Supplementary Methods

Cell sorting

Single cell suspensions of duodenal biopsies were used fresh for cell sorting. PCs were stained with multimerised TG2 and the DGP PLQPEQPFP using APC-labelled streptamers (IBA) and PerCP-conjugated SA (Biolegend), respectively[1, 2]. Following multimers staining, the cells were labelled with anti-CD3 BV570, anti-CD14 Pacific Blue (Biolegend), anti-CD4 APC-H7, anti-CD11c Horizon V450 (BD Bioscience), anti-CD27 PE-Cy7 (eBioscience) and anti-IgA FITC (Southern Biotech) and violet viability dye (ThermoFisher Scientific). All antibodies apart from anti-IgA were monoclonal. IgA intestinal PCs were defined as live, large CD11c/CD14/CD3-negative with high expression of CD27 (Supplementary Figure 5). TG2-PCs and non-TG2-PCs were sorted using a FACSAria II (BD Bioscience). The numbers of sorted PCs are given in Table 1. To verify the purity of the sorted PCs we have searched for gene markers of potential contamination from T cells, monocytes, dendritic cells, eosinophils and stromal cells; such genes were not found in the data (Figure 1).

RNA-seq

RNA was extracted from sorted PCs subsets using RNeasy micro kit (Qiagen) and RNA quality and quantity were determined using RNA 6000 Pico kit (Agilent Technologies). The number of sorted cell and RNA quantity and integrity are shown in Table 1. Approximately 600 ng of RNA were used for cDNA synthesis. cDNA synthesis and amplification (15x cycles) was performed using SMARTer® Ultra® Low Input RNA Kit for Sequencing-v3 (Clontech Laboratories).Amplified cDNA was quantified using High Sensitivity DNA Kit (Agilent Technologies). Tagmentation and ligation of indexed adapter was achieved using NexteraXT library preparation kit (Illumina, Inc). Differently indexed samples were pooled and amplicon libraries were sequenced on NextSeq500 (Illumina, Inc). Two separated libraries (i.e. batches) were generated and were sequenced independently. Four patients were sequenced in the first batch and three patients and four disease controls (i.e. healthy donors) were sequenced in the second batch.

Gene expression quantification from RNA-seq libraries

Salmon version 0.7.2 was used for transcript quantification in quasi-mapping mode (k-mer length=31) with variational Bayesian EM algorithm for optimising abundance estimates. The index was built on the transcriptome of Ensembl genome build GRCh38 release version 86 that includes alternative loci. Read counts of transcripts (including those on alternative loci) were aggregated to gene-level. Built-in models in the Salmon tool that correct the sequence-specific biases and fragment-level GC biases were used.

To deal with the highly variable sequences of the Ig genes the following approach was taken: First, to capture as much known variation as possible into the mapping step, we incorporated the known variation patches (alternate sequences) of the GRCh38 genome build into the mapping step. The Salmon tool that we used for quantification uses a concept named “chain of maximal exact matches” to deal with mismatches such as point mutations as well as insertions and deletions (InDels). The tool would find a reasonably specific anchoring match between a read and a transcript unless point mutations are evenly distributed in a very small region (e.g., a point mutation occurring every 5 bases or 10 bases). If this is not the case, (i.e., if point mutations are evenly distributed in a very small region), then mapping of reads from that specific region might suffer. Overall, more than 80% of the total reads produced were mapped on average. We expect that the mapping may not suffer for the majority of non-IG protein coding genes. For those non-Ig genes, there might be a spurious mapping if there is a really high sequence similarity with Ig sequences.

Pre-filtering of genes for downstream analyses

To identify genes (other than IG genes) that are expressed in PCs, we followed a similar approach as described in reference[3]. Briefly, using finite normal mixture models implemented in Mclust package[4], we performed a model-based clustering of the regularised log-transformed expression values of all protein-coding genes in each sample to categorise the genes into two classes that can be considered as either “expressed” or “not expressed”. Next, only genes that

were considered “expressed” in at least half of the samples were defined as “expressed genes” and used for downstream analyses. This exercise has been carried out independently for each subgroup of PCs.

Functional annotation using gene ontology categories

The enrichment of gene ontology (GO) categories was tested using Bioconductor package GoSeq version 1.32[5]; testing for categories belonging only to “Biological Process”. The analysis was restricted to GO level 3 when obtaining a global overview and to levels 3-9 to get detailed information of biological process. After multiple testing correction, significant GO terms were filtered further for redundancy using REVIGO[6].

