SUPPLEMENTARY DATA

Checklist for Reporting Human Islet Preparations Used in Research

Adapted from Hart NJ, Powers AC (2018) Progress, challenges, and suggestions for using human islets to understand islet biology and human diabetes. Diabetologia https://doi.org/10.1007/s00125-018-4772-2.

Manuscript DOI: https://doi.org/10.2337/**[insert manuscript submission number]** **(**Example, https://doi.org/10.2337/db18-1234)	
Title: Early detection of peripheral blood cell signature in children developing beta-cell autoimmunity at a young age	
Author list: Henna Kallionpää, Juhi Somani, Soile Tuomela, Ubaid Ullah, Rafael de Albuquerque, Tapio Lönnberg, Elina Komsi, Heli Siljander, Jarno Honkanen, Taina Härkönen, Aleksandr Peet, Vallo Tillmann, Vikash Chandra, Mahesh Kumar Anagandula, Gun Frisk, Timo Otonkoski, Omid Rasool, Riikka Lund, Harri Lähdesmäki, Mikael Knip, Riitta Lahesmaa	
Corresponding author: Riitta Lahesmaa	**Email address:** rilahes@utu.fi

Islet preparation	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**[a]
MANDATORY INFORMATION								
Unique identifier	H1883	H1900	H1994	H1997				
Donor age (years)	69	58	75	64				
Donor sex (M/F)	F	M	M	M				
Donor BMI (kg/m^2)								
Donor HbA_{1c} or other measure of blood glucose control								
Origin/source of islets[b]								

Islet isolation centre	NORDIC NETWORK FOR ISLET TRANSPLANTATION Uppsala,Sweden	NORDIC NETWORK FOR ISLET TRANSPLANTATION Uppsala,Sweden	NORDIC NETWORK FOR ISLET TRANSPLANTATION Uppsala,Sweden	NORDIC NETWORK FOR ISLET TRANSPLANTATION Uppsala,Sweden				
Donor history of diabetes? Yes/No								
If Yes, complete the next two lines if this information is available								
Diabetes duration (years)								
Glucose-lowering therapy at time of death[c]								

RECOMMENDED INFORMATION								
Donor cause of death								
Warm ischaemia time (h)								
Cold ischaemia time (h)								
Estimated purity (%)	95%	95%	98%	90%				
Estimated viability (%)								
Total culture time (h)[d]								
Glucose-stimulated insulin secretion or other functional measurement[e]								

SUPPLEMENTARY DATA

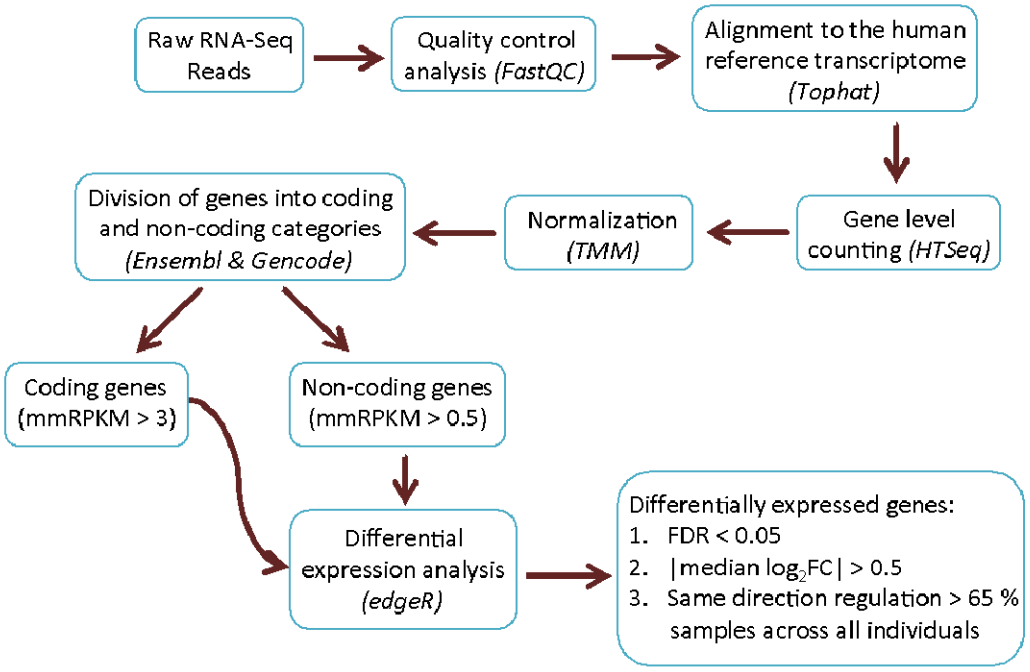
Handpicked to purity? Yes/No	Yes	Yes	Yes	Yes				
Additional notes								

[a]If you have used more than eight islet preparations, please complete additional forms as necessary
[b]For example, IIDP, ECIT, Alberta IsletCore
[c]Please specify the therapy/therapies
[d]Time of islet culture at the isolation centre, during shipment and at the receiving laboratory
[e]Please specify the test and the results

SUPPLEMENTARY DATA

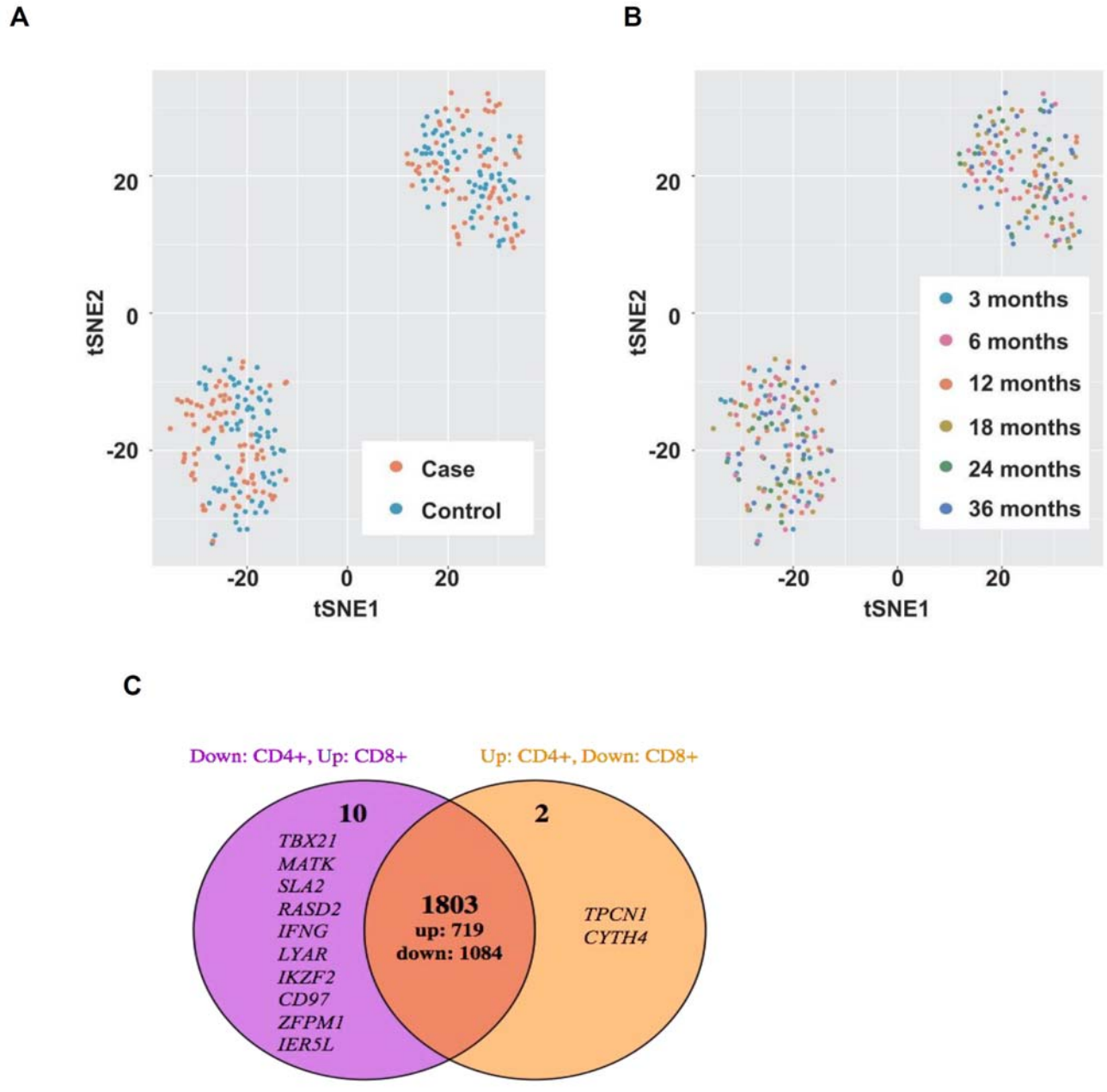
Early detection of peripheral blood cell signature in children developing beta cell autoimmunity at a young age

Supplementary Figure S1. Related to Figure 2.
Flow chart depicting the steps taken in the differential expression analyses of the RNA-seq data in this study.



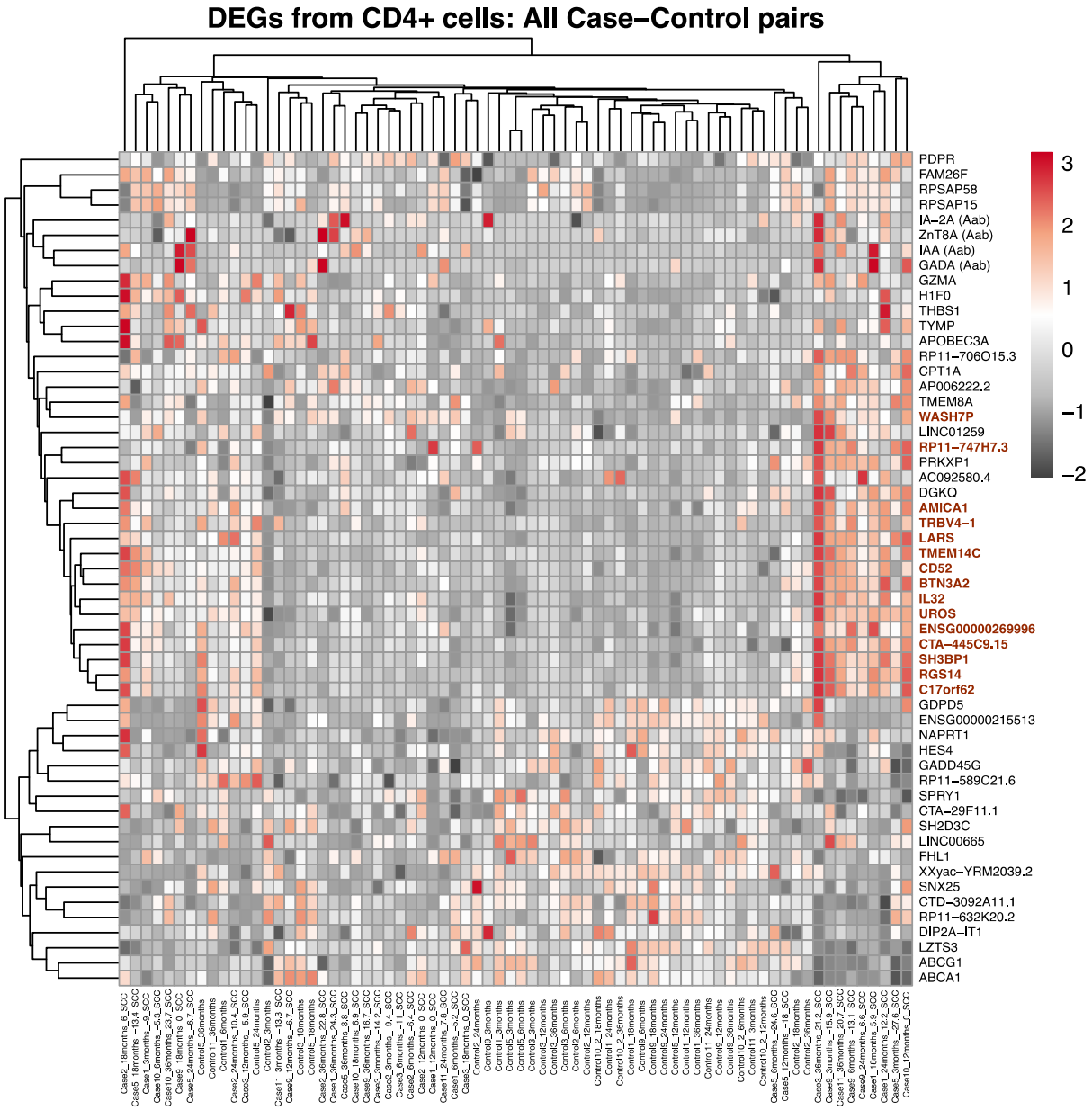
SUPPLEMENTARY DATA

Supplementary Figure S2. Related to Figure 1.
A-B) tSNE (t-Distributed Stochastic Neighbor Embedding) visualization of the log2-transformed expression data from all cell fractions and all genes. **Figure 1** was colored according to cell fractions, and here the colouring of the samples is done according to **A)** Case and Control status and **B)** age at sample collection. For further sample information, see **Table 1** and **Supplementary Table 1**. **C)** Venn diagram expanding on the 1815 genes found DE in both CD4+ vs PBMC and CD8+ vs PBMC analyses (**Figure 1C**). Here, the intersection represents the genes regulated in the same direction. For full lists of genes, see **Supplementary Table 2.**



