooth quantile normalization. *Biostatistics (Oxford, England)* 19(2), 185-198. doi:
1217 10.1093/biostatistics/kxx028.
1218 Huang, T.H., Uthe, J.J., Bearson, S.M., Demirkale, C.Y., Nettleton, D., Knetter, S., et al. (2011).
1219 Distinct peripheral blood RNA responses to Salmonella in pigs differing in Salmonella

shedding levels: intersection of IFNG, TLR and miRNA pathways. *PLoS One* 6(12), e28768. doi: 10.1371/journal.pone.0028768.

Ihn, H.J., Kim, D.H., Oh, S.S., Moon, C., Chung, J.W., Song, H., et al. (2011). Identification of Acvr2a as a Th17 cell-specific gene induced during Th17 differentiation. *Biosci Biotechnol Biochem* 75(11), 2138-2141. doi: 10.1271/bbb.110436.

Kapetanovic, R., Fairbairn, L., Downing, A., Beraldi, D., Sester, D.P., Freeman, T.C., et al. (2013). The impact of breed and tissue compartment on the response of pig macrophages to lipopolysaccharide. *BMC Genomics* 14, 581. doi: 10.1186/1471-2164-14-581.

Kesarwani, P., Murali, A.K., Al-Khami, A.A., and Mehrotra, S. (2013). Redox regulation of T-cell function: from molecular mechanisms to significance in human health and disease. *Antioxid Redox Signal* 18(12), 1497-1534. doi: 10.1089/ars.2011.4073.

Kim, C., Jin, J., Weyand, C.M., and Goronzy, J.J. (2020). The Transcription Factor TCF1 in T Cell Differentiation and Aging. *Int J Mol Sci* 21(18). doi: 10.3390/ijms21186497.

Knetter, S.M., Bearson, S.M., Huang, T.H., Kurkiewicz, D., Schroyen, M., Nettleton, D., et al. (2015). Salmonella enterica serovar Typhimurium-infected pigs with different shedding levels exhibit distinct clinical, peripheral cytokine and transcriptomic immune response phenotypes. *Innate Immun* 21(3), 227-241. doi: 1753425914525812 [pii]
10.1177/1753425914525812.

Kobayashi, E., Hanazono, Y., and Kunita, S. (2018). Swine used in the medical university: overview of 20 years of experience. *Exp Anim* 67(1), 7-13. doi: 10.1538/expanim.17-0086.

Koltes, J.E., Cole, J.B., Clemmens, R., Dilger, R.N., Kramer, L.M., Lunney, J.K., et al. (2019). A Vision for Development and Utilization of High-Throughput Phenotyping and Big Data Analytics in Livestock. *Front Genet* 10, 1197. doi: 10.3389/fgene.2019.01197.

Krueger, J., Jules, F., Rieder, S.A., and Rudd, C.E. (2017). CD28 family of receptors inter-connect in the regulation of T-cells. *Receptors Clin Investig* 4.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15. doi: 10.1186/gb-2014-15-2-r29.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6), 882-883. doi: 10.1093/bioinformatics/bts034.

Li, S., Labaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotech* 32(9), 888-895. doi: 10.1038/nbt.3000
http://www.nature.com/nbt/journal/v32/n9/abs/nbt.3000.html#supplementary-information.

Liu, G., Wang, B., Chen, Q., Li, Y., Li, B., Yang, N., et al. (2020). Interleukin (IL)-21 Promotes the Differentiation of IgA-Producing Plasma Cells in Porcine Peyer's Patches via the JAK-STAT Signaling Pathway. *Front Immunol* 11, 1303. doi: 10.3389/fimmu.2020.01303.

Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* 5. doi: 10.12688/f1000research.7223.1.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12), 550. doi: 10.1186/s13059-014-0550-8.

1265 Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., Marioni, J.C., et al. (2019).
1266 EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA
1267 sequencing data. *Genome Biol* 20(1)**,** 63. doi: 10.1186/s13059-019-1662-y.
1268 Mach, N., Gao, Y., Lemonnier, G., Lecardonnel, J., Oswald, I.P., Estellé, J., et al. (2013). The
1269 peripheral blood transcriptome reflects variations in immunity traits in swine: towards the
1270 identification of biomarkers. *BMC Genomics* 14**,** 894. doi: 10.1186/1471-2164-14-894.
1271 Maroilley, T., Lemonnier, G., Lecardonnel, J., Esquerré, D., Ramayo-Caldas, Y., Mercat, M.J.,
1272 et al. (2017). Deciphering the genetic regulation of peripheral blood transcriptome in pigs
1273 through expression genome-wide association study and allele-specific expression
1274 analysis. *BMC Genomics* 18(1)**,** 967. doi: 10.1186/s12864-017-4354-6.
1275 Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., et al. (2019). RNA-Seq
1276 Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human
1277 Immune Cell Types. *Cell Rep* 26(6)**,** 1627-1640.e1627. doi:
1278 10.1016/j.celrep.2019.01.041.
1279 Munyaka, P.M., Kommadath, A., Fouhse, J., Wilkinson, J., Diether, N., Stothard, P., et al.
1280 (2019). Characterization of whole blood transcriptome and early-life fecal microbiota in
1281 high and low responder pigs before, and after vaccination for Mycoplasma
1282 hyopneumoniae. *Vaccine* 37(13)**,** 1743-1755. doi: 10.1016/j.vaccine.2019.02.016.
1283 Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., et al.
1284 (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell
1285 differentiation. *Blood* 128(8)**,** e20-31. doi: 10.1182/blood-2016-05-716480.
1286 Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., et al. (2019).
1287 Determining cell type abundance and expression from bulk tissues with digital cytometry.
1288 *Nat Biotechnol* 37(7)**,** 773-782. doi: 10.1038/s41587-019-0114-2.
1289 Palmer, C., Diehn, M., Alizadeh, A.A., and Brown, P.O. (2006). Cell-type specific gene
1290 expression profiles of leukocytes in human peripheral blood. *BMC Genomics* 7**,** 115. doi:
1291 10.1186/1471-2164-7-115.
1292 Paulson, J.N., Chen, C.Y., Lopes-Ramos, C.M., Kuijjer, M.L., Platig, J., Sonawane, A.R., et al.
1293 (2017). Tissue-aware RNA-Seq processing and normalization for heterogeneous and
1294 sparse data. *BMC Bioinformatics* 18(1)**,** 437. doi: 10.1186/s12859-017-1847-x.
1295 Piriou-Guzylack, L., and Salmon, H. (2008). Membrane markers of the immune cells in swine:
1296 an update. *Vet Res* 39(6)**,** 54. doi: 10.1051/vetres:2008030.
1297 Pizzolato, G., Kaminski, H., Tosolini, M., Franchini, D.M., Pont, F., Martins, F., et al. (2019).
1298 Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of
1299 human TCRVδ1 and TCRVδ2 γδ T lymphocytes. *Proc Natl Acad Sci U S A* 116(24)**,**
1300 11906-11915. doi: 10.1073/pnas.1818488116.
1301 Qian Liu, M.M. (2016). Evaluation of methods in removing batch effects on RNA-seq data.
1302 *Infect Dis Transl Med* 2(1)**,** 3-9. doi: 10.11979/idtm.201601002.
1303 Rincon-Orozco, B., Kunzmann, V., Wrobel, P., Kabelitz, D., Steinle, A., and Herrmann, T.
1304 (2005). Activation of V gamma 9V delta 2 T cells by NKG2D. *J Immunol* 175(4)**,** 2144-
1305 2151. doi: 10.4049/jimmunol.175.4.2144.
1306 Rodríguez-Gómez, I.M., Talker, S.C., Käser, T., Stadler, M., Reiter, L., Ladinig, A., et al.
1307 (2019). Expression of T-Bet, Eomesodermin, and GATA-3 Correlates With Distinct
1308 Phenotypes and Functional Properties in Porcine γδ T Cells. *Front Immunol* 10**,** 396. doi:
1309 10.3389/fimmu.2019.00396.

1310 Rusmini, M., Griseri, P., Lantieri, F., Matera, I., Hudspeth, K.L., Roberto, A., et al. (2013).
1311 Induction of RET dependent and independent pro-inflammatory programs in human
1312 peripheral blood mononuclear cells from Hirschsprung patients. *PLoS One* 8(3), e59066.
1313 doi: 10.1371/journal.pone.0059066.
1314 Rusmini, M., Griseri, P., Matera, I., Pontarini, E., Ravazzolo, R., Mavilio, D., et al. (2014).
1315 Expression variability and function of the RET gene in adult peripheral blood
1316 mononuclear cells. *J Cell Physiol* 229(12), 2027-2037. doi: 10.1002/jcp.24660.
1317 Saalmüller, A., Werner, T., and Fachinger, V. (2002). T-helper cells from naive to committed.
1318 *Vet Immunol Immunopathol* 87(3-4), 137-145. doi: 10.1016/s0165-2427(02)00045-4.
1319 Schroyen, M., and Tuggle, C.K. (2015). Current transcriptomics in pig immunity research.
1320 *Mamm Genome* 26(1-2), 1-20. doi: 10.1007/s00335-014-9549-4.
1321 Sedlak, C., Patzl, M., Saalmüller, A., and Gerner, W. (2014). CD2 and CD8α define porcine γδ T
1322 cells with distinct cytokine production profiles. *Dev Comp Immunol* 45(1), 97-106. doi:
1323 10.1016/j.dci.2014.02.008.
1324 Shi, W., Liao, Y., Willis, S.N., Taubenheim, N., Inouye, M., Tarlinton, D.M., et al. (2015).
1325 Transcriptional profiling of mouse B cell terminal differentiation defines a signature for
1326 antibody-secreting plasma cells. *Nat Immunol* 16(6), 663-673. doi: 10.1038/ni.3154.
1327 Siegers, G.M. (2018). Integral Roles for Integrins in γδ T Cell Function. *Front Immunol* 9, 521.
1328 doi: 10.3389/fimmu.2018.00521.
1329 Song, L., and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina
1330 RNA-seq reads. *Gigascience* 4, 48. doi: 10.1186/s13742-015-0089-y.
1331 Steiner, J., Marquardt, N., Pauls, I., Schiltz, K., Rahmoune, H., Bahn, S., et al. (2011). Human
1332 CD8(+) T cells and NK cells express and secrete S100B upon stimulation. *Brain Behav*
1333 *Immun* 25(6), 1233-1241. doi: 10.1016/j.bbi.2011.03.015.
1334 Stepanova, K., and Sinkora, M. (2013). Porcine γδ T lymphocytes can be categorized into two
1335 functionally and developmentally distinct subsets according to expression of CD2 and
1336 level of TCR. *J Immunol* 190(5), 2111-2120. doi: 10.4049/jimmunol.1202890.
1337 Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., et al. (2019).
1338 Comprehensive Integration of Single-Cell Data. *Cell* 177(7), 1888-1902.e1821. doi:
1339 10.1016/j.cell.2019.05.031.
1340 Summerfield, A., Meurens, F., and Ricklin, M.E. (2015). The immunology of the porcine skin
1341 and its value as a model for human skin. *Mol Immunol* 66(1), 14-21. doi:
1342 10.1016/j.molimm.2014.10.023.
1343 Summerfield, A., Rziha, H.J., and Saalmüller, A. (1996). Functional characterization of porcine
1344 CD4+CD8+ extrathymic T lymphocytes. *Cell Immunol* 168(2), 291-296. doi:
1345 10.1006/cimm.1996.0078.
1346 Summers, K.M., Bush, S.J., and Hume, D.A. (2020). Network analysis of transcriptomic
1347 diversity amongst resident tissue macrophages and dendritic cells in the mouse
1348 mononuclear phagocyte system. *PLoS Biol* 18(10), e3000859. doi:
1349 10.1371/journal.pbio.3000859.
1350 Sutermaster, B.A., and Darling, E.M. (2019). Considerations for high-yield, high-throughput cell
1351 enrichment: fluorescence versus magnetic sorting. *Sci Rep* 9(1), 227. doi:
1352 10.1038/s41598-018-36698-1.
1353 Swindle, M.M., Makin, A., Herron, A.J., Clubb, F.J., Jr., and Frazier, K.S. (2012). Swine as
1354 models in biomedical research and toxicology testing. *Vet Pathol* 49(2), 344-356. doi:
1355 10.1177/0300985811402846.