Differential expression analysis

To examine if batch variations were present in the generated libraries, we analysed the linear relationship, and distances between libraries. Specifically, Pearson’s correlation coefficient between all the libraries and Euclidean distance based unsupervised clustering were generated and examined prior to our analyses of the data. The Pearson’s correlation coefficient between all PC samples ranged between 0.92 to 1.0 (Supplementary Figure 6A), while the samples of batch-1 had somewhat higher intra-correlation. This picture of batch-1 samples having less intra-batch distances among themselves is also evident through the distance-based unsupervised clustering (Supplementary Figure 6B). These batch variations and interindividual variations were paid due attention and accounted for in the differential expression analyses as described in relevant sections below.

Differential expression analysis was performed using Bioconductor package DESeq2 version 1.20[7]. When comparing the transcriptional profiles of TG2-PCs and non-TG2-PCs of CeD patients, we added the paired sample information as one of the covariates to the design formula. This accounted for both the inter-individual and batch variations, as patient samples were assayed in two different batches. Each patient’s TG2-PC sample and non-TG2-PC sample however, were processed on the same batch.

When comparing the transcriptional profiles of CeD patients and disease controls, as there was no paired information, the contribution of batch variation towards differential expression testing was minimised through a different approach as described below. Disease control samples were processed in batch-2, whereas the CeD patient samples were processed in both bacth-1 and batch-2. To minimise the batch-induced effect, we computed the differences in average expression levels of control and batch-1 CeD samples and similarly for batch-2 CeD samples and retained only those genes that did not exhibit a substantial difference between both those numbers. Genes were called differentially expressed at a FDR of 10%.

References:

1. Snir O, Mesin L, Gidoni M, et al. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J Immunol* 2015; 194: 5703-5712. DOI: 10.4049/jimmunol.1402611.

2. Snir O, Chen X, Gidoni M, et al. Stereotyped antibody responses target posttranslationally modified gluten in celiac disease. *JCI Insight* 2017; 2. DOI: 10.1172/jci.insight.93961.

3. Hebenstreit D, Fang M, Gu M, et al. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* 2011; 7: 497. DOI: 10.1038/msb.2011.28.

4. Scrucca L, Fop M, Murphy TB, et al. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* 2016; 8: 289-317.

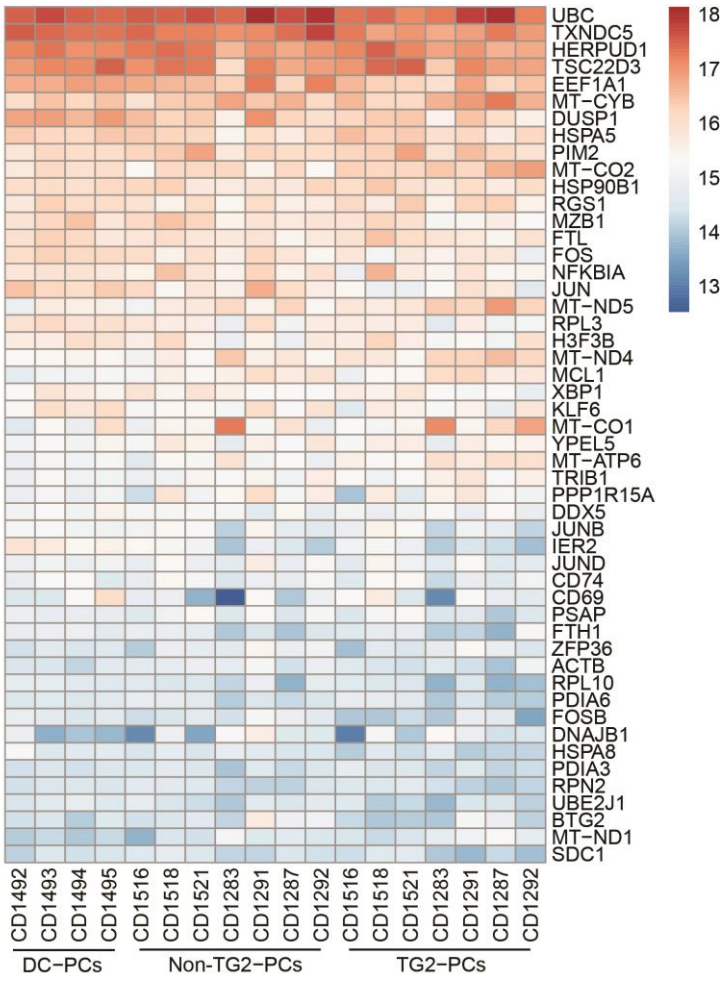
5. Young MD, Wakefield MJ, Smyth GK, et al. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology* 2010; 11: R14. DOI: 10.1186/gb-2010-11-2-r14.