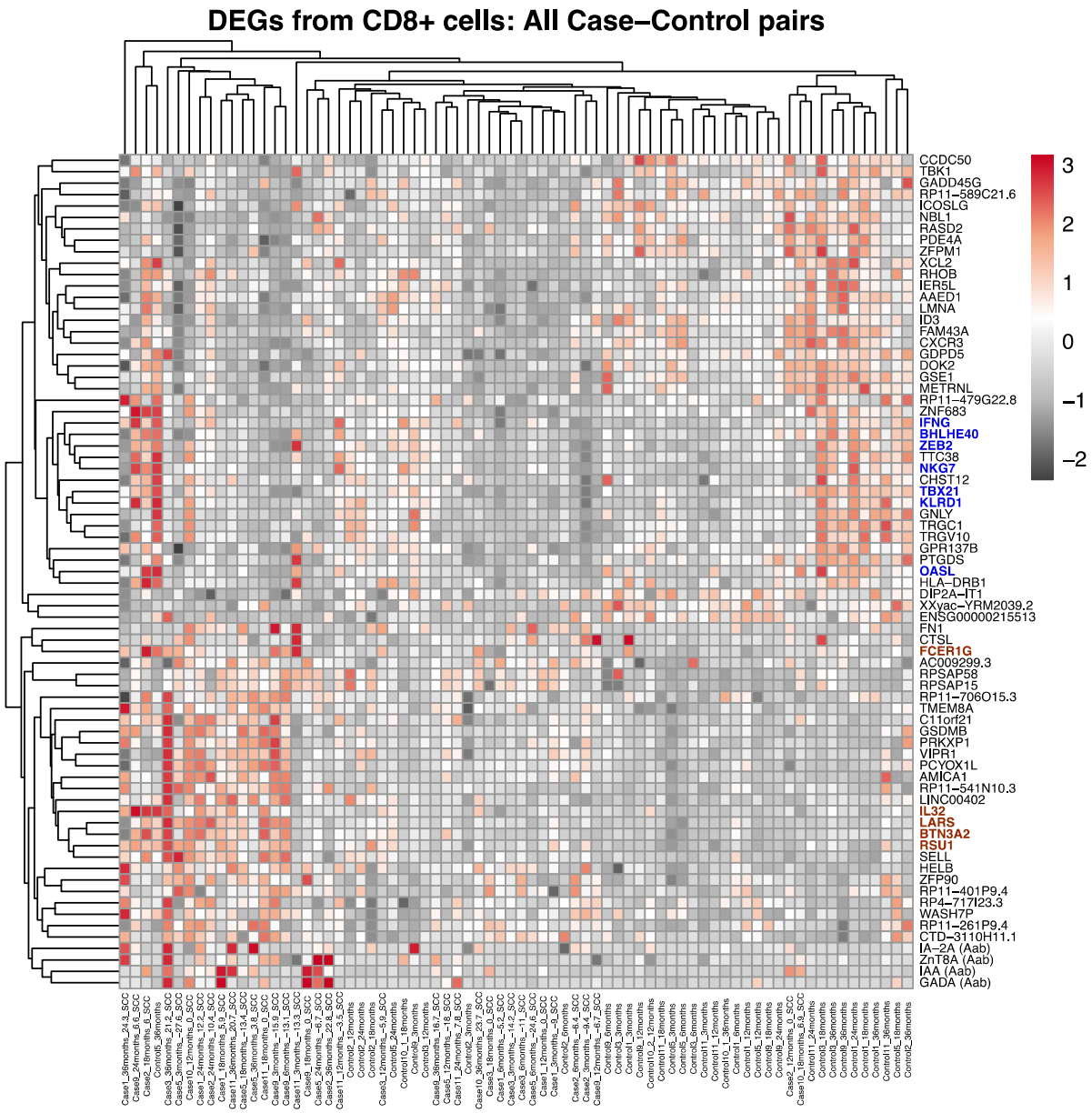
SUPPLEMENTARY DATA

Supplementary Figure S3A. Related to Figure 2.
Hierarchical clustering of the levels of standardized autoantibodies (IAA, IA-2A, ZnT8A, and GADA) and the 51 differentially expressed (DE) genes between the Cases and Controls detected in the CD4+ fraction. Each gene's expression was standardized across samples from each case-control pair individually. Genes with an Euclidean distance (ED)< 2.5 to IL-32 (co-clustering results from k-means clustering) are marked with red text (Supplementary table 4). The samples labels along the x-axis include the sample number, case/control indicator, age of sampling in months, and months to (negative no. of months) or from (positive no. of months) seroconversion time. Here, SCC stands for seroconversion-centered, which is why the months to/from seroconversion are negative or positive.



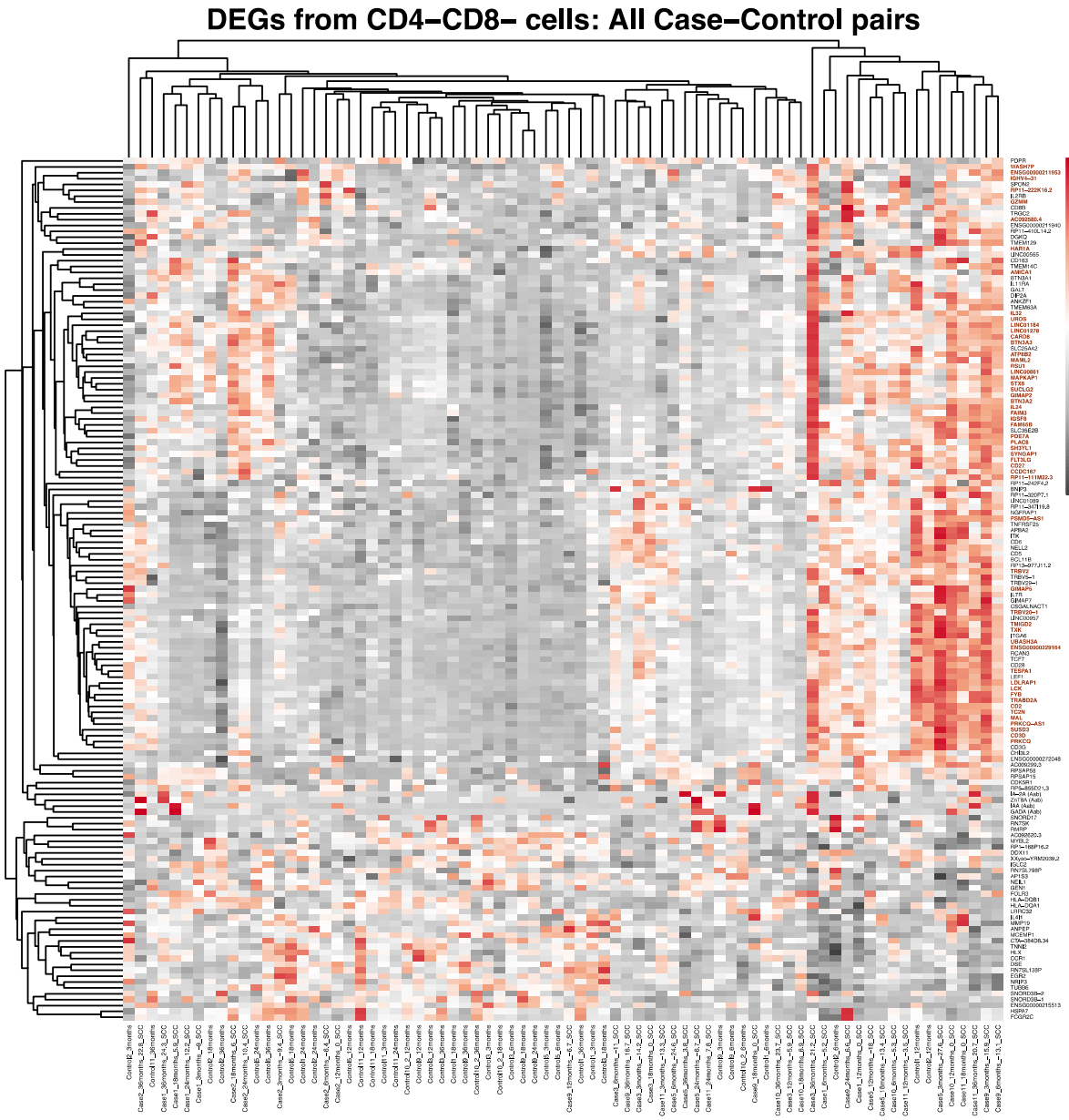
SUPPLEMENTARY DATA

Supplementary Figure S3B. Related to Figure 2.
Hierarchical clustering of the levels of standardized autoantibodies (IAA, IA-2A, ZnT8A, and GADA) and the 69 DE genes between the Cases and Controls detected in the CD8+ fraction. Each gene's expression was standardized across samples from each case-control pair individually. Genes with an Euclidean distance (ED) < 2.5 to IL-32 (co-clustering results from k-means clustering) are marked with red text and those with ED < 2.5 to IFNG are marked with blue text (Supplementary table 4). The samples labels along the x-axis include the sample number, case/control indicator, age of sampling in months, and months to (negative no. of months) or from (positive no. of months) seroconversion time. Here, SCC stands for seroconversion-centered, which is why the months to/from seroconversion are negative or positive.



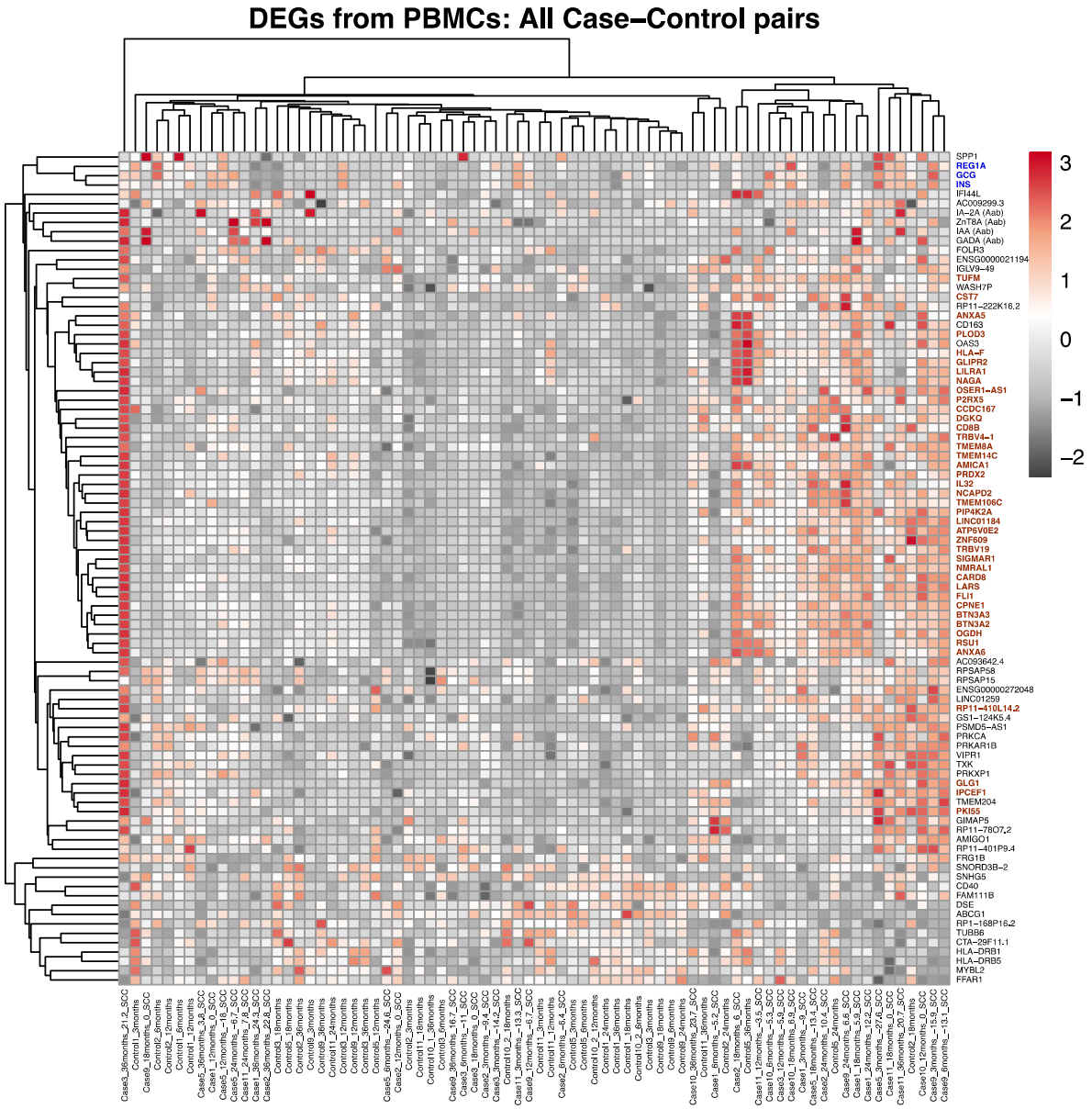
SUPPLEMENTARY DATA

Supplementary Figure S3C. Related to Figure 2.
Hierarchical clustering of the levels of standardized autoantibodies (IAA, IA-2A, ZnT8A, and GADA) and the 143 DE genes between the Cases and Controls detected in the CD4-CD8- fraction. The expression of each gene was standardized across samples from each case-control pair individually. Genes with an Euclidean distance (ED) < 2.5 to IL-32 (co-clustering results from k-means clustering) are marked with red text (Supplementary table 4). The samples labels along the x-axis include the sample number, case/control indicator, age of sampling in months, and months to (negative no. of months) or from (positive no. of months) seroconversion time. Here, SCC stands for seroconversion-centered, which is why the months to/from seroconversion are negative or positive.