1356 Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., et al. (2019). Single-
1357 cell transcriptomics of human T cells reveals tissue and activation signatures in health
1358 and disease. *Nat Commun* 10(1)**,** 4706. doi: 10.1038/s41467-019-12464-3.
1359 Takamatsu, H.H., Denyer, M.S., Stirling, C., Cox, S., Aggarwal, N., Dash, P., et al. (2006).
1360 Porcine gammadelta T cells: possible roles on the innate and adaptive immune responses
1361 following virus infection. *Vet Immunol Immunopathol* 112(1-2)**,** 49-61. doi:
1362 10.1016/j.vetimm.2006.03.011.
1363 Toka, F.N., Nfon, C.K., Dawson, H., and Golde, W.T. (2009). Accessory-cell-mediated
1364 activation of porcine NK cells by toll-like receptor 7 (TLR7) and TLR8 agonists. *Clin*
1365 *Vaccine Immunol* 16(6)**,** 866-878. doi: 10.1128/CVI.00035-09.
1366 Tsang, J.S., Schwartzberg, P.L., Kotliarov, Y., Biancotto, A., Xie, Z., Germain, R.N., et al.
1367 (2014). Global analyses of human immune variation reveal baseline predictors of
1368 postvaccination responses. *Cell* 157(2)**,** 499-513. doi: S0092-8674(14)00406-1 [pii]
1369 10.1016/j.cell.2014.03.031.
1370 Tu, Y.H., Cooper, A.J., Teng, B., Chang, R.B., Artiga, D.J., Turner, H.N., et al. (2018). An
1371 evolutionarily conserved gene family encodes proton-selective ion channels. *Science*
1372 359(6379)**,** 1047-1050. doi: 10.1126/science.aao3264.
1373 Upadhyay, G. (2019). Emerging Role of Lymphocyte Antigen-6 Family of Genes in Cancer and
1374 Immune Cells. *Front Immunol* 10**,** 819. doi: 10.3389/fimmu.2019.00819.
1375 van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J.P. (2018). Gene co-
1376 expression analysis for functional classification and gene-disease predictions. *Brief*
1377 *Bioinform* 19(4)**,** 575-592. doi: 10.1093/bib/bbw139.
1378 Van Goor, A., Pasternak, A., Walker, K., Hong, L., Malgarin, C., MacPhee, D.J., et al. (2020).
1379 Differential responses in placenta and fetal thymus at 12 days post infection elucidate
1380 mechanisms of viral level and fetal compromise following PRRSV2 infection.
1381 Vieira Braga, F.A., Teichmann, S.A., and Chen, X. (2016). Genetics and immunity in the era of
1382 single-cell genomics. *Hum Mol Genet* 25(R2)**,** R141-R148. doi: 10.1093/hmg/ddw192.
1383 Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., et al. (2017).
1384 Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and
1385 progenitors. *Science* 356(6335). doi: 10.1126/science.aah4573.
1386 Wang, H., Horbinski, C., Wu, H., Liu, Y., Sheng, S., Liu, J., et al. (2016). NanoStringDiff: a
1387 novel statistical method for differential expression analysis based on NanoString
1388 nCounter data. *Nucleic Acids Res* 44(20)**,** e151. doi: 10.1093/nar/gkw677.
1389 Wang, M., Windgassen, D., and Papoutsakis, E.T. (2008). Comparative analysis of
1390 transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune
1391 response players in T-cell activation. *BMC Genomics* 9**,** 225. doi: 10.1186/1471-2164-9-
1392 225.
1393 Ware, C.F. (2005). Network communications: lymphotoxins, LIGHT, and TNF. *Annu Rev*
1394 *Immunol* 23**,** 787-819. doi: 10.1146/annurev.immunol.23.021704.115719.
1395 Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D.M., Billis, K., et al. (2020). An improved
1396 pig reference genome sequence to enable pig genetics and genomics research.
1397 *Gigascience* 9(6). doi: 10.1093/gigascience/giaa051.
1398 Wilson, N.K., and Göttgens, B. (2018). Single-Cell Sequencing in Normal and Malignant
1399 Hematopoiesis. *Hemasphere* 2(2)**,** e34. doi: 10.1097/HS9.0000000000000034.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* 8(4)**,** 281-291.e289. doi: 10.1016/j.cels.2018.11.005.

Yang, H., and Parkhouse, R.M. (1996). Phenotypic classification of porcine lymphocyte subpopulations in blood and lymphoid tissues. *Immunology* 89(1)**,** 76-83. doi: 10.1046/j.1365-2567.1996.d01-705.x.

Yang, H., and Parkhouse, R.M. (1997). Differential expression of CD8 epitopes amongst porcine CD8-positive functional lymphocyte subsets. *Immunology* 92(1)**,** 45-52. doi: 10.1046/j.1365-2567.1997.00308.x.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Alvarez-Jarreta, J., Amode, M.R., et al. (2020). Ensembl 2020. *Nucleic Acids Res* 48(D1)**,** D682-D688. doi: 10.1093/nar/gkz966.

Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9(12). doi: 10.1093/gigascience/giaa151.

Zemmour, D., Zilionis, R., Kiner, E., Klein, A.M., Mathis, D., and Benoist, C. (2018). Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat Immunol* 19(3)**,** 291-301. doi: 10.1038/s41590-018-0051-0.

Zhang, B., Lin, Y.Y., Dai, M., and Zhuang, Y. (2014). Id3 and Id2 act as a dual safety mechanism in regulating the development and population size of innate-like γδ T cells. *J Immunol* 192(3)**,** 1055-1063. doi: 10.4049/jimmunol.1302694.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8**,** 14049. doi: 10.1038/ncomms14049.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 10(1)**,** 1523. doi: 10.1038/s41467-019-09234-6.

Zhu, L., Yang, P., Zhao, Y., Zhuang, Z., Wang, Z., Song, R., et al. (2020). Single-Cell Sequencing of Peripheral Mononuclear Cells Reveals Distinct Immune Response Landscapes of COVID-19 and Influenza Patients. *Immunity* 53(3)**,** 685-696.e683. doi: 10.1016/j.immuni.2020.07.009.

Zhu, Y., Yao, S., Iliopoulou, B.P., Han, X., Augustine, M.M., Xu, H., et al. (2013). B7-H5 costimulates human T cells via CD28H. *Nat Commun* 4**,** 2043. doi: 10.1038/ncomms3043.

Zuckermann, F.A. (1999). Extrathymic CD4/CD8 double positive T cells. *Veterinary immunology and immunopathology* 72(1-2)**,** 55-66.

Tables

Table 1. Abbreviations and phenotype information of pig sorted immune cells. [a] Refers to gate in Figure 1. [b] Reagents listed in materials and methods.

			CD3ε$^+$ MACS Fraction			
Gate[a]	**Population Abbreviation**	**Marker**	**Clone**	**Fluorophore**	**Company (Catalog #)b**	**FACS Sort Criteria**
-	-	Anti mouse IgG1	RMG1-1	PE-Cy7	BioLegend(406614)	-
1	SWC6gdT	SWC6gdT	MAC320	APC	BD(561482)	CD3ε$^+$SWC6$^+$
2	CD4T	CD4	74-12-4	FICT	BD(559585)	CD3ε$^+$SWC6$^-$CD4$^+$CD8α$^-$
3	CD4CD8T	CD4CD8α	74-12-4 / 76-2-11	FICT / PE	BD(559585) / BD(559584)	CD3ε$^+$SWC6$^-$CD4$^+$CD8α$^+$
4	CD8T	CD8α	76-2-11	PE	BD(559584)	CD3ε$^+$SWC6$^-$CD4$^-$CD8α$^+$
			CD3ε$^-$ MACS Fraction			
Gate[a]	**Population Abbreviation**	**Marker**	**Clone**	**Fluorophore**	**Company (Catalog #)b**	**FACS Sort Criteria[c]**
5	Myeloid	CD172	74-22-15A	FICT	BD(561498)	CD3ε$^-$CD172α$^+$CD8α$^-$
6	NK	CD8α	76-2-11	PE	BD(559584)	CD3ε$^-$CD172α$^-$CD8α$^+$
7	CD21pB	CD21	BB6-11C9.6	AF647	Southern Biotech(4530-31)	CD3ε$^-$CD172α$^-$CD8α$^-$CD21$^+$
8	CD21nB	-	-	-	-	CD3ε$^-$CD172α$^-$CD8α$^-$CD21$^-$

1441 **Table 2.** Cell type-enriched and cell type specific genes identified in pig sorted immune cells.

Cell type	Enriched genes	Genes without symbol	Specific genes	Genes without symbol
SWC6gdT	3591	481	8	1
CD8T	3318	566	2	-
CD4CD8T	2271	312	0	-
CD4T	2606	374	0	-
NK	1855	304	29	5
Myeloid	3440	432	397	60
CD21pB	2383	432	5	2
CD21nB	2456	414	0	-

1442

1443

Table 3. Specific highly enriched genes in myeloid, NK, CD21pB, SWC6gdT and CD4CD8T-cells.