6. Supek F, Bosnjak M, Skunca N, et al. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 2011; 6: e21800. DOI: 10.1371/journal.pone.0021800.

7. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014; 15: 550. DOI: 10.1186/s13059-014-0550-8.

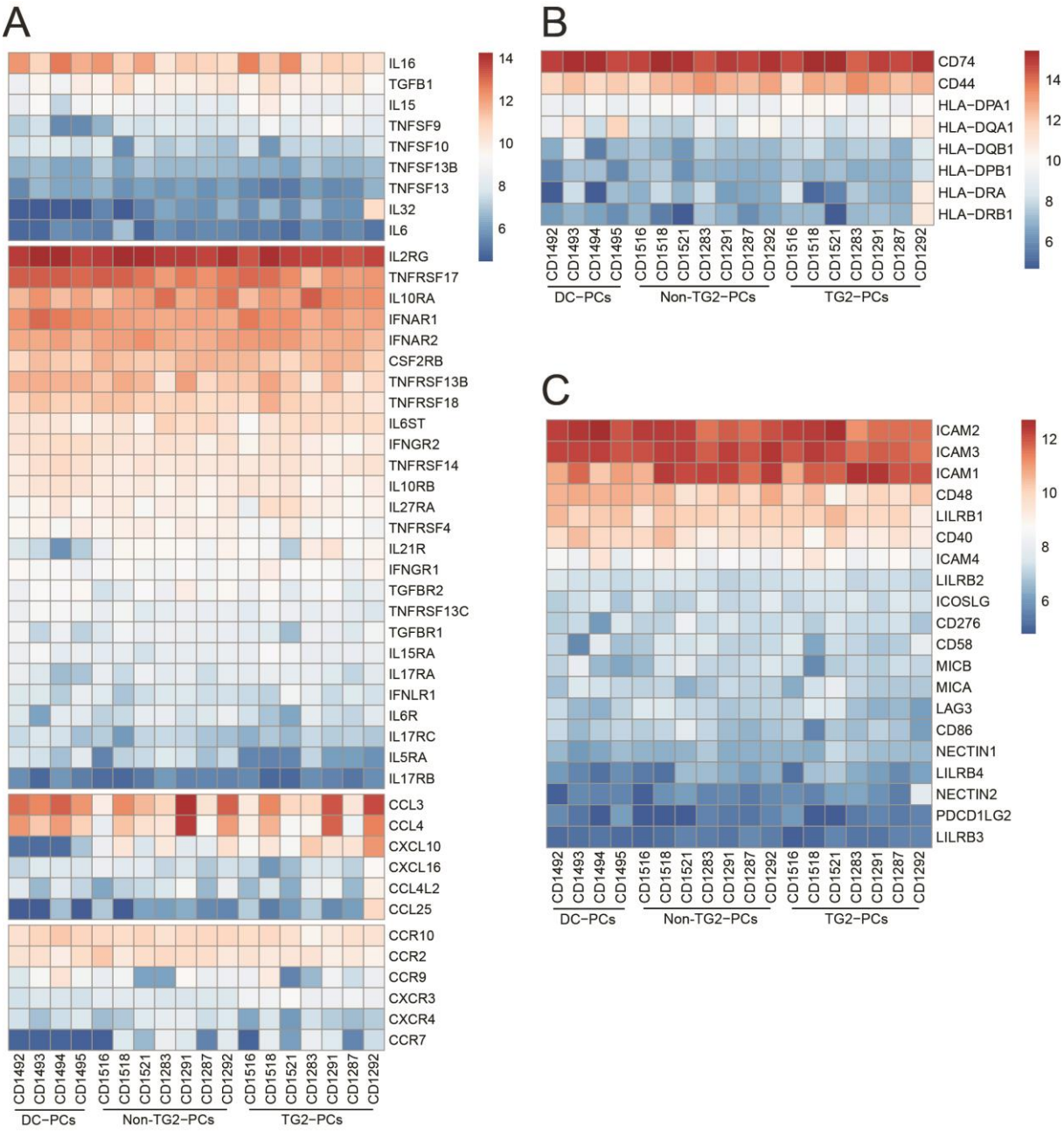
Supplementary figure 1 – Individual representation of the top fifty-expressed genes

The top fifty-expressed genes in small-intestinal of IgA PCs of DC-PCs, non-TG2-PCs and TG2-PCs. The colour scale shows the normalized expression counts on a log2 scale (regularized logarithmic values reported by DESeq2).