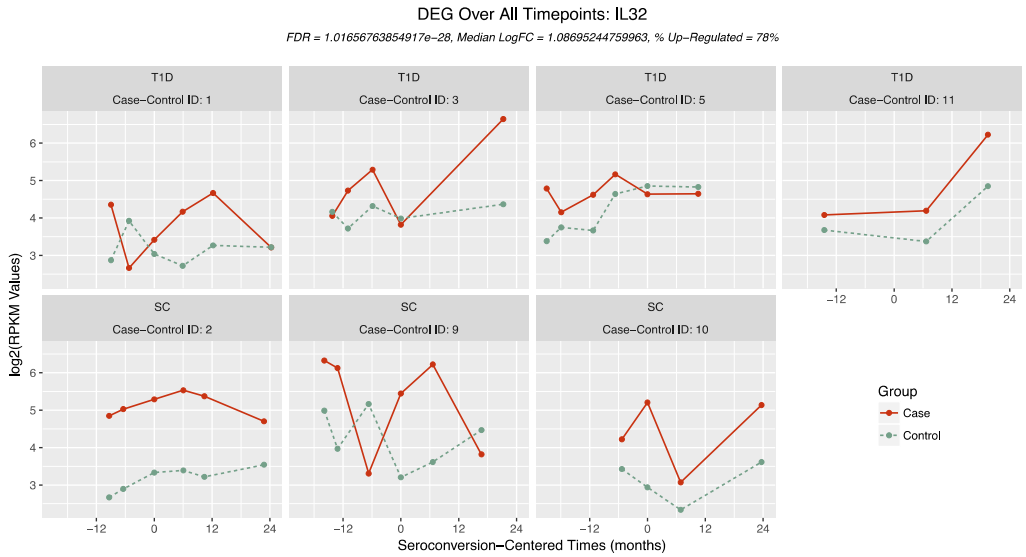
SUPPLEMENTARY DATA

Supplementary Figure S3D. Related to Figure 2.
Hierarchical clustering of the levels of standardized autoantibodies (IAA, IA-2A, ZnT8A, and GADA) and the 85 DE genes between the Cases and Controls detected in the PBMC population. Each gene's expression was standardized across samples from each case-control pair individually. Genes with an Euclidean distance (ED) < 2.5 to IL-32 (co-clustering results from k-means clustering) are marked with red text and those with ED < 2.5 to INS are marked with blue text (Supplementary table 4). The samples labels along the x-axis include the sample number, case/control indicator, age of sampling in months, and months to (negative no. of months) or from (positive no. of months) seroconversion time. Here, SCC stands for seroconversion-centered, which is why the months to/from seroconversion are negative or positive.

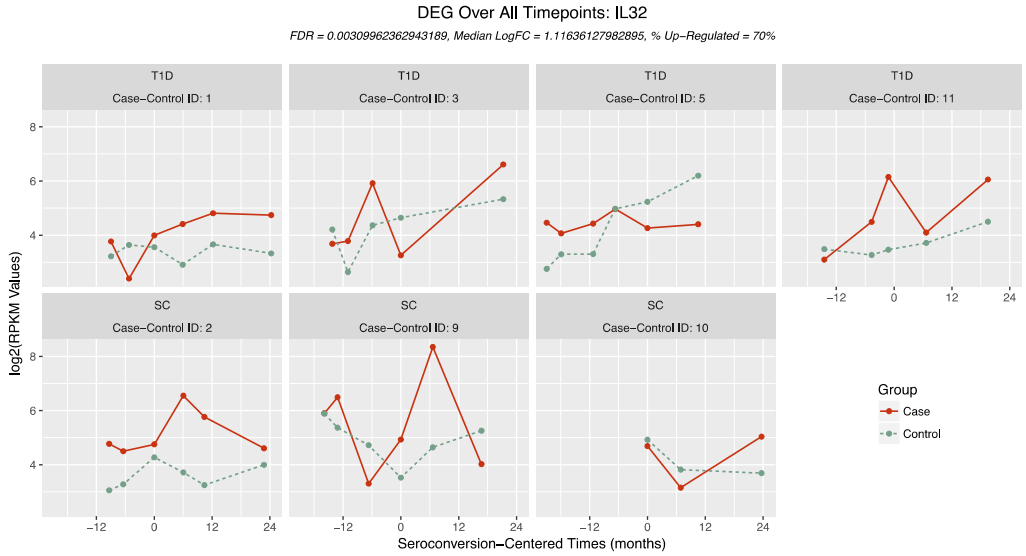


SUPPLEMENTARY DATA

Supplementary Figure S4A-W. Related to Figure 2.
Expression profile plots of genes highlighted in the manuscript:

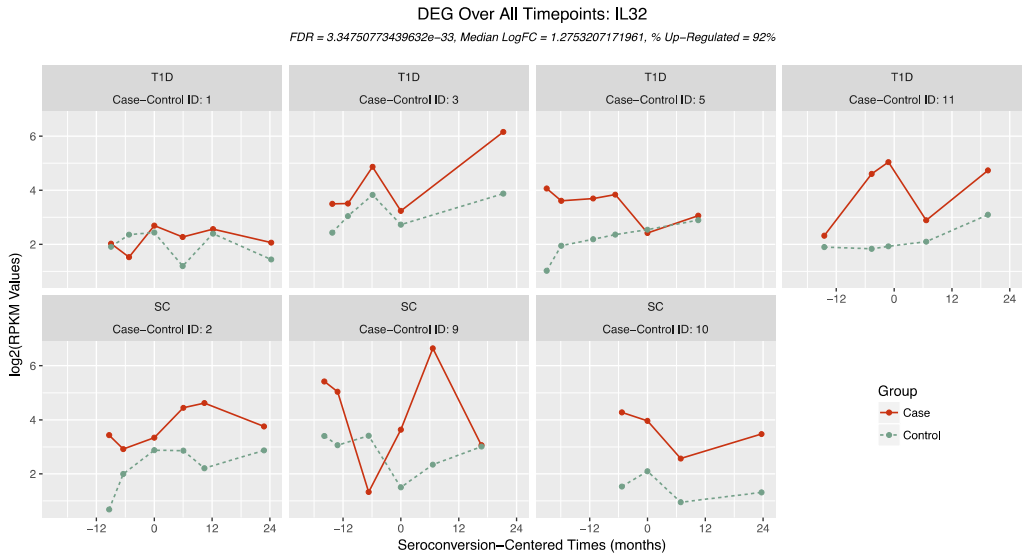


A) Expression levels of IL-32 gene in CD4+ cells.

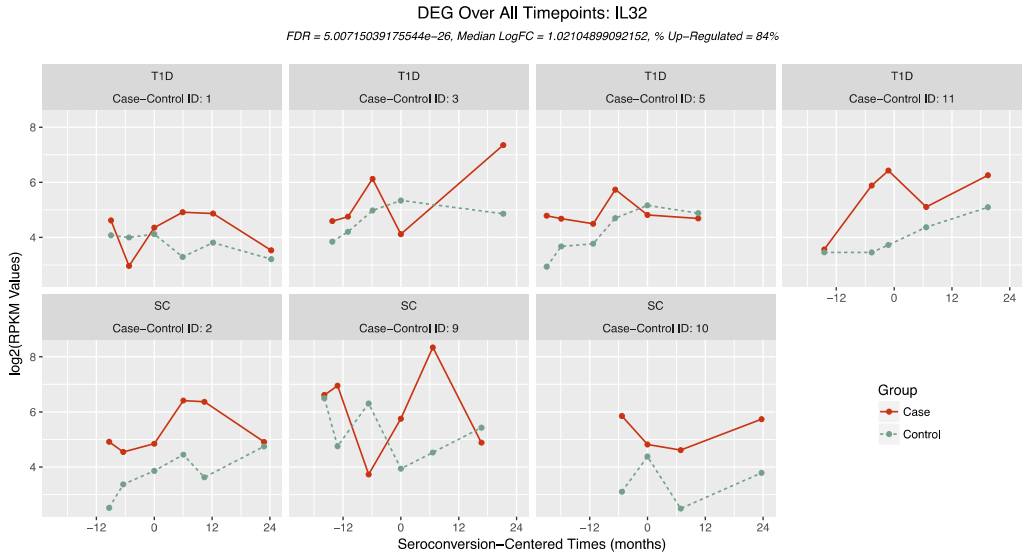


B) Expression levels of IL-32 gene in CD8+ cells.

SUPPLEMENTARY DATA

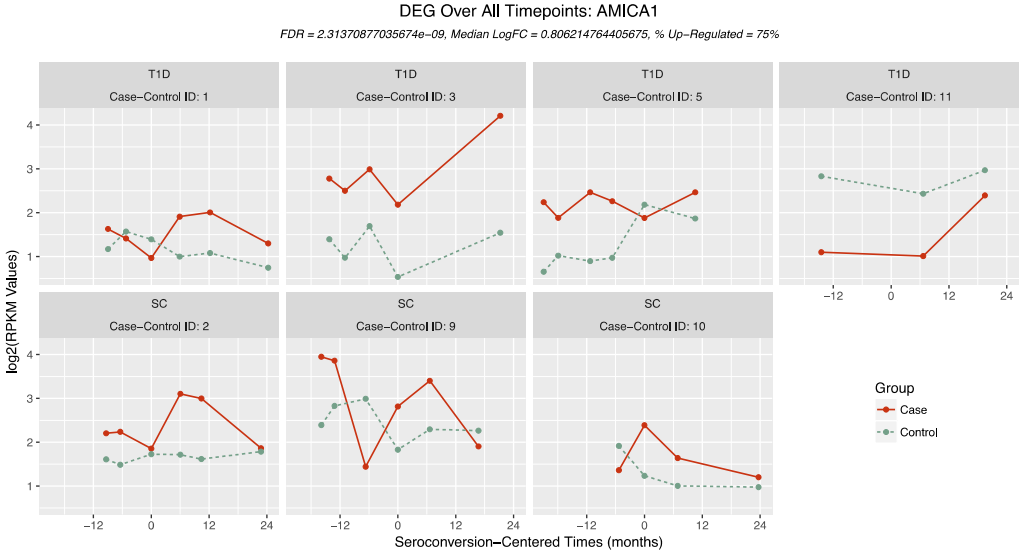


C) Expression levels of IL-32 gene in CD4-CD8- cells.

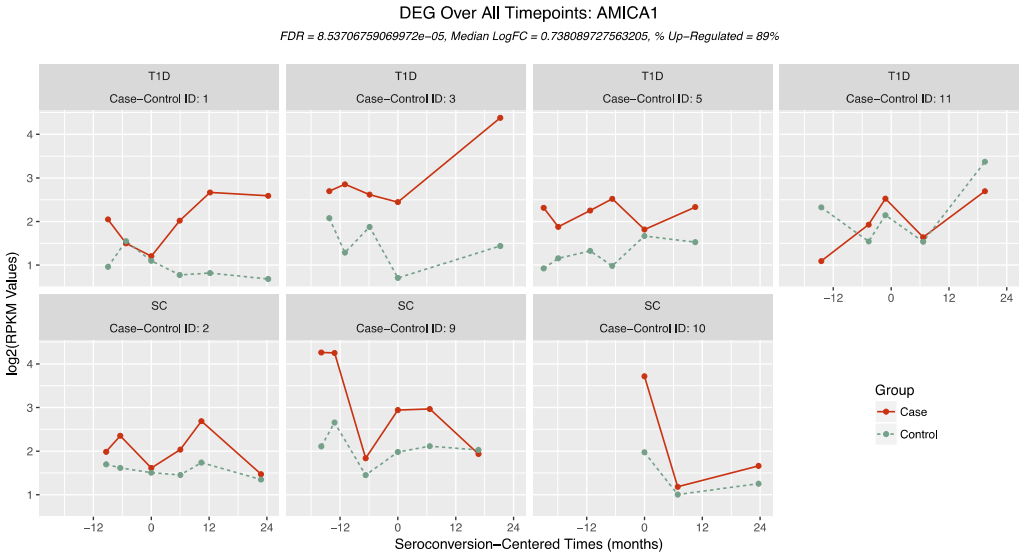


D) Expression levels of IL-32 gene in PBMCs.