Group	Total genes	Gene names
Specific Myeloid + Top 25% myeloid	271	*SLC18A1, ENSSSCG00000025687, KLHL13, PAK1, C1RL, MITF, SIRPB2, ENSSSCG00000014997, HNMT, C5AR1, A2M, TEK, SEL1L3, TSPAN13, ENSSSCG00000035960, ENSSSCG00000039214, ENSSSCG00000003226, APOE, CHST15, DNM1, GAS2L1, SERPING1, COL18A1, CDS1, ENSSSCG00000016184, CRHBP, KCNE3, NCAM1, ABHD12, ENSSSCG00000001850, ENSSSCG00000023479, ASAH1, FN1, ENSSSCG00000003554, ENSSSCG00000038429, GAA, ECE1, SLC46A2, UBTD1,CEBPD, CTSB, ENSSSCG00000031640, ENSSSCG00000037466, PLAC9, CCDC60, DOPEY2, TALDO1, ADAMTSL4, ENSSSCG00000034555, STK3, ENSSSCG00000021675, FAM129B, SIGLEC1, SULF2, TRPM2, MGP, CMKLR1, TNFRSF19, DOCK4, ENSSSCG00000027991, ULBP1, SLC11A1, SFXN3, TNFSF13, ENSSSCG00000013380, CD68, KCNQ1, RPS6KA2, CD14, MCF2L, ENSSSCG00000037541, ENSSSCG00000015839, PAM, SERPINB8, TSPAN12, F13A1, SASH1, C9orf72, PLCB4, SH3PXD2B, BLVRA, CXCL2, ADAM28, GPBAR1, CHI3L2, SNX9, LGALS3, SLC2A6, ENSSSCG00000035675, EHD4, ENSSSCG00000039758, UNC13A, ENSSSCG00000038418, C2, PLA2G7, FUCA1, ENSSSCG00000037426, ENSSSCG00000025271, ABCA9, RASGRP4, SLC7A7, VCAN, SLC39A8, ADAP2, SMIM5, DAGLA, RAB11FIP5, ZNF768, ENSSSCG00000007644, CTNND1, ENSSSCG00000022258, ENSSSCG00000017754, STXBP1, ENSSSCG00000027665, MANSC1, RND2, IGSF6, BMX, NLRP12, TPST1, NOD2, TREM1, SEMA6B, JDP2, FAM111B, CIDEB, ENSSSCG00000033457, MMP19, SGK3, CTTNBP2NL, MAPK4, PLAUR, INSIG1, RNASE4, FLVCR2, SCARF1, BCL2L14, ENSSSCG00000026196, MCTP1, WLS, ENSSSCG00000017920, PLOD1, CHPT1, PRCP, ENSSSCG00000013842, SH2D6, CA13, PLCB2, CAPN3, PRAM1, ENSSSCG00000038616, ALOX5, GPNMB, ACVRL1, SMIM3, GPR137B, LAMP1, NR1H3, ARL11, ITGB4, CYSLTR2, CCSER1, NCF2, GPCPD1, PDXK, NACC2, FOLR1, ADGRL2, MERTK, OLFM1, PLXNC1, ECM1, LRRC25, IFIT2, CORO1B, ASAP3, SLC43A3, STEAP4, CAMKK1, CTSS, TMEM47, TTLL7, AKR7A2, ENSSSCG00000036342, V1M, TLR8, LIN7A, MPP1, TBXAS1, LIPA, DRAM1, MRC2, TGM3, HEXB, GALM, EREG, JPH4, ANG, QPCT, PPT1, ARRDC4, RAB31, ABHD17C, NFAM1, TLR3, LTB4R, HSD3B7, VDR, ENSSSCG00000010497, CD163, OSCAR, DSC3, LRP6, ENSSSCG00000031951, ENSSSCG00000028635, PSAP, SCPEP1, EPB41L2, ZDHHC9, IL1R2, EXPH5, ENSSSCG00000023264, IFIT5, AGPAT2, NKD2, GUCY1B1, GLUL, COL14A1, TNFRSF1B, SLC16A3, GRN, ENSSSCG00000013100, CEBPA, OLFML2B, TLR4, XG, CCL21, ATF6, SLC49A3, HFE, ACVR1B, IFNGR1, ENSSSCG00000022925, SERPINB10, TCF7L2, ENSSSCG00000008769, ENSSSCG00000016093, UNC93B1, TIMP2, RAMP2, F11R, LGALS8, ENSSSCG00000032723, CFP, ZNF385A, CLIC2, TDRD1, HIP1, ENSSSCG00000026653, GSDMD, CSF1R, NAGK, GAB1, PGD, ENSSSCG00000034639, LRPAP1, DAPK1, ENSSSCG00000039956, GPAT3, GALNTL5, ENSSSCG00000029414*
Specific NK + Top 25% NK	14	*OTOP2, B3GNT7, OSBPL3, NR4A3, IGF2BP2, OTOP3, ENSSSCG00000010703, LY6D, RET, TUBB6, ENSSSCG00000033385, ENSSSCG00000036743, PTH1R, SUSD1*
Specific CD21pB + Top 25% CD21pB	2	*GP2, CR2*
Specific SWC6gdT + Top 25% SWC6gdT	5	*TMEM87B, ACVR2A, ENSSSCG00000028443, SLC4A4, CASS4*
Specific CD8T + Top 25% CD8T	2	*TMIGD2, JAML*

1447 **Table 4.** Genes differentially expressed between both CD2- gd T-cell clusters (clusters 6, 21)

1448 and both CD2+ gd T-cell clusters (clusters 24, 31). * Refer to gene name replacement in methods

Population with greater gene expression	Genes
CD2- γδ T cells (clusters 6, 21)	*AP3S1*, ANXA1, BLK, CAPG, CDNF*, CD163L1*, EMP3, ENSSSCG00000032017, ENSSSCG00000033734, FCER1A, GATA3, IL6R, ITM2B, LGALS1, LTB, MAN2B1*, MYL12A*, PARK7, PIK3AP1, PLEKHF2, PPP1CC, RCAN3, RHEX, RPS19*, SAMSN1, SELL, SLC25A24, SRGN, TIMP1, VIM, YBX3*
CD2+ γδ T cells (clusters 24, 31)	*ABI3*, ACTG1, ARPC1B, ARPC5L, BIN1, CAMK4, CCDC12*, CD2, COTL1, CTSD, DYNLRB1, ENSSSCG00000023584, ENSSSCG00000027196, ENSSSCG00000029596, ENSSSCG00000038825, FAM49B, FSCN1, FYB1, GBP7*, GIMAP4*, H2AFV, IFITM1*, IFI6, IKZF2, IKZF3, IL2RB, ISG15, ITGA4, ITGB2, ITM2C, KRAS, LCK, MAGOHB*, NT5C3A*, PIK3R1, PRKCH*, PSIP1, PTPRC, RESF1, S100A1, SLC9A3R1, SMC4, SNRK, STK17B, STMN3, TRAT1, UBAC2, WCR1, WIPF1**