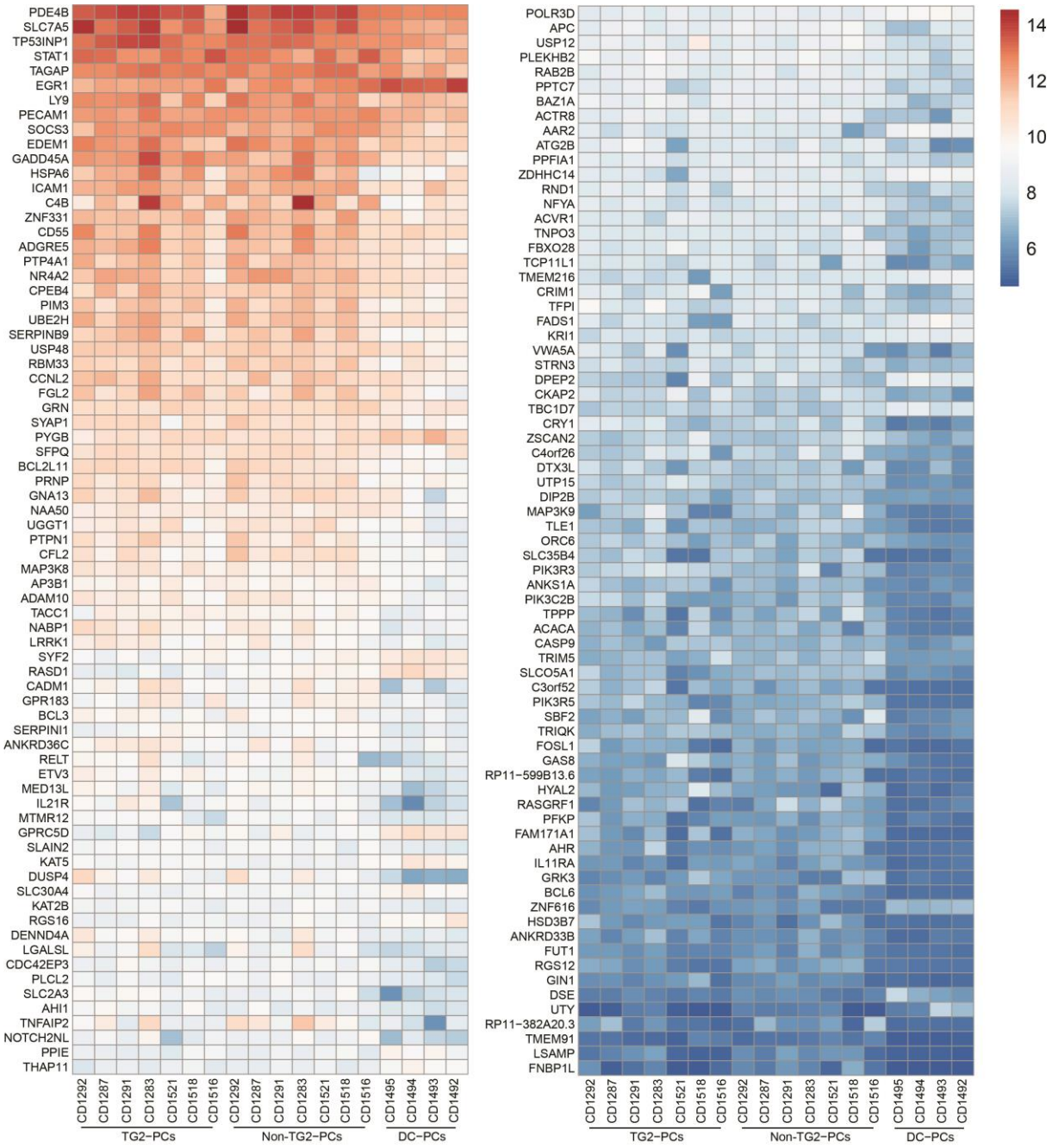
Supplementary figure 2 – Individual representation of immune-related mediators and receptor genes that are expressed in PCs

Expression profile of the immune related genes in IgA PCs of DC-PCs, non-TG2-PCs and TG2-PCs: (A) cytokine, cytokine receptor as well as chemokines and chemokines receptors. (B) Co-stimulatory molecules and (C) HLA class II genes and associated. The colour scale shows the normalized expression counts on a log2 scale (regularized logarithmic values reported by DESeq2).