SUPPLEMENTARY DATA

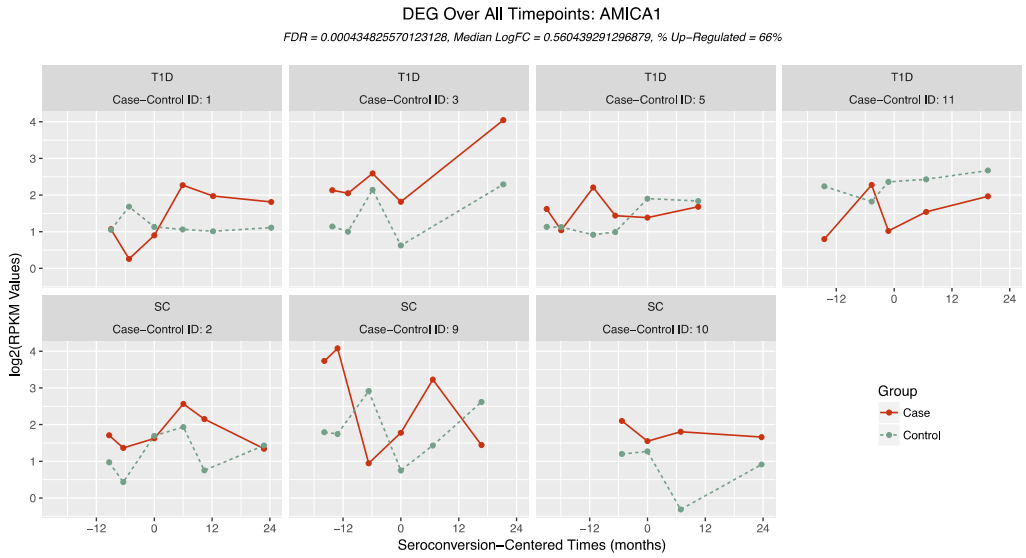


E) Expression levels of AMICA1 gene in CD4+ cells.

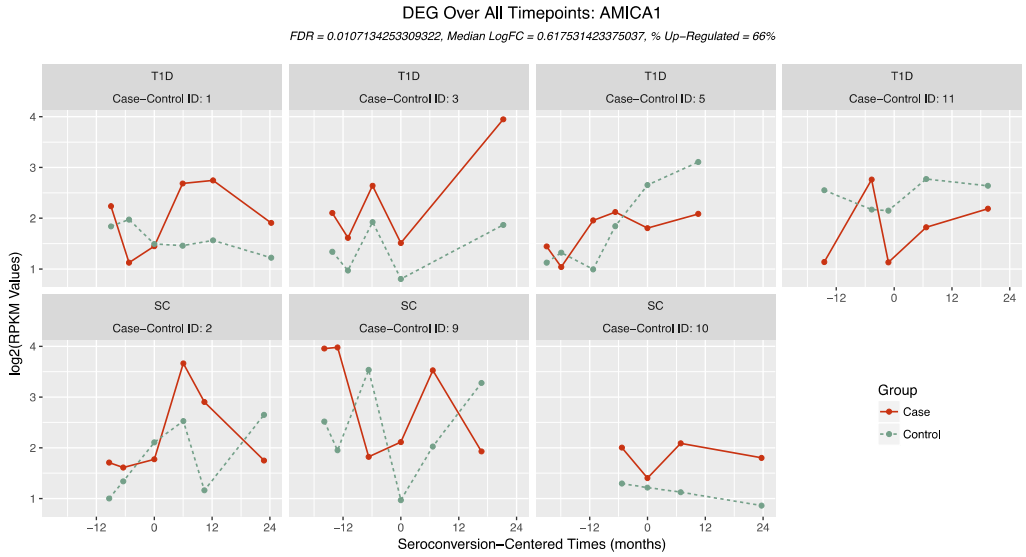


F) Expression levels of AMICA1 gene in CD8+ cells.

SUPPLEMENTARY DATA

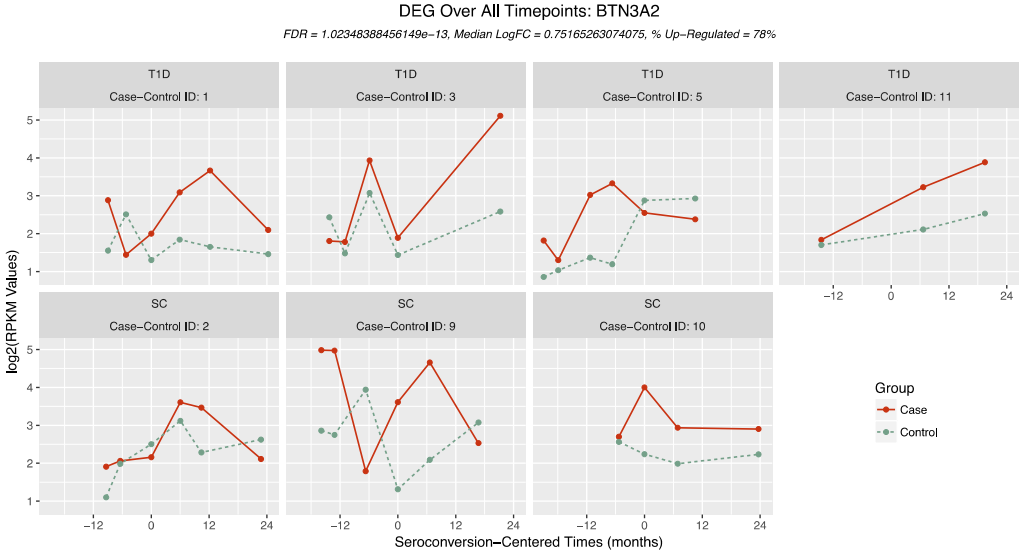


G) Expression levels of AMICA1 gene in CD4-CD8- cells.

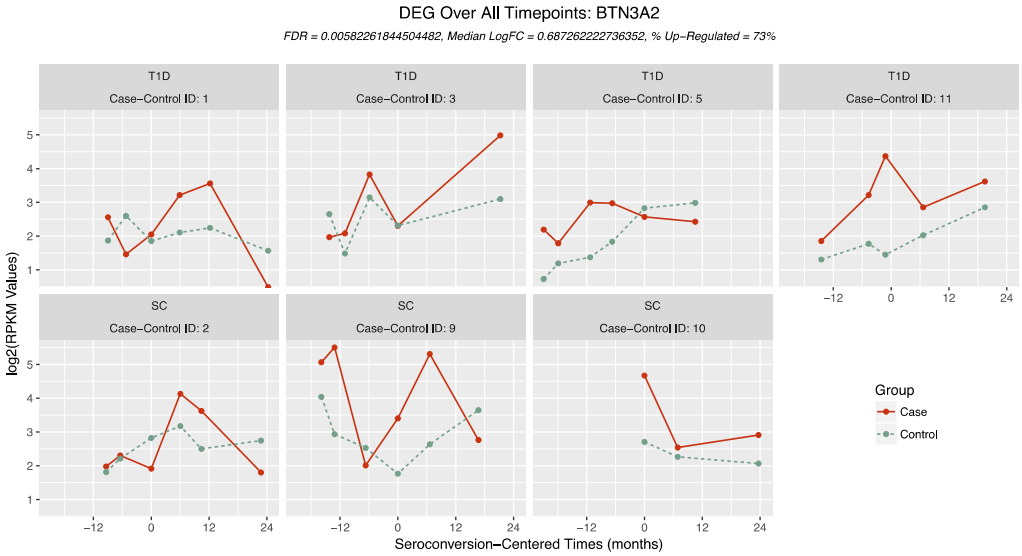


H) Expression levels of AMICA1 gene in PBMCs.

SUPPLEMENTARY DATA

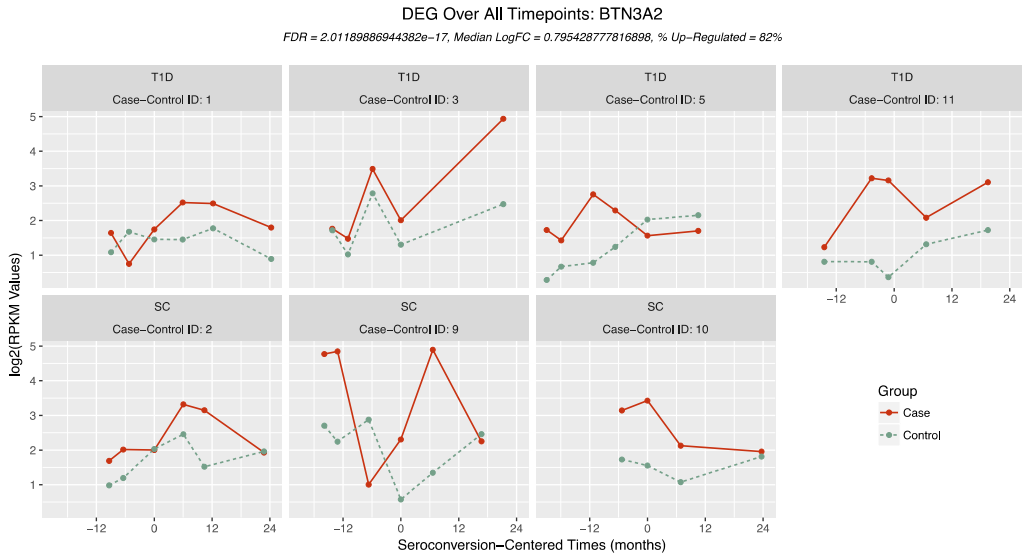


I) Expression levels of BTN3A2 gene in CD4+ cells.

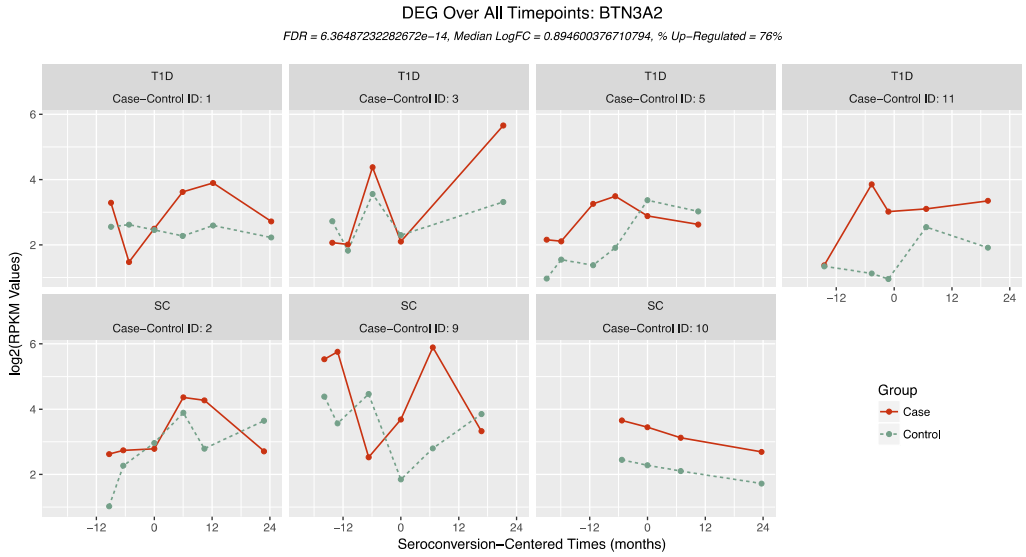


J) Expression levels of BTN3A2 gene in CD8+ cells.