1449

1450

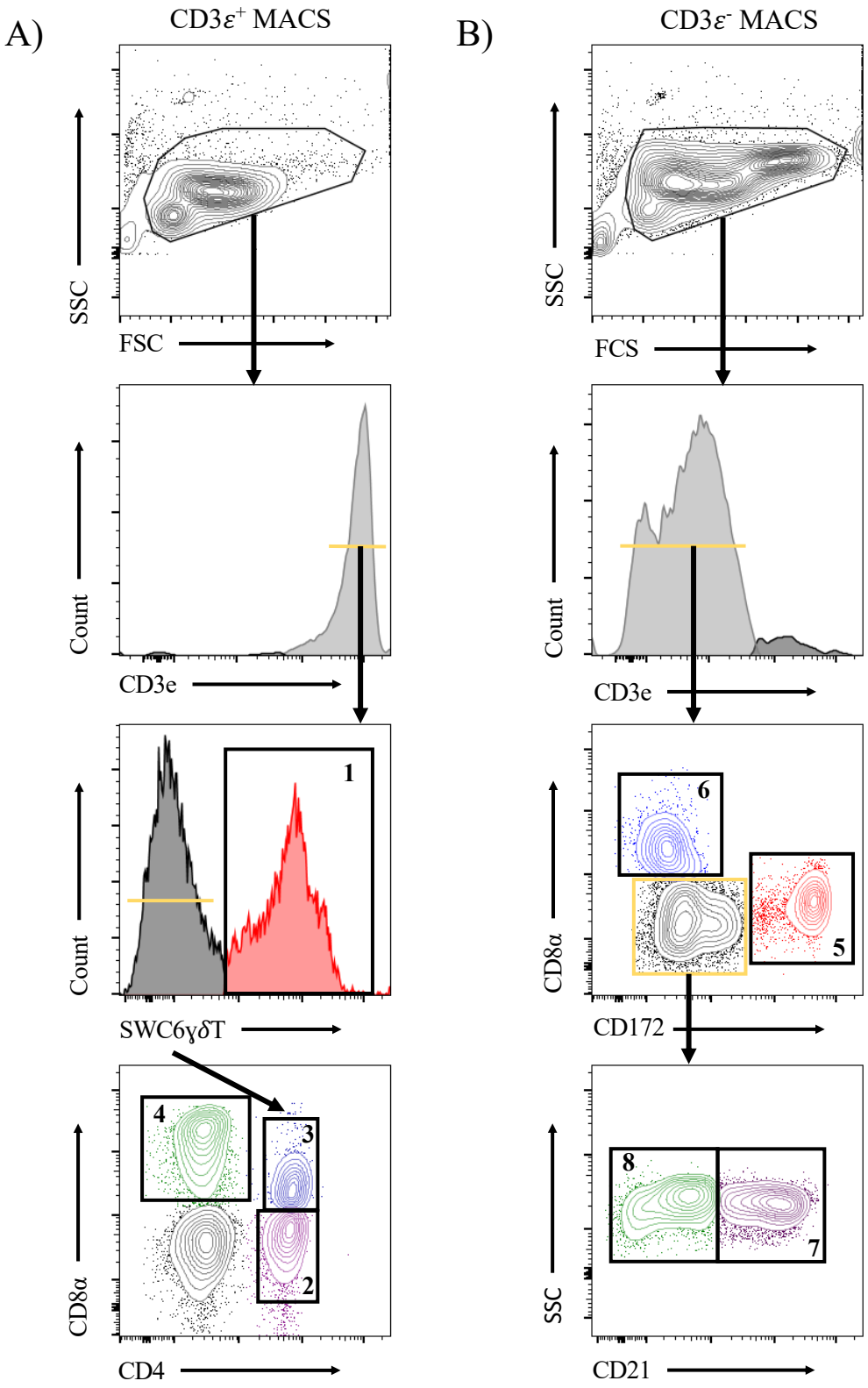
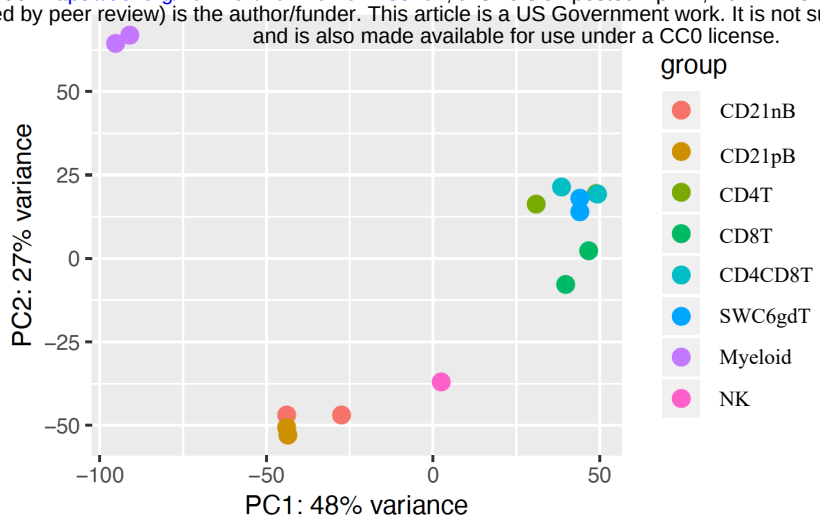
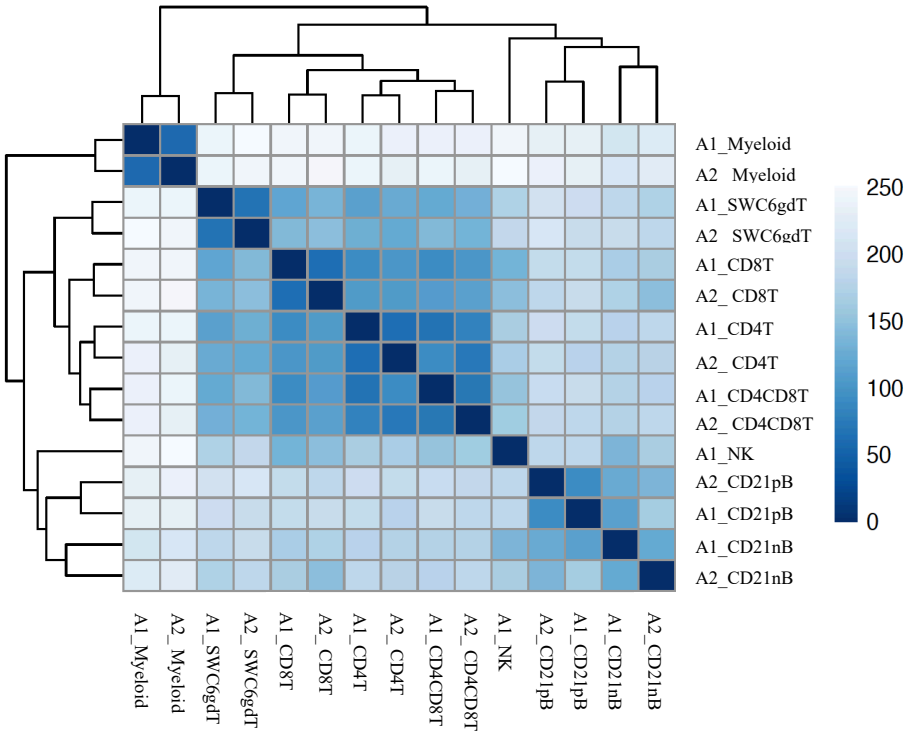

A)
CD3ε+ MACS
SSC
FSC
Count
CD3e
1
Count
SWC6γδT
4
3
2
CD8α
CD4
B)
CD3ε- MACS
SSC
FCS
Count
CD3e
6
5
CD8α
CD172
8
7
SSC
CD21

Figure 1. Representative plots for fluorescence-activated cell sorting (FACS) isolation of 8 leukocyte populations from pig peripheral blood mononuclear cells (PBMCs). Porcine PBMCs were first subjected to magnetic-activated cell sorting (MACS) to enrich for CD3ε+ and CD3ε- fractions. **A)** Cells in $CD3\varepsilon^{+}$ MACS fraction were FACS gated on FSC vs SSC, doublets removed (not shown), and $CD3\varepsilon^{+}$ cells were isolated into 4 population: $SWC6^{+}$ γδ T-cells (gate 1), and the $SWC6^{-}$ cells sorted as $CD4^{+}CD8\alpha^{-}$ (gate 2), $CD4^{+}CD8\alpha^{+}$ (gate 3), $CD4^{-}CD8\alpha^{+}$ (gate 4) T-cells. **B)** Cells in $CD3\varepsilon^{-}$ MACS fraction were FACS gated on FSC vs SSC, doublets removed (not shown), and $CD3\varepsilon^{-}$ cells were isolated into 4 populations: $CD172\alpha^{+}$ myeloid lineage leukocytes (gate 5), $CD8\alpha^{+}CD172^{-}$ NK cells (gate 6), and the remaining $CD8\alpha^{-}$ $CD172\alpha^{-}$, cells were isolated as $CD21^{+}$ (gate 7) and $CD21^{-}$ (gate 8) B-cells. Table 1 outlines abbreviations and sort criteria for each population.

A)



B)



C)

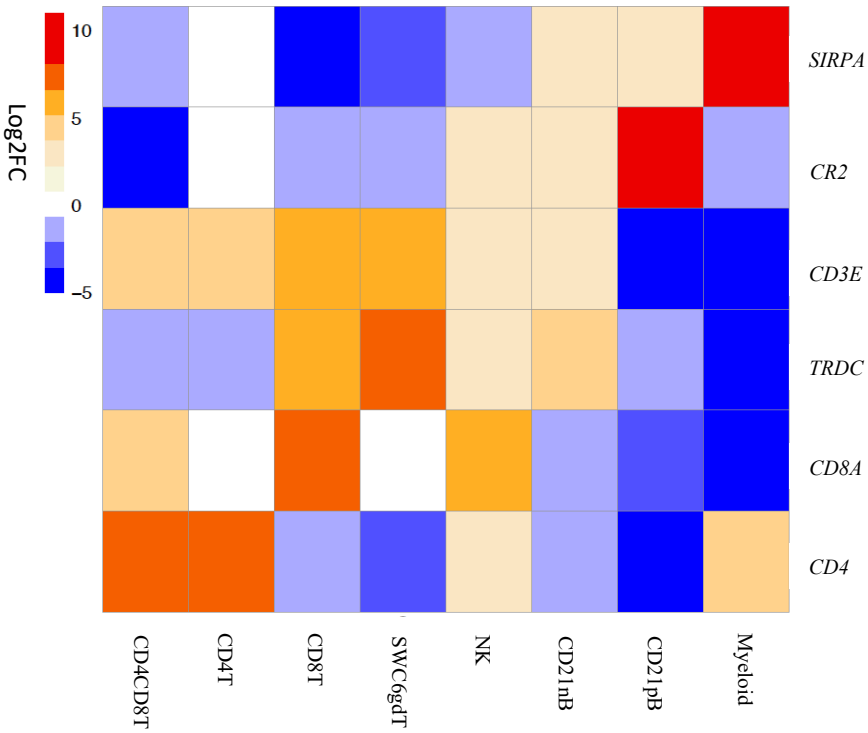


Figure 2. Transcriptional expression patterns of immune cells are distinct and cluster more by progenitors. **A)** Principal component analysis of transformed RNA-seq reads counts for whole transcriptomes. Axis indicate component scores. **B)** Heat map depicting hierarchical clustering of sample-to-sample distance. Gene expression for whole transcriptomes were used to calculate sample to sample Euclidean distance (color scale) for hierarchical clustering. **C)** Heatmap showing cell-type enriched gene values (Log2FC) between sorted immune cells. Gene coding proteins that were used for cell sorting were display.