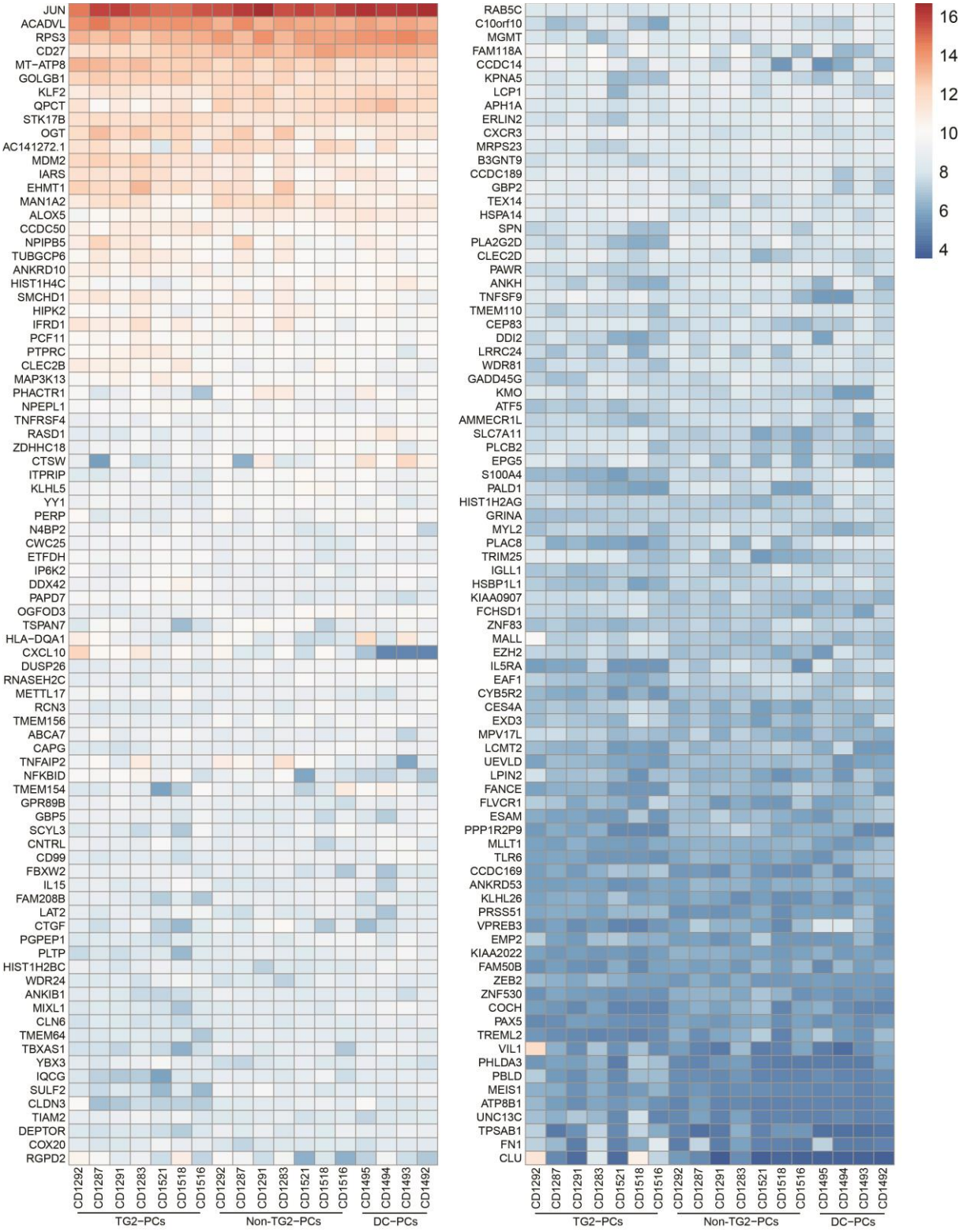
Supplementary figure 3 – Individual representation of differently expressed genes in CD vs. disease controls

An individual representation of the differentially expressed genes (n=141 based on fold-difference) in PCs from CeD patients (non-TG2- and TG2-PCs) and disease controls. The colour scale shows the normalized expression counts on a log2 scale (regularized logarithmic values reported by DESeq2).