SUPPLEMENTARY DATA

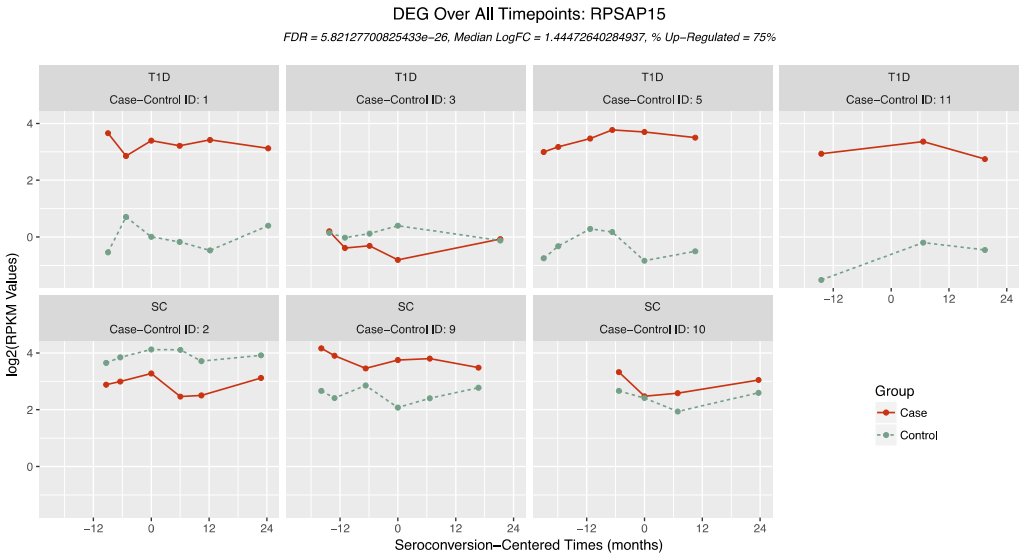


K) Expression levels of BTN3A2 gene in CD4-CD8- cells.

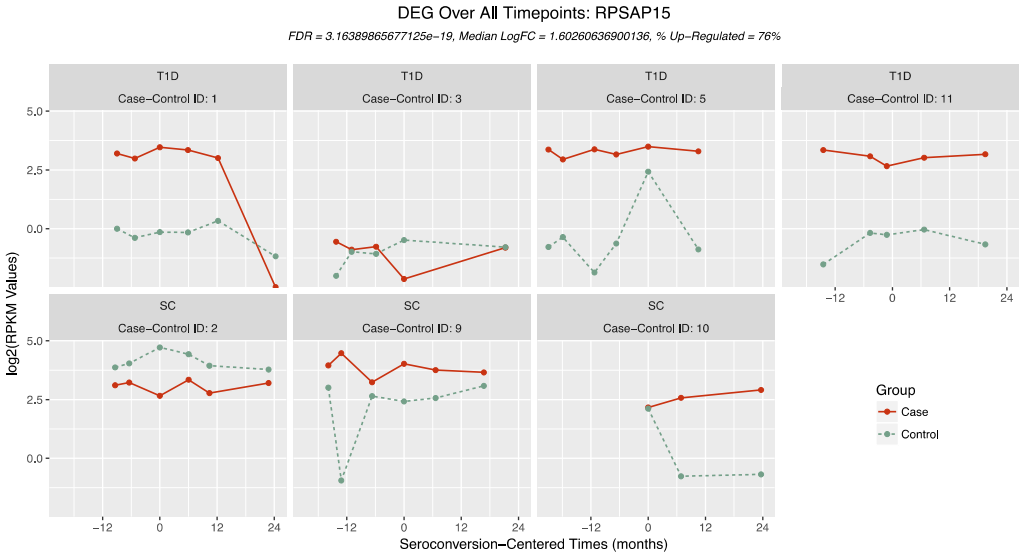


L) Expression levels of BTN3A2 gene in PBMCs.

SUPPLEMENTARY DATA

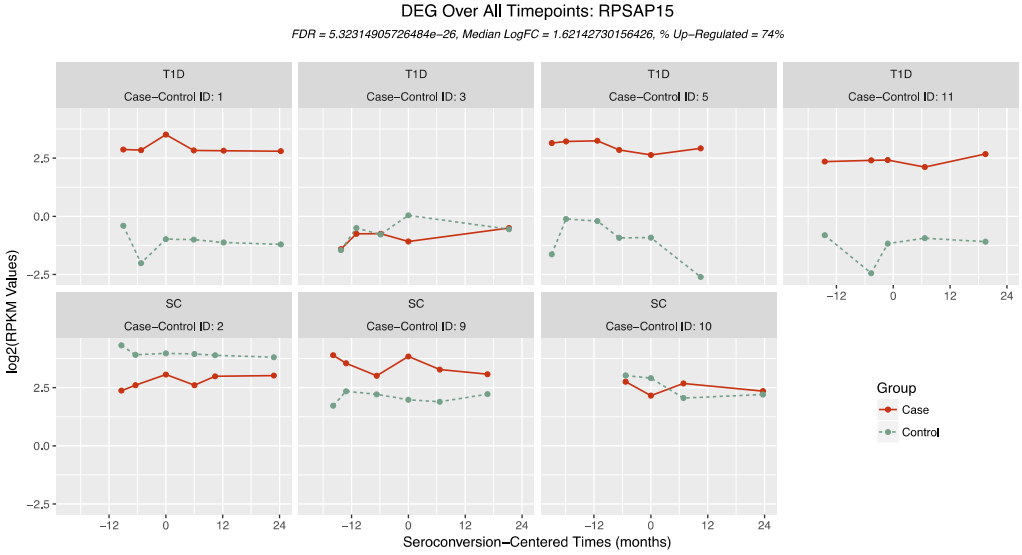


M) Expression levels of RPSAP15 gene in CD4+ cells.

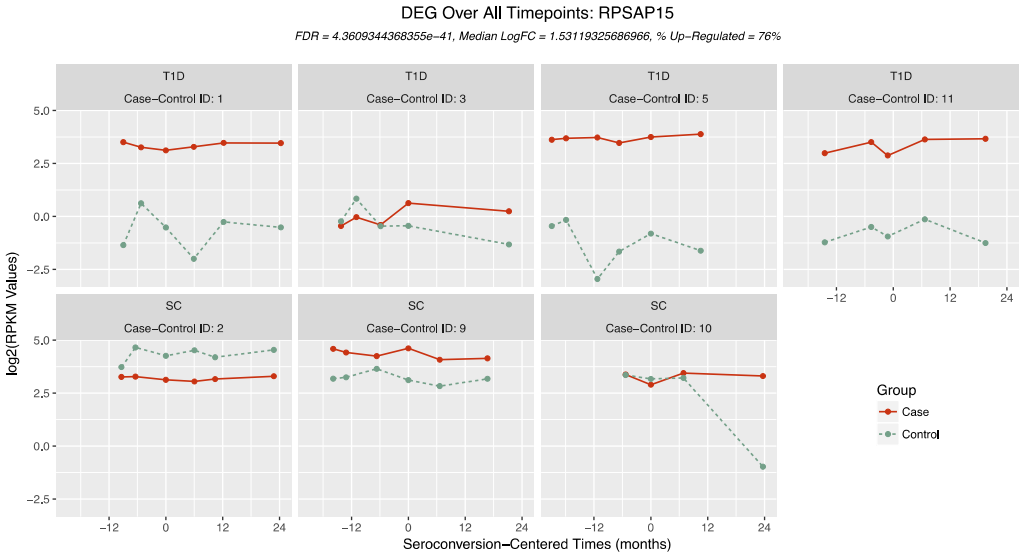


N) Expression levels of RPSAP15 gene in CD8+ cells.

SUPPLEMENTARY DATA

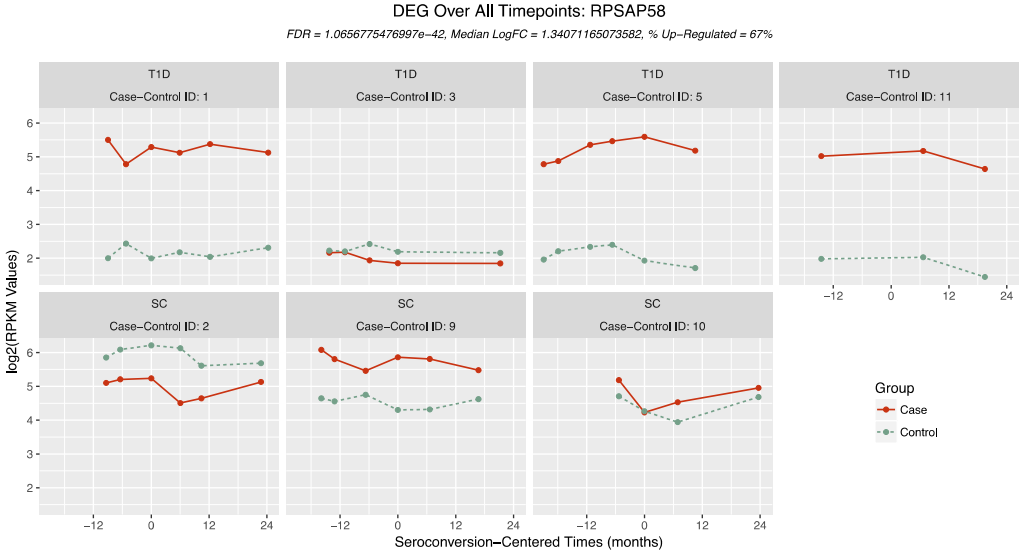


O) Expression levels of RPSAP15 gene in CD4-CD8- cells.

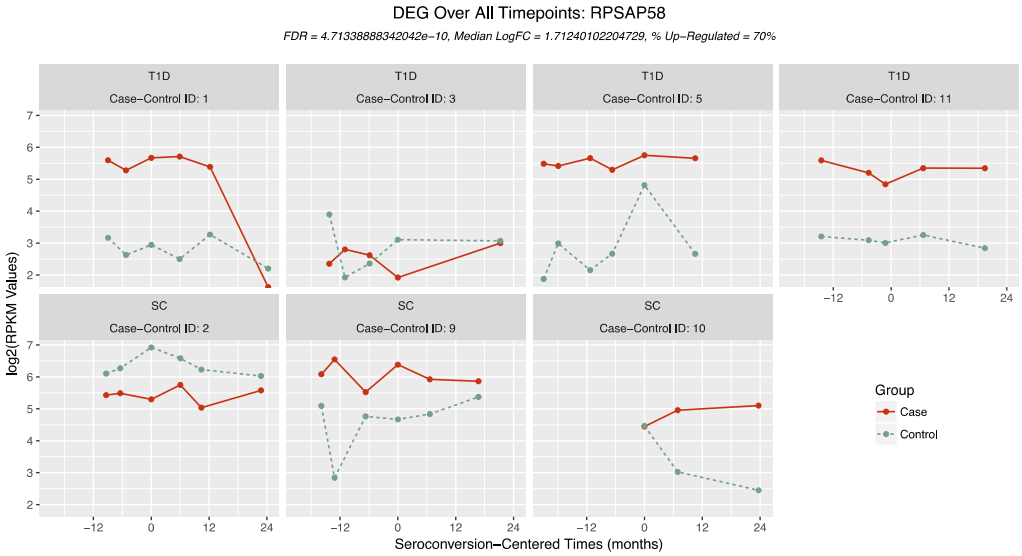


P) Expression levels of RPSAP15 gene in PBMCs.

SUPPLEMENTARY DATA

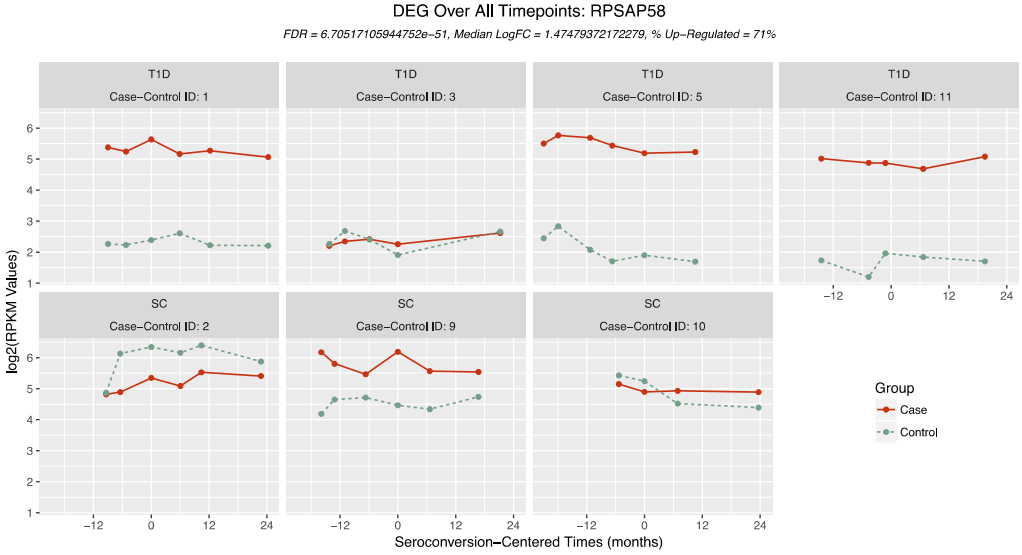


Q) Expression levels of RPSAP58 gene in CD4+ cells.