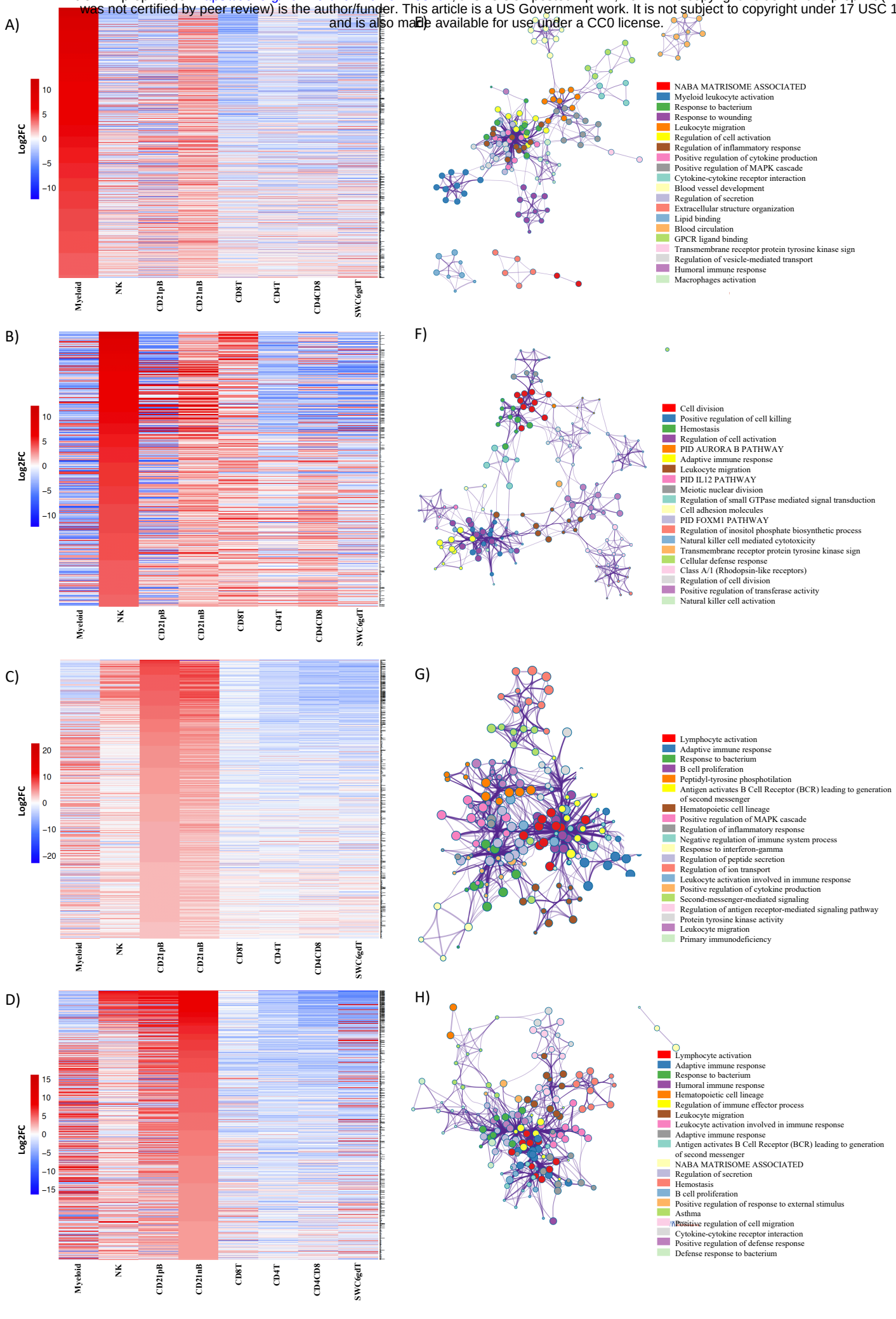
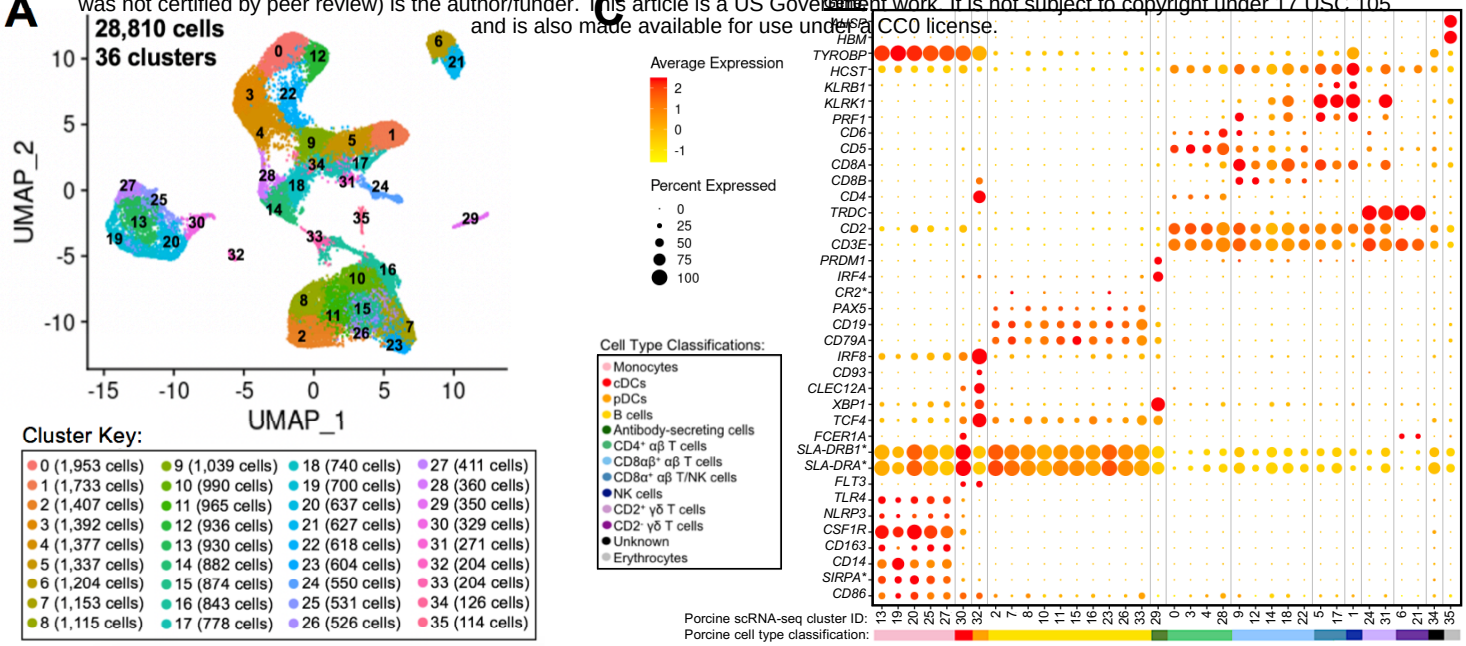
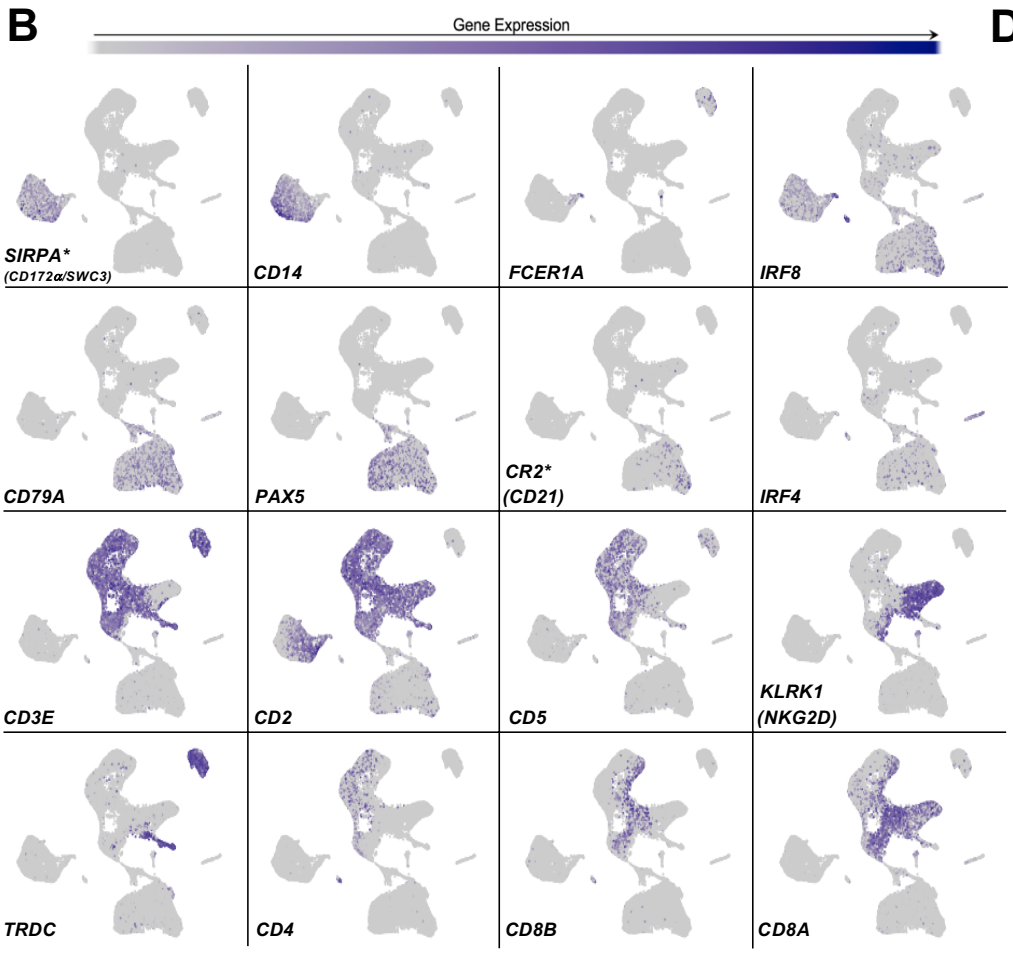


Figure 3. Top 25% highly enriched genes in CD3- sorted cells. Heatmap showing in decreasing order the top 25% of highly enriched genes in **A**) myeloid, **B**) NK, **C**) CD21pB and **D)** CD21nB-cells. Ontology enrichment clusters of the top 25% highly enriched genes of **E**) myeloid, **F**) NK, **G**) CD21pB and **H)** CD21nB-cells. The most statistically significant term within similar term cluster was chosen to represent the cluster. Term color is given by cluster ID and the size of the terms is given by –log10 P-value. The stronger the similarity among terms, the thicker the edges between them.