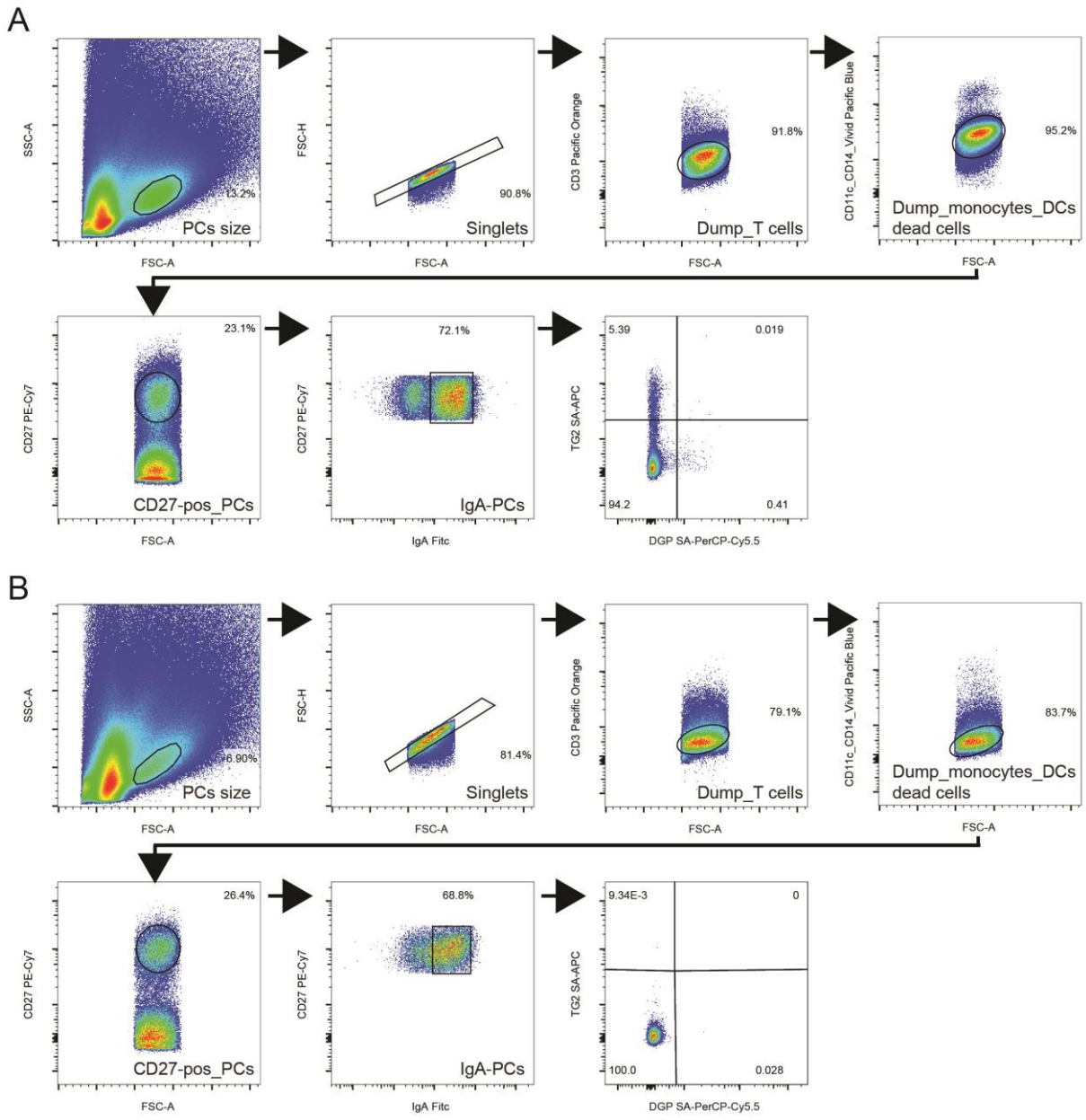
Supplementary figure 4 – Differently expressed genes in TG2-PCs vs. non-TG2-PCs

An individual representation of the expression profile of the differentially expressed genes (n=151) in TG2-PCs in comparison with non-TG2-PCs. The colour scale shows the normalized expression counts on a log2 scale (regularized logarithmic values reported by DESeq2).