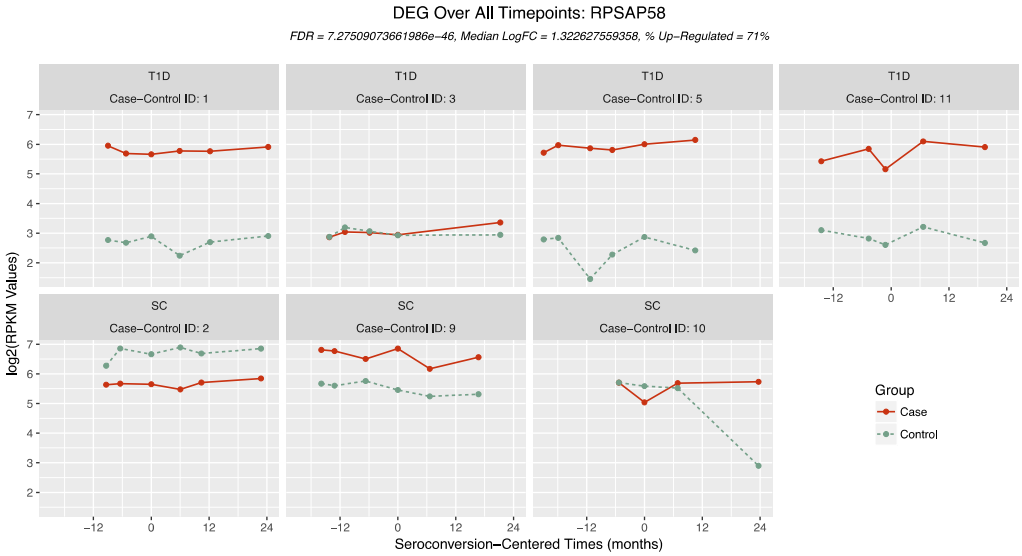


R) Expression levels of RPSAP58 gene in CD8+ cells.

SUPPLEMENTARY DATA

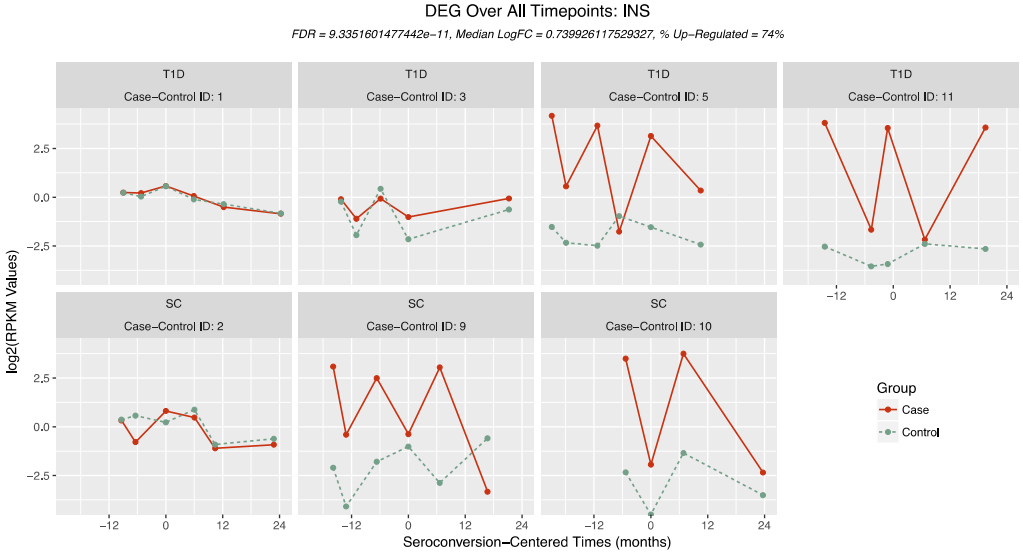


S) Expression levels of RPSAP58 gene in CD4-CD8- cells.

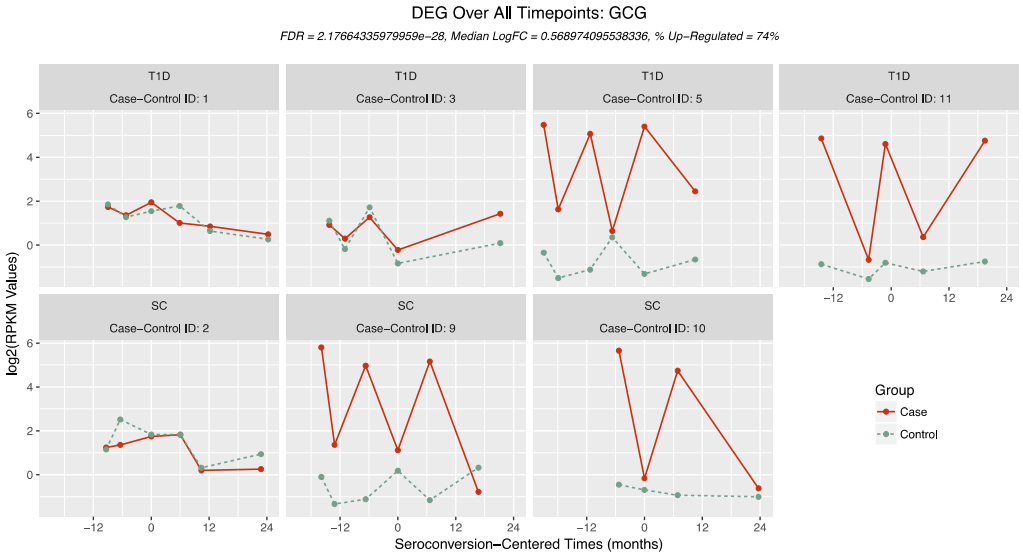


T) Expression levels of RPSAP58 gene in PBMCs.

SUPPLEMENTARY DATA

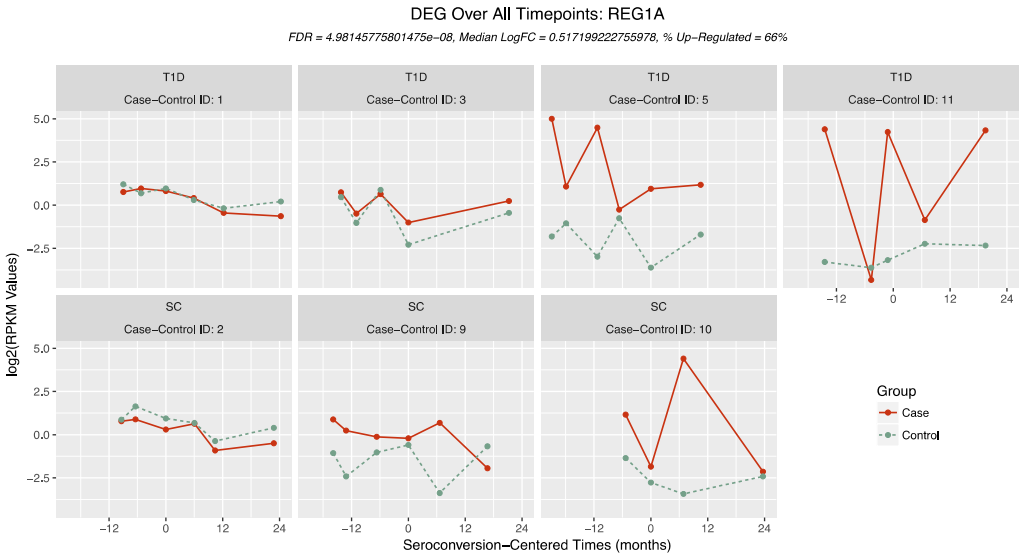


U) Expression levels of INS gene in PBMCs.

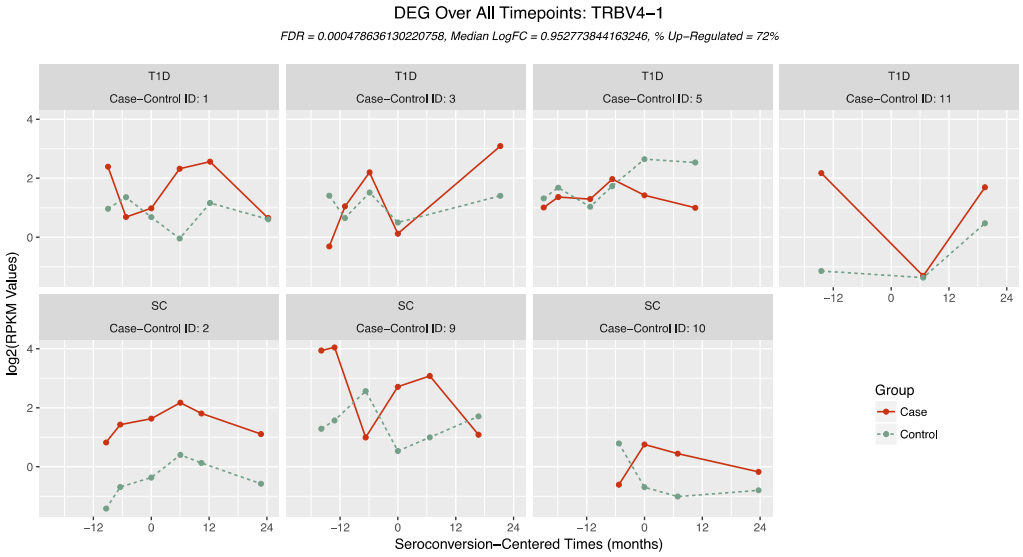


V) Expression levels of GCG gene in PBMCs.

SUPPLEMENTARY DATA

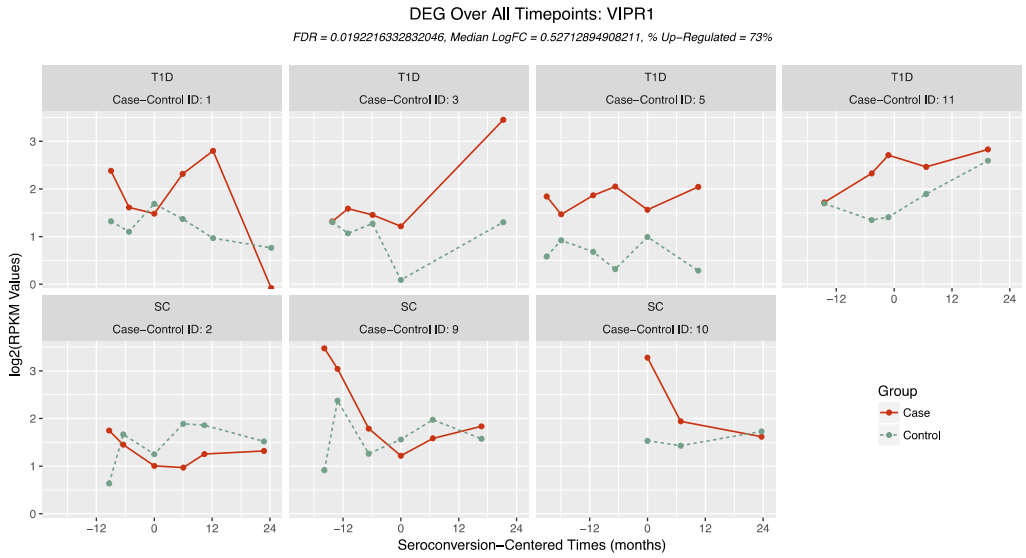


W) Expression levels of REG1A gene in PBMCs.

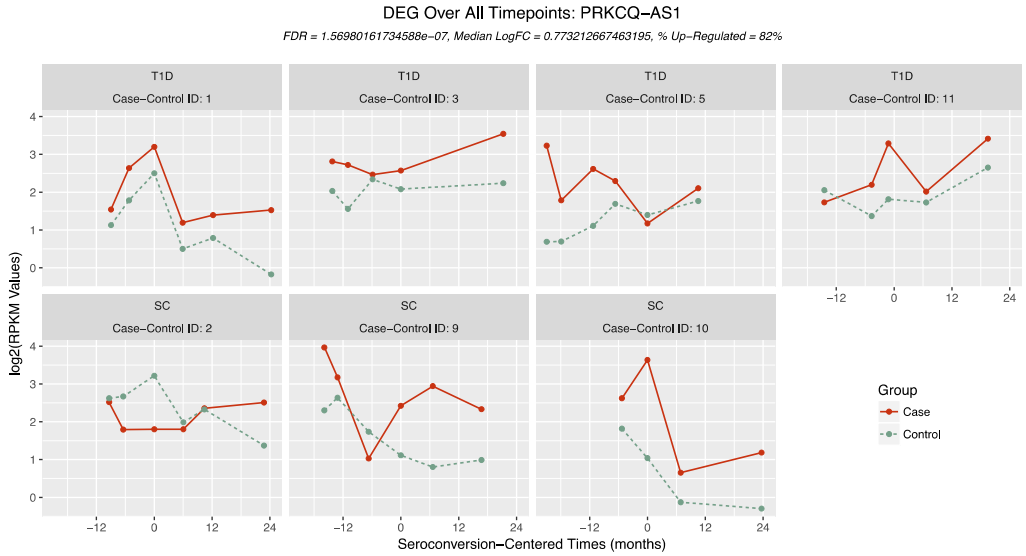


X) Expression levels of TRBV4-1 gene in CD4+ cells.