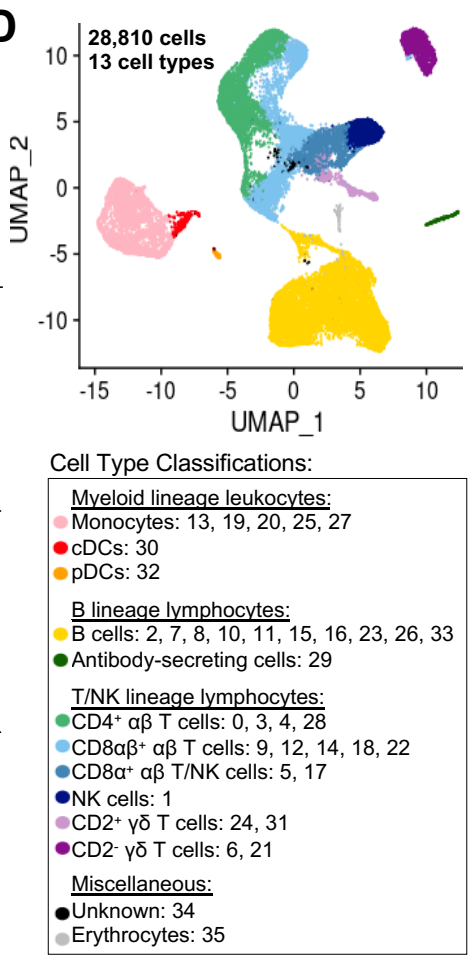
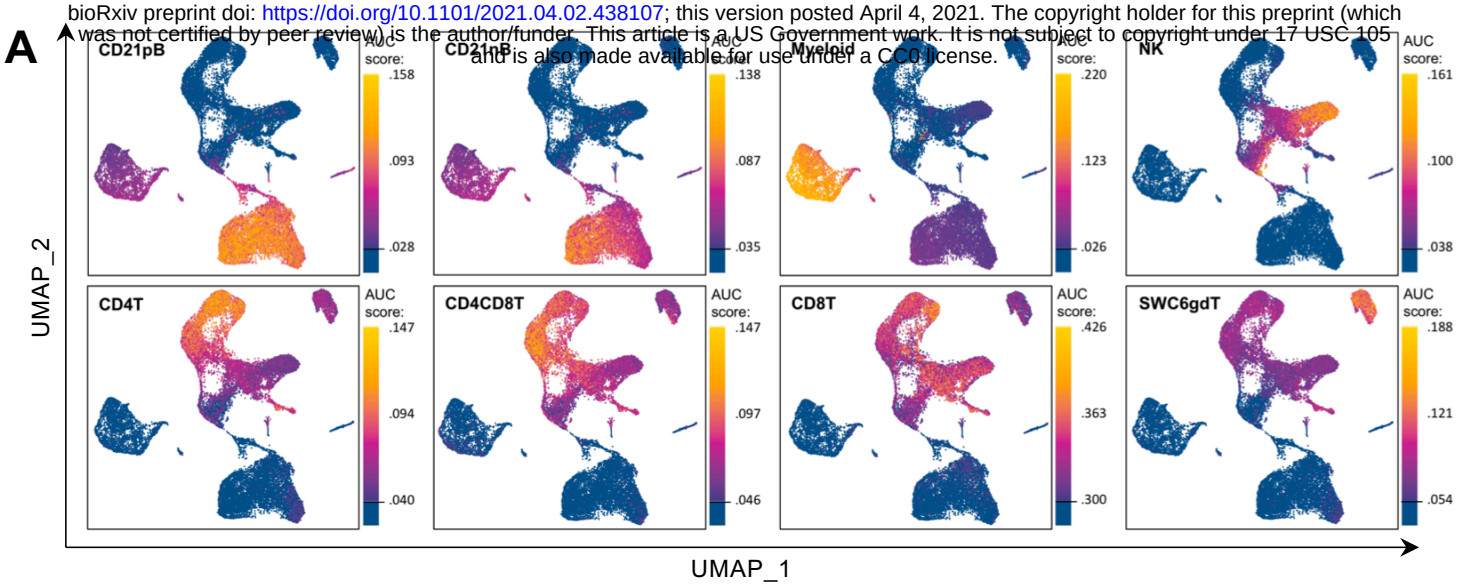
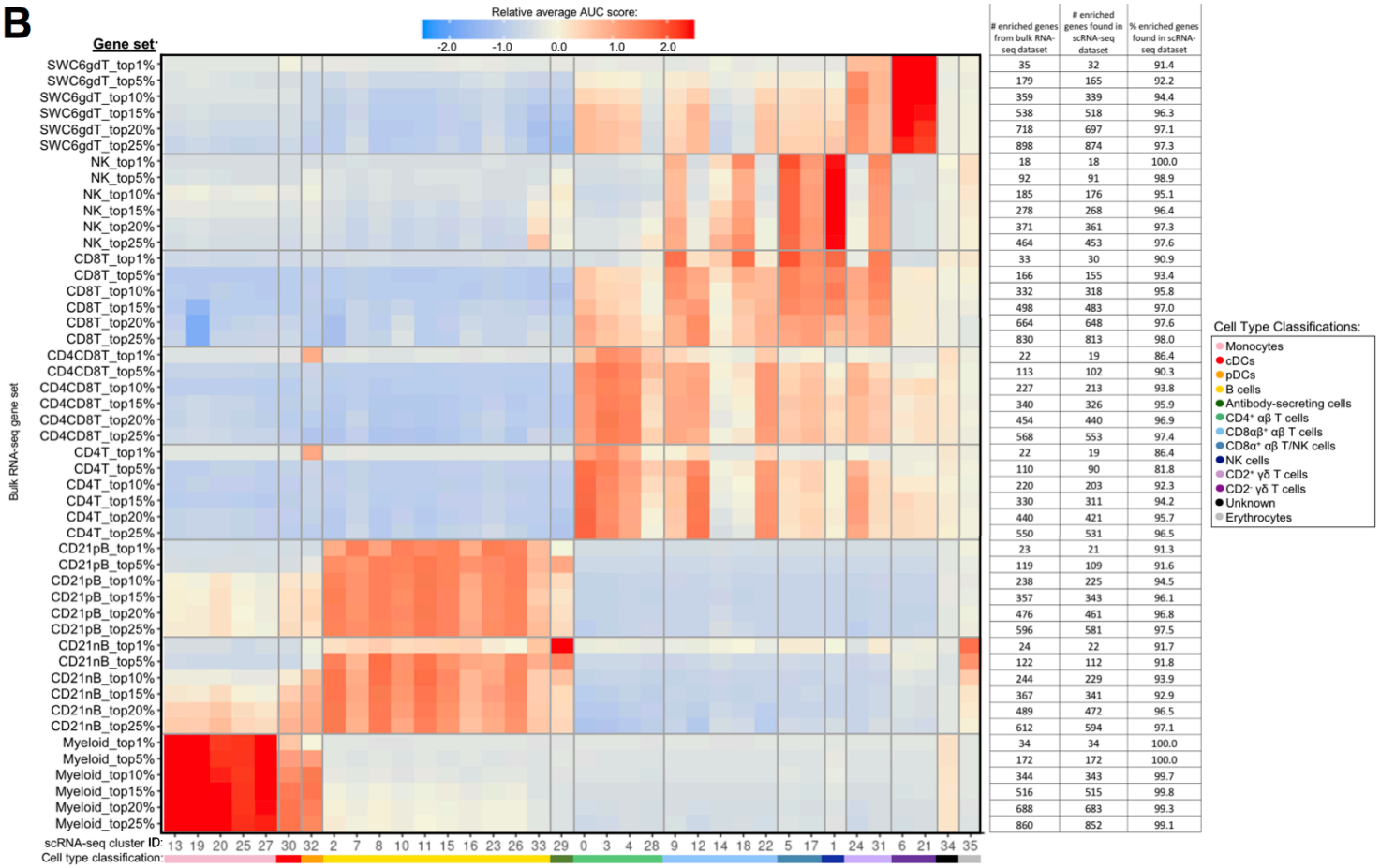
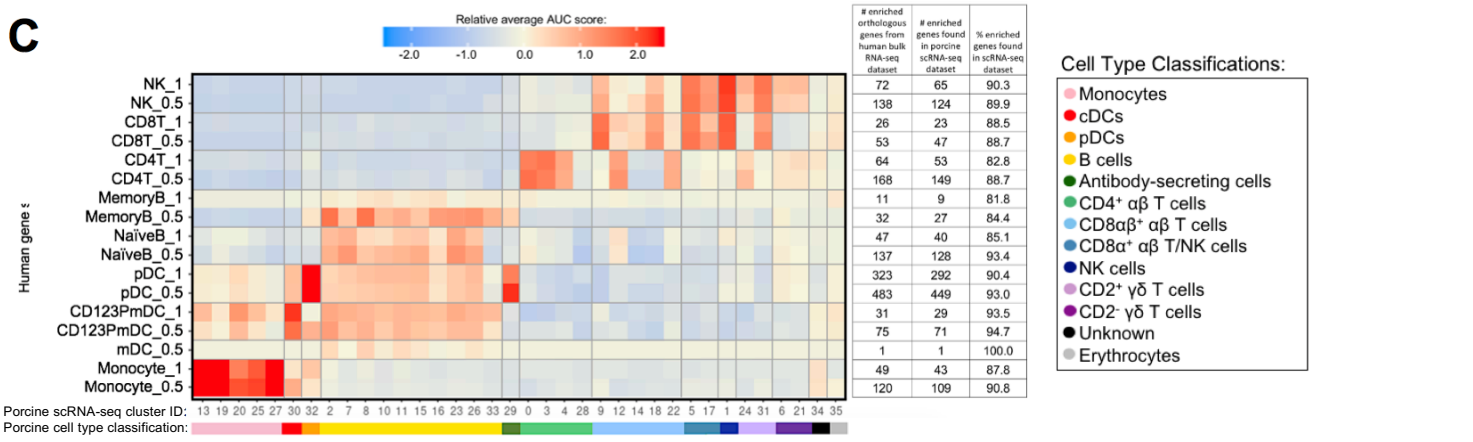


Figure 4. Classification of porcine PBMC scRNA-seq clusters based on known cell type-specific gene expression. **A)** Two-dimensional UMAP visualization of 28,810 single cells from porcine PBMCs classified into 36 designated clusters. Each point represents a single cell. Color of the point corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **B)** Visualization of selected cell type-specific gene expression overlaid onto two-dimensional UMAP coordinates of single cells. Each point represents a single cell. Color of the point corresponds to relative expression of a specified gene (bottom left of each UMAP plot) within a cell. Grey corresponds to little/no gene expression, while navy corresponds to increased gene expression. **C)** Dot Plot visualization of selected cell type-specific gene expression for each single-cell cluster shown in A. Clusters are listed on the x-axis, while selected genes are listed on the y-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene. The color of a dot corresponds to the average relative expression level for the gene in the cells expressing the gene within a cluster. Color bar below the x-axis corresponds to porcine cell type each cluster was classified as. **D)** Two-dimensional UMAP visualization of single cells from porcine PBMCs classified into major porcine cell types. Each point represents a single cell. Color of the cell corresponds to porcine cell type the respective cluster was designated as based on gene expression patterns for the cluster it belonged to in C. Seven PBMC samples used for scRNA-seq analysis were derived from each of three separate experiments (experiment B, n=2; experiment C, n = 3; experiment D, n = 2). Between 3,042 and 6,518 cells were derived from each PBMC sample. *Refer to 'Gene name replacement' methods.