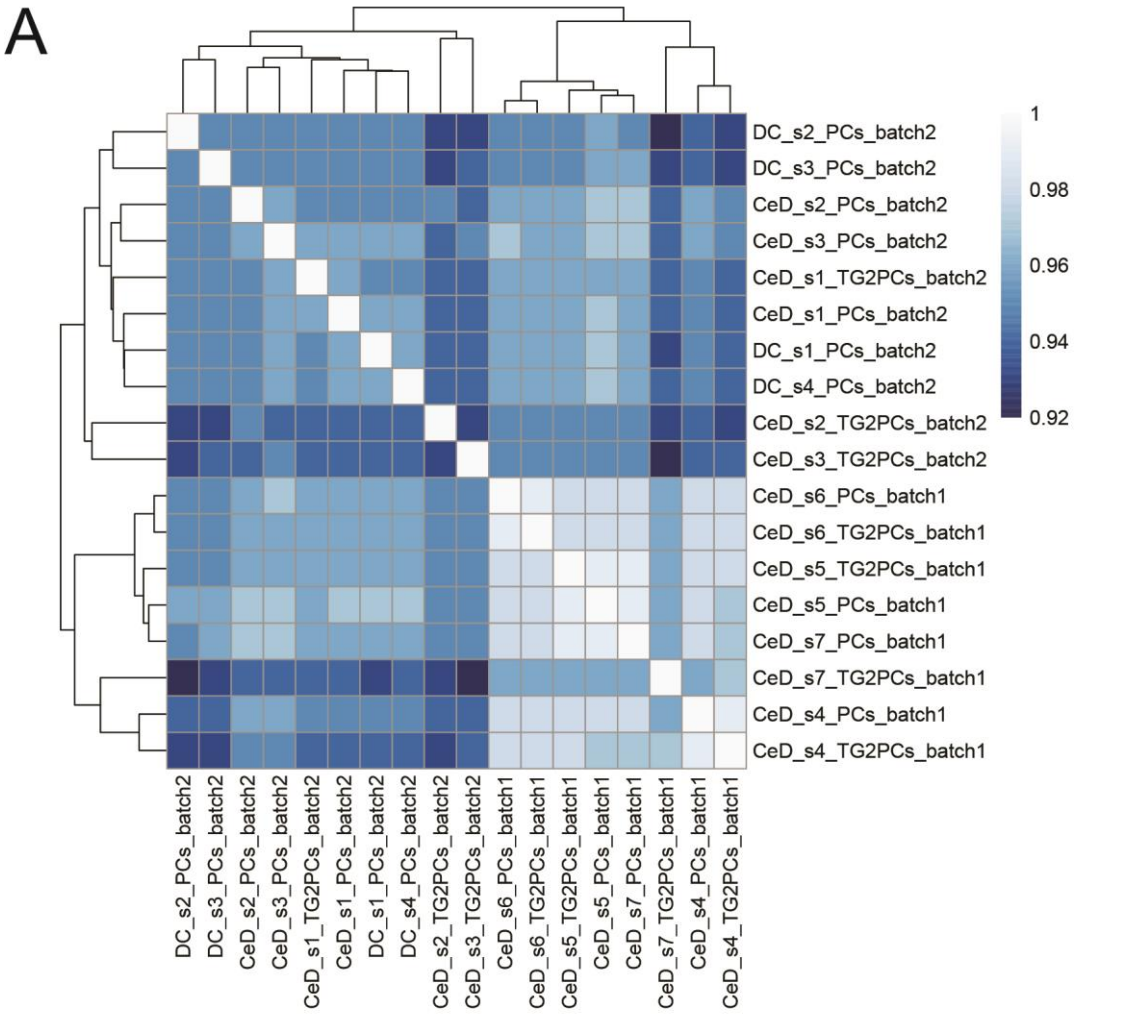
Supplementary figure 5 - Sorting IgA-PCs for transcription analysis – gating strategy