SUPPLEMENTARY DATA

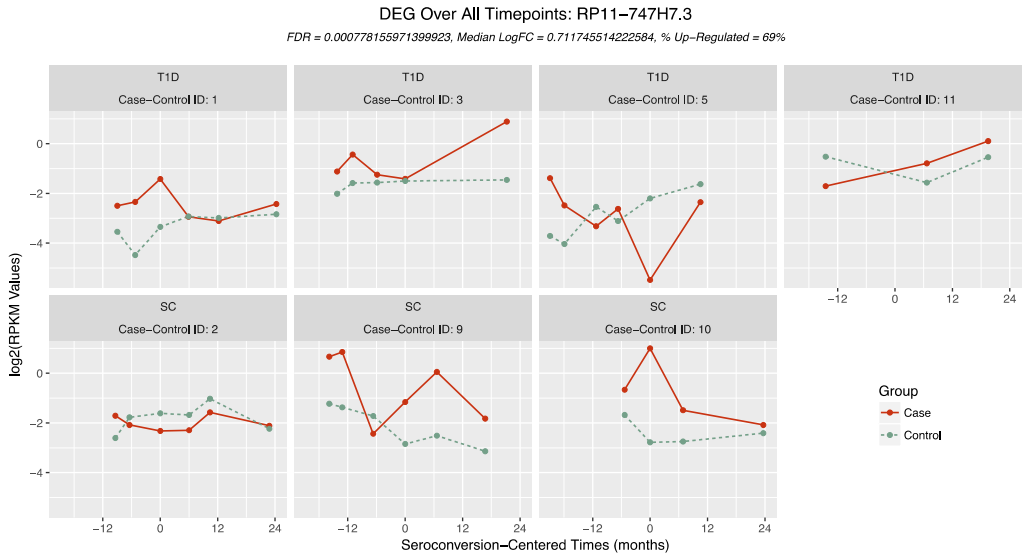


Y) Expression levels of VIPR1 gene in CD8+ cells.

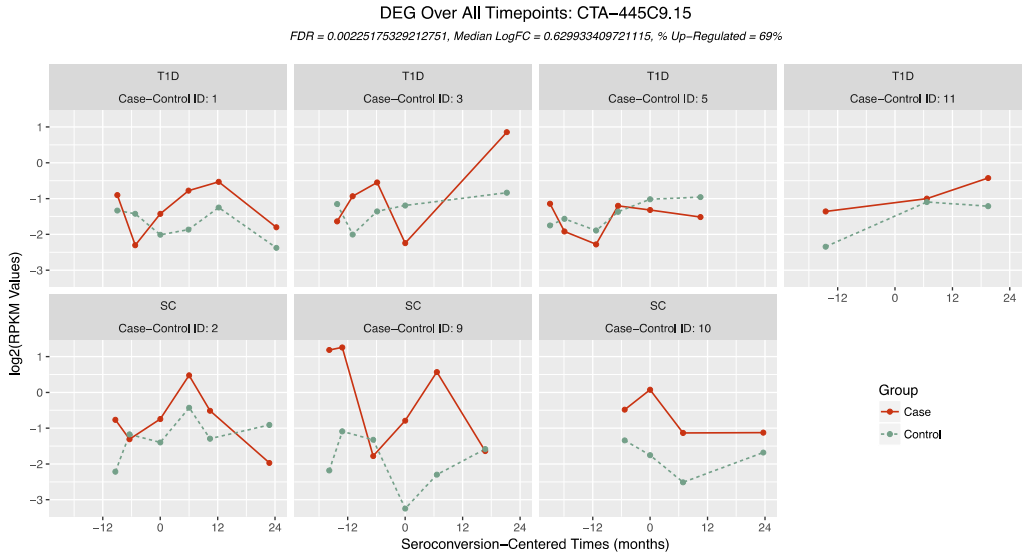


Z) Expression levels of PRKCQ-AS1 gene in CD4-CD8- cells.

SUPPLEMENTARY DATA



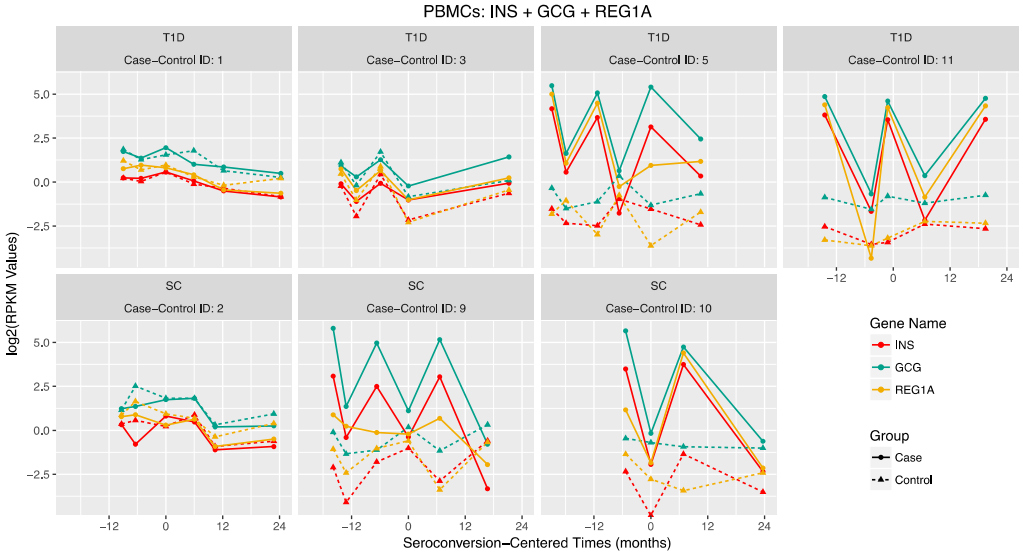
AA) Expression levels of RP11-747H7.3 gene in CD4+ cells.



AB) Expression levels of CTA-445C9.15 gene in CD4+ cells.

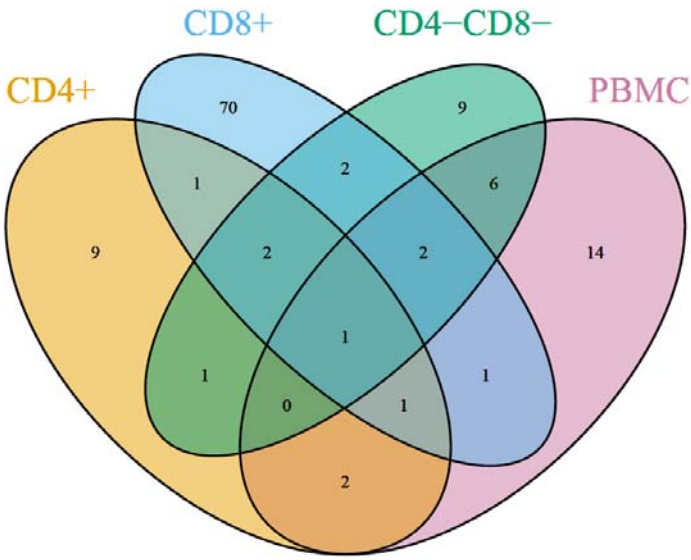
SUPPLEMENTARY DATA

Supplementary Figure S5. Related to Figure 2.
PBMC-specific co-regulation of pancreatic transcripts Insulin (INS), Glucagon (CGC) and Regulin 1 alpha (REG1A). Profiles of *INS*, *CGC* and *REG1A* show concerted gene expression profiles. For individual profiles, see Supplementary Figure 4.



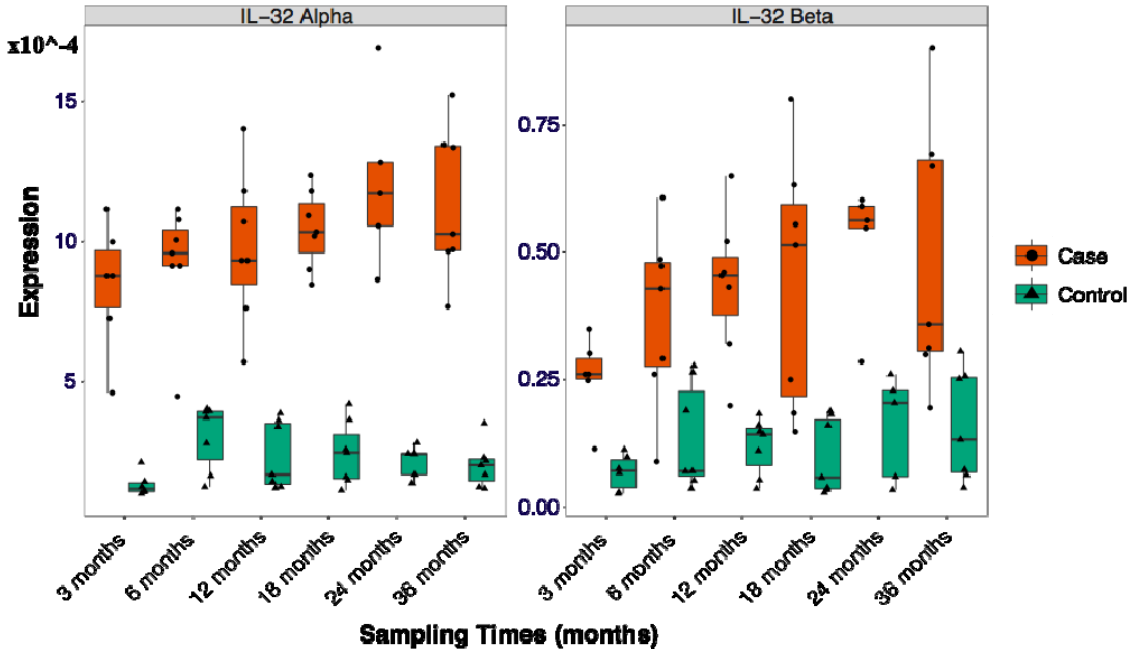
SUPPLEMENTARY DATA

Supplementary Figure S6. Related to Figure 2.
Number of DE genes between the Cases and Controls in the time-window of 12 months before seroconversion (RPKM > 3 for coding genes and RPKM < 0.5 for non-coding genes, Up- or downregulated in ≥ 65% of the Cases). For complete listing, see Supplementary Table 3 columns "12 mo before SC".



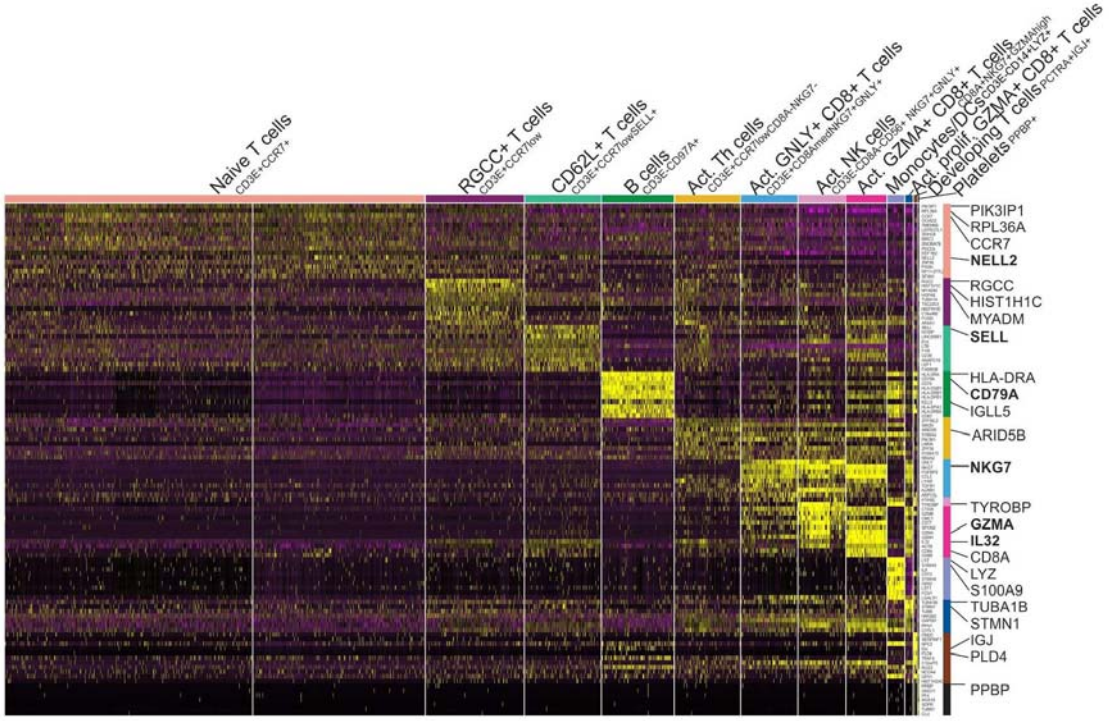
SUPPLEMENTARY DATA

Supplementary Figure S7. Related to Figure 3A.
mRNA expression of IL32 isoforms analysed in PBMC samples of Cases (n=7) and their matched Controls (n=7) by qRT-PCR. For expression level plot of IL32γ isoform, please refer to **Figure 3A.**