# enriched genes from bulk RNA-seq dataset	# enriched genes found in scRNA-seq dataset	% enriched genes found in scRNA-seq dataset
35	32	91.4
179	165	92.2
359	339	94.4
538	518	96.3
718	697	97.1
898	874	97.3
18	18	100.0
92	91	98.9
185	176	95.1
278	268	96.4
371	361	97.3
464	453	97.6
33	30	90.9
166	155	93.4
332	318	95.8
498	483	97.0
664	648	97.6
830	813	98.0
22	19	86.4
113	102	90.3
227	213	93.8
340	326	95.9
454	440	96.9
568	553	97.4
22	19	86.4
110	90	81.8
220	203	92.3
330	311	94.2
440	421	95.7
550	531	96.5
23	21	91.3
119	109	91.6
238	225	94.5
357	343	96.1
476	461	96.8
596	581	97.5
24	22	91.7
122	112	91.8
244	229	93.9
367	341	92.9
489	472	96.5
612	594	97.1
34	34	100.0
172	172	100.0
344	343	99.7
516	515	99.8
688	683	99.3
860	852	99.1



# enriched orthologous genes from human bulk RNA-seq dataset	# enriched genes found in porcine scRNA-seq dataset	% enriched genes found in scRNA-seq dataset
72	65	90.3
138	124	89.9
26	23	88.5
53	47	88.7
64	53	82.8
168	149	88.7
11	9	81.8
32	27	84.4
47	40	85.1
137	128	93.4
323	292	90.4
483	449	93.0
31	29	93.5
75	71	94.7
1	1	100.0
49	43	87.8
120	109	90.8

Figure 5. Enrichment of gene signatures from bulkRNA-seq in porcine single-cell clusters. **A)** Gene set enrichment scores calculated by AUCell analysis of enriched gene sets from the top 25% of SEGs in pig bulkRNA-seq sorted populations overlaid onto cells of the porcine scRNA-seq dataset visualized in two-dimensional UMAP plot. Each point represents a single cell. The color of the point corresponds to the AUC score calculated for each respective cell. Higher AUC scores correspond to a greater percentage of cells from a gene set being detected in the top 5% of expressed genes in a cell. A threshold for AUC score detection within each gene set was set as shown in Supplementary Figure 10A and is indicated by a horizontal line on the gradient fill scale for each plot. **B)** Relative average gene set enrichment scores of scRNA-seq clusters calculated by AUCell analysis of enriched gene sets from porcine bulkRNA-seq sorted data. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across different gene sets (across columns in the heatmap). Gene sets were created from the top 1, 5, 10, 15, 20, or 25% of SEGs from sorted populations, as determined by highest log2FC values in the porcine bulkRNA-seq data. The number of genes included from the bulkRNA-seq dataset and the number and percent of genes detected in the scRNA-seq dataset is listed on the right of the heatmap. A color bar under scRNA-seq cluster IDs indicates the cell type classification, as according to Figure 4D. **C)** Relative average gene set enrichment scores of scRNA-seq clusters calculated by AUCell analysis of enriched gene sets from human bulkRNA-seq sorted data. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across different gene sets (across columns in the heatmap). Gene sets were created from genes with high expression scores > 0.5 or >1 for each respective sorted population of cells, with a greater high expression score indicating greater enrichment. The number of genes

1524 included from the bulkRNA-seq dataset and the number and percent of genes detected in the

1525 scRNA-seq dataset is listed on the right of the heatmap. A color bar under scRNA-seq cluster

1526 IDs indicates the cell type classification, as according to Figure 4D.

1527

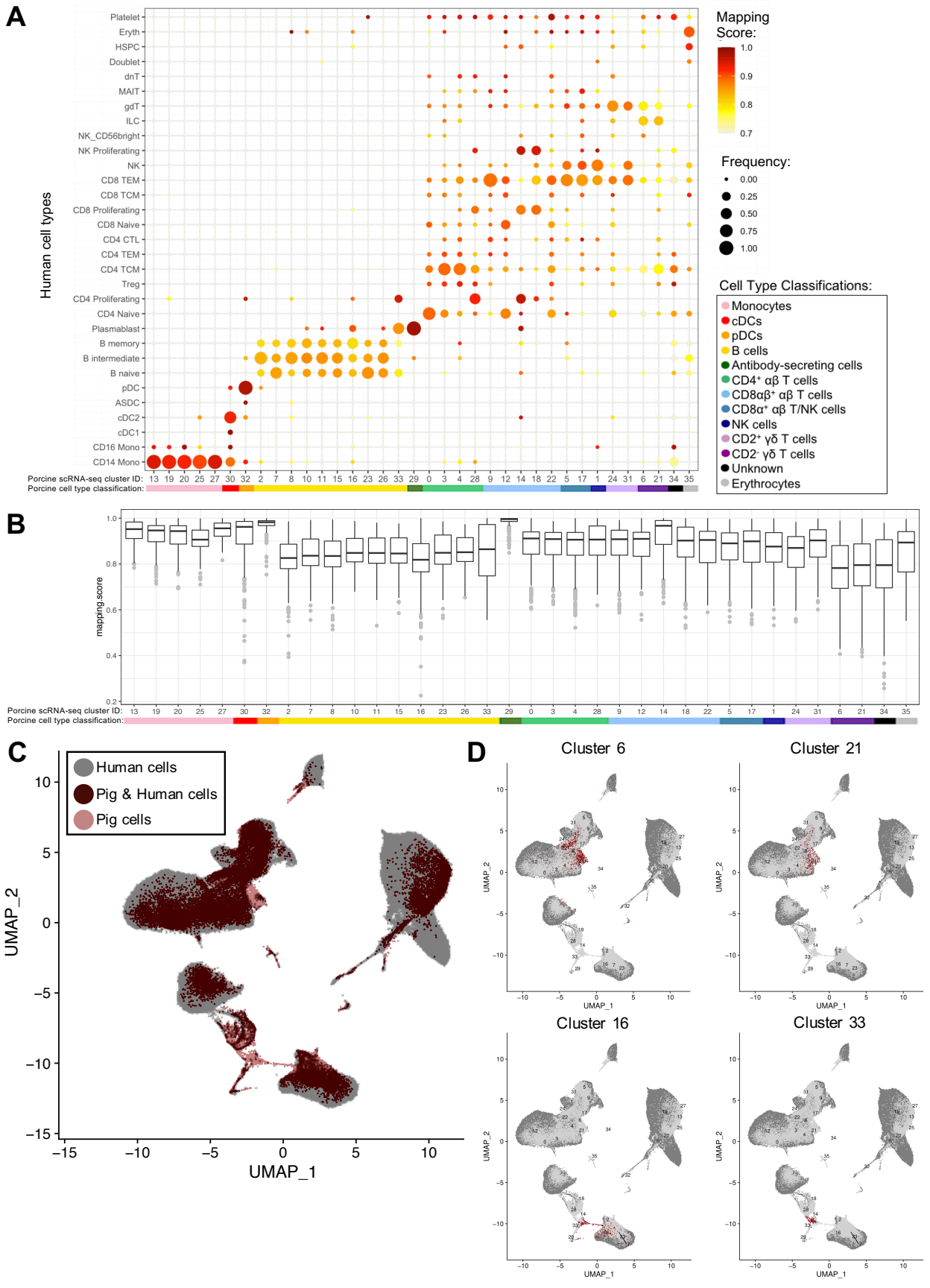

A
Human cell types
Platelet
Eryth
HSPC
Doublet
dnT
MAIT
gdT
ILC
NK_CD56bright
NK Proliferating
NK
CD8 TEM
CD8 TCM
CD8 Proliferating
CD8 Naive
CD4 CTL
CD4 TEM
CD4 TCM
Treg
CD4 Proliferating
CD4 Naive
Plasmablast
B memory
B intermediate
B naive
pDC
ASDC
cDC2
cDC1
CD16 Mono
CD14 Mono
Porcine scRNA-seq cluster ID: 13 19 20 25 27 30 32 2 7 8 10 11 15 16 23 26 33 29 0 3 4 28 9 12 14 18 22 5 17 1 24 31 6 21 34 35
Porcine cell type classification:
Mapping Score:
1.0
0.9
0.8
0.7
Frequency:
0.00
0.25
0.50
0.75
1.00
Cell Type Classifications:
Monocytes
cDCs
pDCs
B cells
Antibody-secreting cells
CD4+ αβ T cells
CD8αβ+ αβ T cells
CD8α+ αβ T/NK cells
NK cells
CD2+ γδ T cells
CD2- γδ T cells
Unknown
Erythrocytes
B
mapping.score
1.0
0.8
0.6
0.4
0.2
Porcine scRNA-seq cluster ID: 13 19 20 25 27 30 32 2 7 8 10 11 15 16 23 26 33 29 0 3 4 28 9 12 14 18 22 5 17 1 24 31 6 21 34 35
Porcine cell type classification:
C
Human cells
Pig & Human cells
Pig cells
UMAP_2
UMAP_1
10
5
0
−5
−10
−15
D
Cluster 6
Cluster 21
Cluster 16
Cluster 33
UMAP_2
UMAP_1

Figure 6. Integration of porcine and human scRNA-seq datasets to further annotate porcine cells**. A)** Mapping scores calculated to determine how well porcine cells were represented by the human dataset. The human cell type specific frequency (size of the circle) and mapping score for that human cell type (color) are shown for each porcine scRNA-seq cluster. Porcine cell type classifications (color) are shown below the porcine scRNA-seq cluster IDs. **B)** Mapping scores calculated to determine how well porcine cells were represented by the human dataset. The mapping scores for each porcine scRNA-seq cluster is represented by a box and whiskers plot. Porcine cell type classifications (color) are shown below the porcine scRNA-seq cluster IDs. **C)** To identify cells in the porcine dataset that were not well represented in the human dataset, a de-novo visualization of the merged porcine and human data was performed. The porcine (pink) and human (grey) were plotted together using UMAP. An overlap of both porcine and human cells is shown as (dark red). Clusters of porcine cells that are not well represented in the human data can be observed by pink regions in the plot. **D)** Two primary regions of porcine cells that were not well represented in the human data were identified in C. In order to clarify which porcine scRNA-seq clusters were represented in these regions, the porcine cluster IDs were projected onto the UMAP and cells from four clusters overlapping the identified regions were colored as dark red.