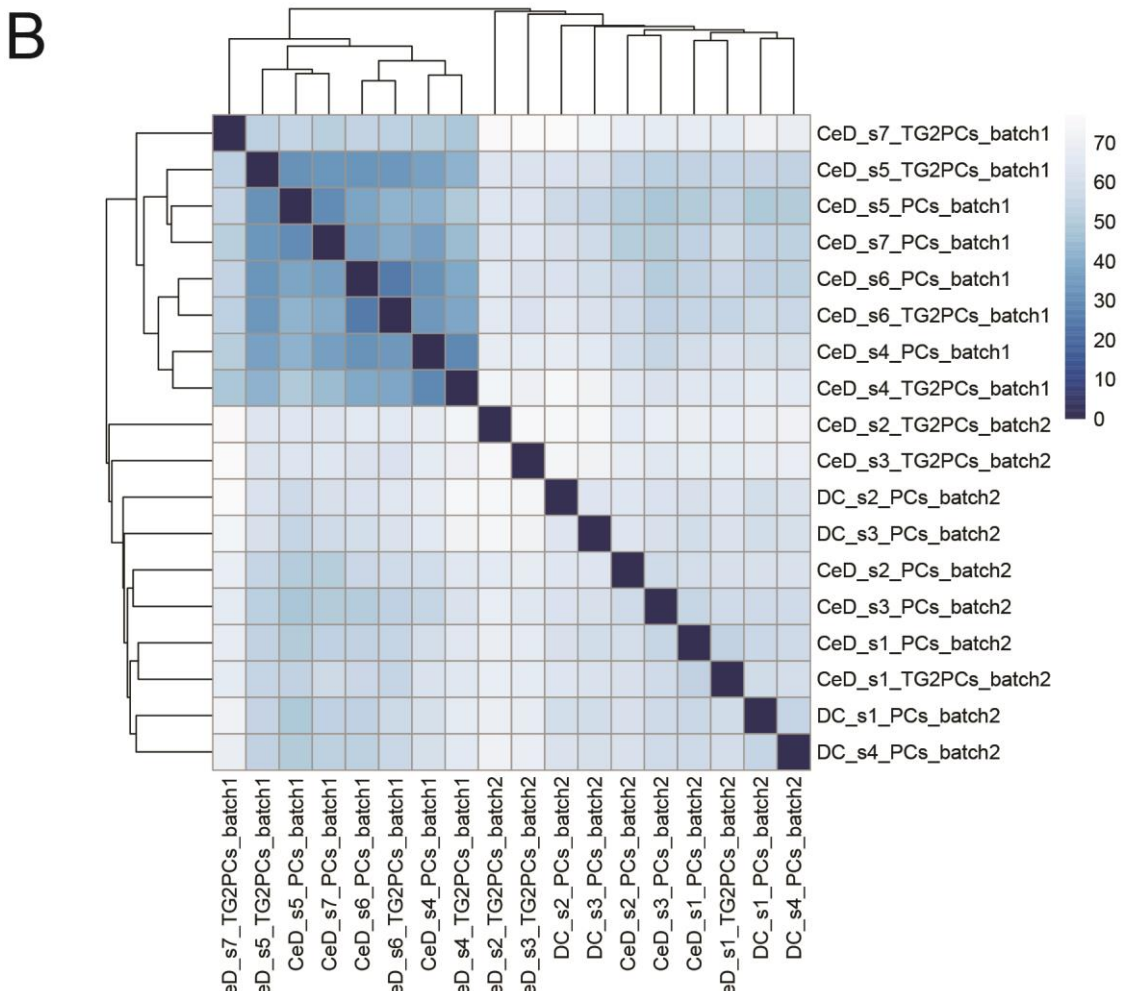
Gating strategy that was taken to isolate IgA PCs from small intestinal biopsies of (A) CD patients and (B) disease controls; PCs that are specific or non-specific to TG2 were sorted (i.e. TG2-PCs and non-TG2-PCs, respectively). PCs were defined as large, live cells expressing CD27 and IgA on their surface. CD3, CD11c and CD14 were used to exclude T cell, monocyte/macrophages and dendritic cells. Dead cells were exploded using viability dye. Arrows indicate sequential gating.



Supplementary figure 6 – Pairwise correlation and unsupervised distance-based clustering between all PC samples

(A) Pearson's correlation coefficient was computed pairwise for all the generated PC libraries using the expressed genes and displayed as a matrix. The correlation coefficient ranged from 0.92 to 1.0. (B) The distances between the transcriptional profiles of all the PC samples were computed and distance-based unsupervised clustering was performed. The euclidean distance, unlike a correlation coefficient does not have a constant range of values. It is low for highly similar samples and high for highly dissimilar elements. The colour scale represents the Euclidean distance between the samples.


A
DC_s2_PCs_batch2
DC_s3_PCs_batch2
CeD_s2_PCs_batch2
CeD_s3_PCs_batch2
CeD_s1_TG2PCs_batch2
CeD_s1_PCs_batch2
DC_s1_PCs_batch2
DC_s4_PCs_batch2
CeD_s2_TG2PCs_batch2
CeD_s3_TG2PCs_batch2
CeD_s6_PCs_batch1
CeD_s6_TG2PCs_batch1
CeD_s5_TG2PCs_batch1
CeD_s5_PCs_batch1
CeD_s7_PCs_batch1
CeD_s7_TG2PCs_batch1
CeD_s4_PCs_batch1
CeD_s4_TG2PCs_batch1
1
0.98
0.96
0.94
0.92


B
CeD_s7_TG2PCs_batch1
CeD_s5_TG2PCs_batch1
CeD_s5_PCs_batch1
CeD_s7_PCs_batch1
CeD_s6_PCs_batch1
CeD_s6_TG2PCs_batch1
CeD_s4_PCs_batch1
CeD_s4_TG2PCs_batch1
CeD_s2_TG2PCs_batch2
CeD_s3_TG2PCs_batch2
DC_s2_PCs_batch2
DC_s3_PCs_batch2
CeD_s2_PCs_batch2
CeD_s3_PCs_batch2
CeD_s1_PCs_batch2
CeD_s1_TG2PCs_batch2
DC_s1_PCs_batch2
DC_s4_PCs_batch2
70
60
50
40
30
20
10
0