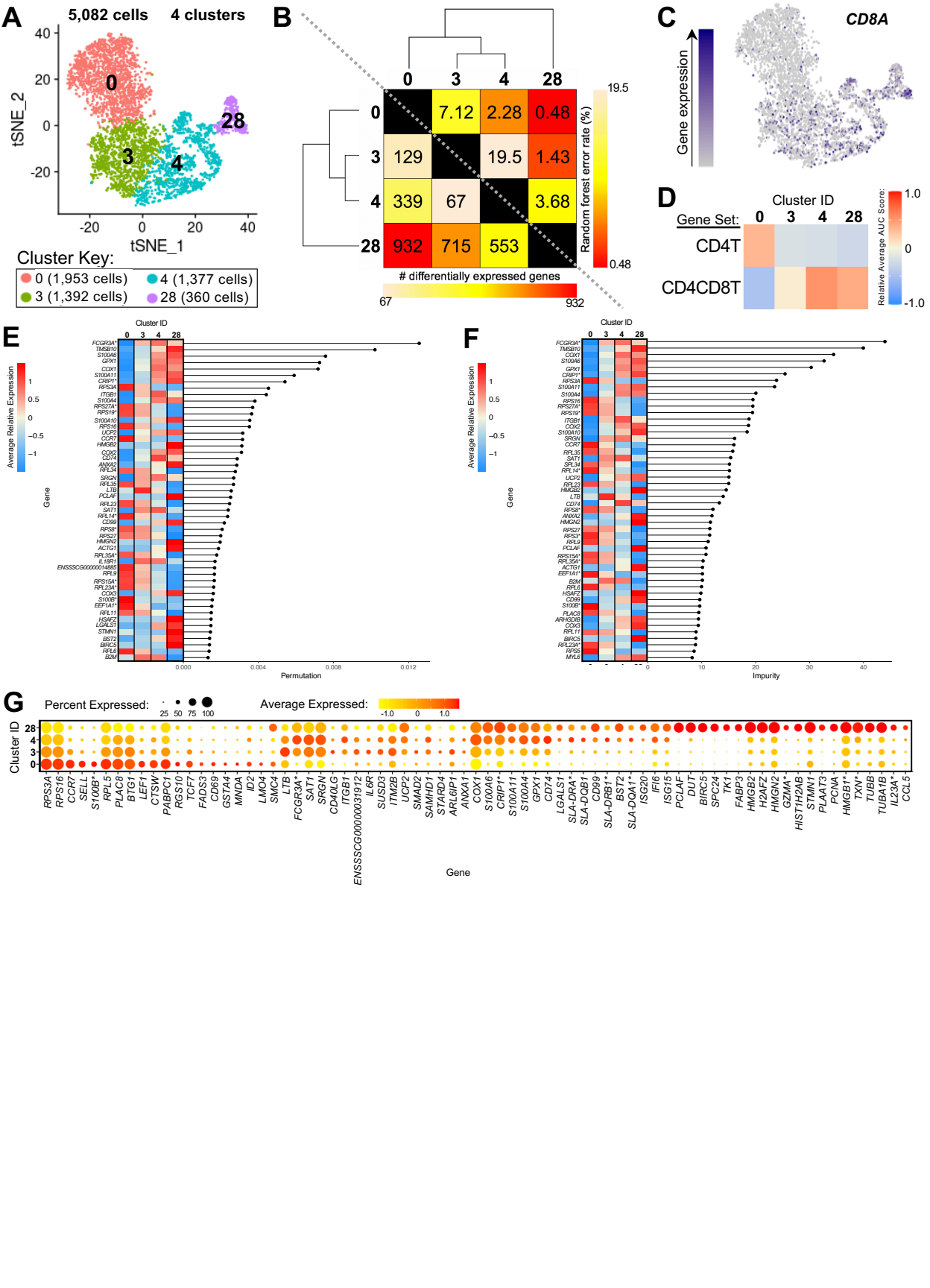

A
5,082 cells
4 clusters
tSNE_2
tSNE_1
Cluster Key:
0 (1,953 cells)
4 (1,377 cells)
3 (1,392 cells)
28 (360 cells)
B
Random forest error rate (%)
19.5
0.48
differentially expressed genes
67
932
C
CD8A
Gene expression
D
Cluster ID
Gene Set:
CD4T
CD4CD8T
Relative Average AUC Score:
1.0
0
-1.0
E
Cluster ID
Average Relative Expression
Gene
Permutation
F
Cluster ID
Average Relative Expression
Gene
Impurity
G
Percent Expressed:
Average Expressed:
Cluster ID
Gene

Figure 7. Transcriptional heterogeneity of porcine CD4+ ab T-cells at single-cell resolution. **A)** Two-dimensional t-SNE plot of 5,082 cells belonging to clusters designated as CD4+ ab T-cells (clusters 0, 3, 4, 28) in Figure 4D. Each point represents a single cell. Color of the cell corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **B)** Transcriptomic relationship amongst CD4+ ab T-cell clusters as calculated by three methods: hierarchical clustering (as seen by hierarchical trees on both axes), pairwise random forest analyses (as seen on top right diagonal); and pairwise DGE analyses (as seen on bottom left diagonal). Longer branches on the hierarchical tree corresponds to greater hierarchical distance. Lower numbers of DEGs by DGE analysis and higher out-of-bag (OOB) error rates from random forest analyses indicate greater pairwise transcriptional similarity. **C)** Visualization of CD8A expression overlaid onto t-SNE coordinates of single CD4+ ab T-cells. Each point represents a single cell. Color of the point corresponds to relative expression of CD8A within a cell. Grey corresponds to little/no gene expression, while navy corresponds to increased gene expression. **D)** Relative average gene set enrichment scores of CD4+ ab T-cell clusters calculated by AUCell analysis of DEG sets from pairwise DGE analysis of the CD4T and CD4CD8T populations from porcine bulkRNA-seq. Scores are relative to other cells within a single gene set comparison (across a row of the heatmap) and are not calculated relative to scores across gene set (across columns in the heatmap). **E&F)** Genes with the largest effects in discriminating CD4+ ab T-cells by cluster identities were determined, as indicated by high permutation **(E)** and/or impurity scores **(F)** calculated from a trained random forest model. Average relative expression for each of these genes within clusters is also depicted by a heatmap. **G)** Dot plot of up to the top 20 DEGs having logFC > 0 from overall DGE analysis of only CD4+ ab T-cell clusters. Clusters are listed on the y-axis, while selected DEGs are listed on the

1569 x-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene.

1570 The color of a dot corresponds to the average relative expression level for the gene in the cells

1571 expressing the gene within a cluster. *Refer to ‘Gene name replacement’ methods.

1572

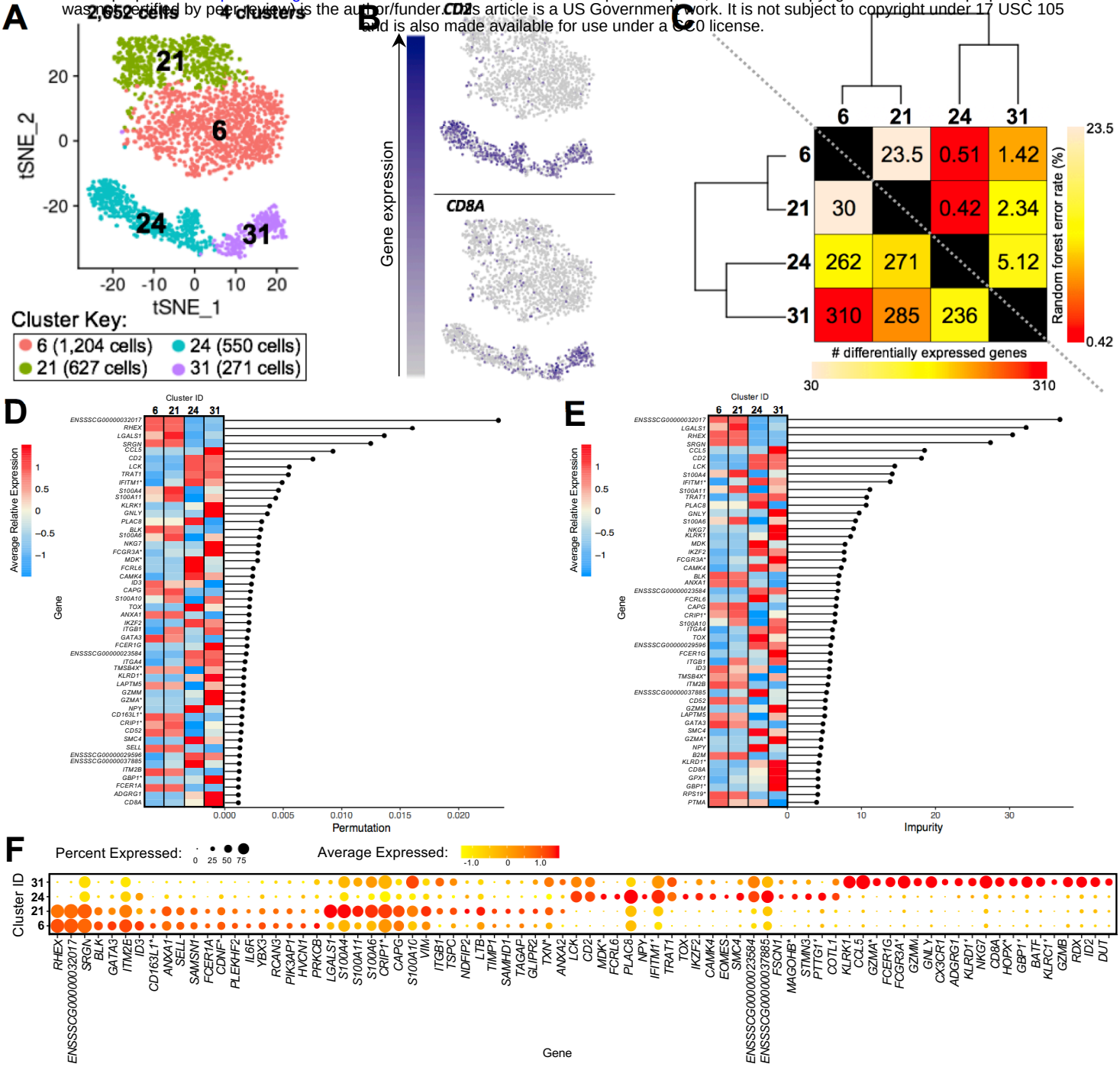


Figure 8. Transcriptional heterogeneity of porcine gd T-cells at single-cell resolution. **A)** Two-dimensional t-SNE plot of 2,652 cells belonging to clusters designated as CD2- gd T-cells (clusters 6, 21) or CD2+ gd T-cells (clusters 24, 31) in Figure 4D. Each point represents a single cell. Color of the cell corresponds to transcriptional cluster a cell belongs to. Cells more transcriptionally similar to each other belong to the same cluster. **B)** Visualization of selected gene expression overlaid onto t-SNE coordinates of single gd T-cells. Each point represents a single cell. Color of the point corresponds to relative expression of a specified gene (top left of each t-SNE plot) within a cell. Grey corresponds to little/no gene expression, while navy corresponds to increased gene expression. **C)** Transcriptomic relationship amongst gd T-cell clusters as calculated by three methods: hierarchical clustering (as seen by hierarchical trees on both axes), pairwise random forest analyses (as seen on top right diagonal); and pairwise DGE analyses (as seen on bottom left diagonal). Longer branches on the hierarchical tree corresponds to greater hierarchical distance. Lower numbers of DEGs by DGE analysis and higher out-of-bag (OOB) error rates from random forest analyses indicate greater pairwise transcriptional similarity. **D&E)** Genes with the largest effects in discriminating gd T-cells by cluster identities were determined, as indicated by high permutation **(D)** and/or impurity scores **(E)** calculated from a trained random forest model. Average relative expression for each of these genes within clusters is also depicted by a heatmap. **F)** Dot plot of up to the top 20 DEGs having logFC > 0 from overall DGE analysis of only gd T-cell clusters. Clusters are listed on the y-axis, while selected DEGs are listed on the x-axis. The size of a dot corresponds to the percent of cells in a cluster that expressed the gene. The color of a dot corresponds to the average relative expression level for the gene in the cells expressing the gene within a cluster. *Refer to 'Gene name replacement' methods.