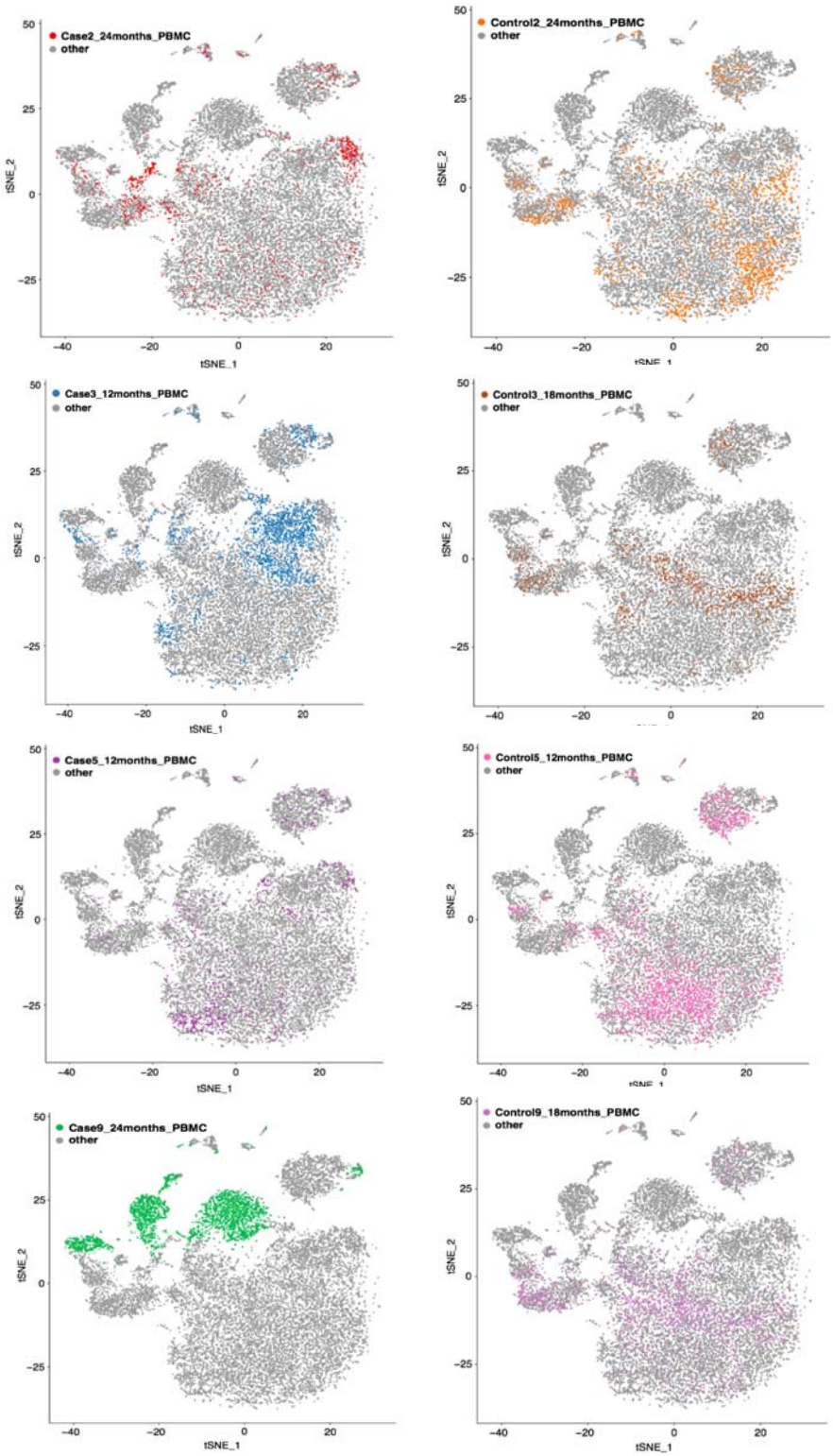
SUPPLEMENTARY DATA

Supplementary Figure S8. Related to Figure 3.
Heatmap of the top 10 most highly expressed genes in the 13 clusters identified after Seurat clustering analysis of the pooled single-cell RNA-Seq data (4 Cases + 4 Controls) represented as a tSNE plot in **Figure 3B**. As the two biggest clusters were similar in their gene expression profiles, they were merged to form the Naive T cell cluster, leaving in total 12 cell clusters. Genes used in the annotation of the cell clusters are marked on the right column, where bolded genes are those that were also found to be DE between Cases and Controls in the bulk RNA-seq data analysis (**Supplementary Table 3**).



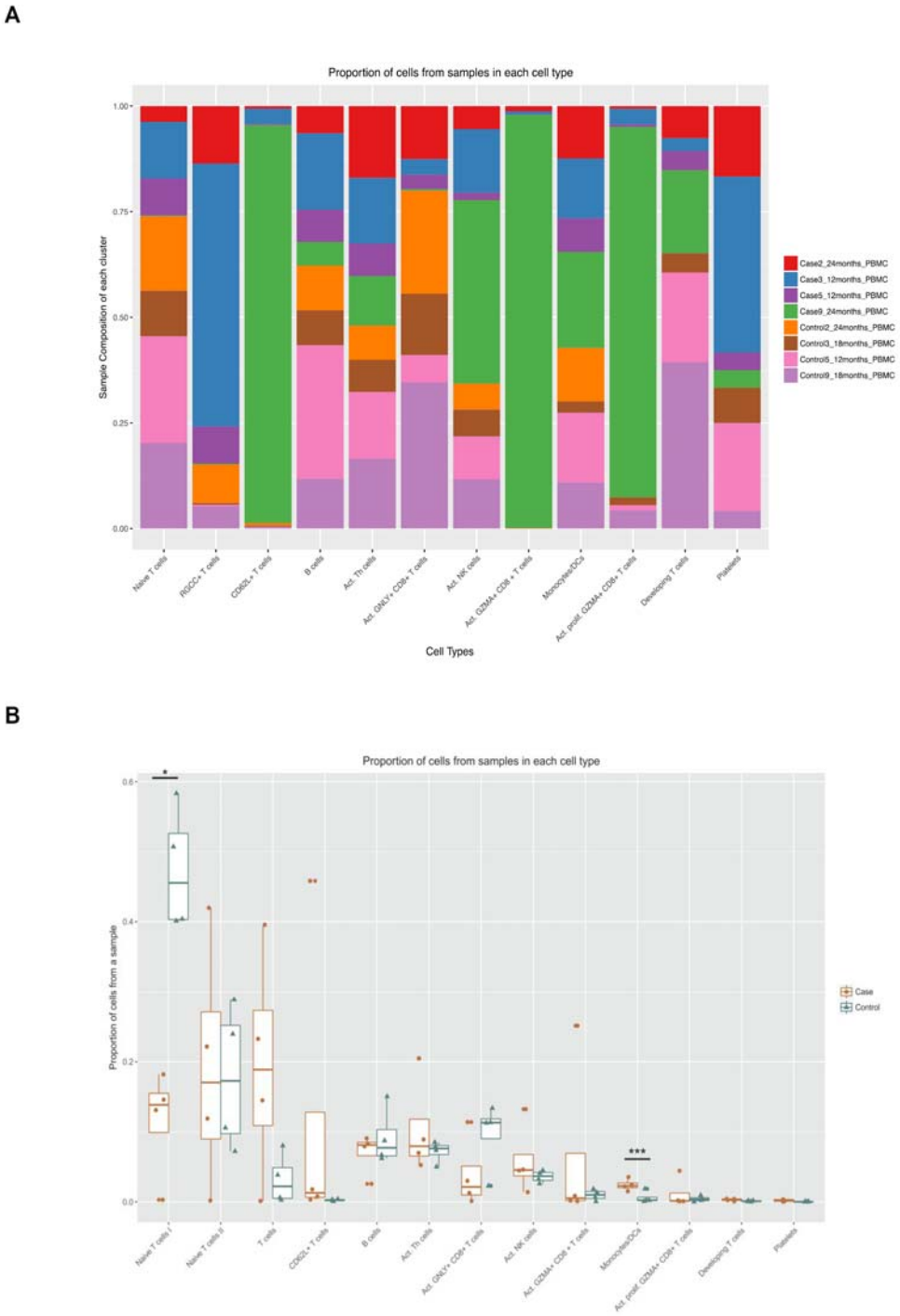
SUPPLEMENTARY DATA

Supplementary Figure S9. Related to Figure 3.
Contribution of individual samples in the t-SNE visualization of pooled single-cell RNA-Seq data, presented in **Figure 3B** and **C**.



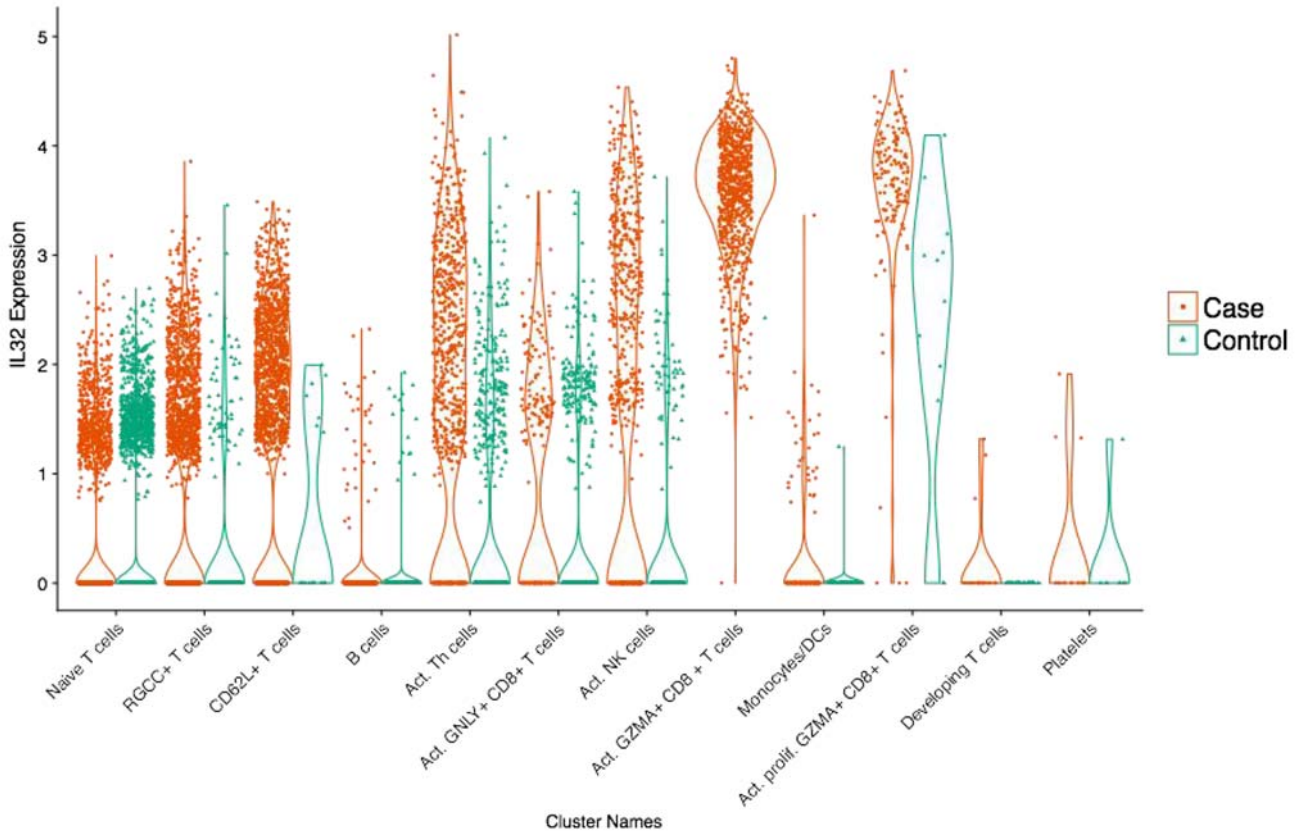
SUPPLEMENTARY DATA

Supplementary Figure S10. Related to Figure 3.
A) Proportion of cells coming from individual samples per cluster (cluster-wise proportioning) **B)** Box-plot highlighting the proportions of cells per cluster in Case (orange) and Control (green) samples. * $p < 0.05$, *** $p < 0.005$ according to paired t-test of the sample-wise proportions of cells per cluster.



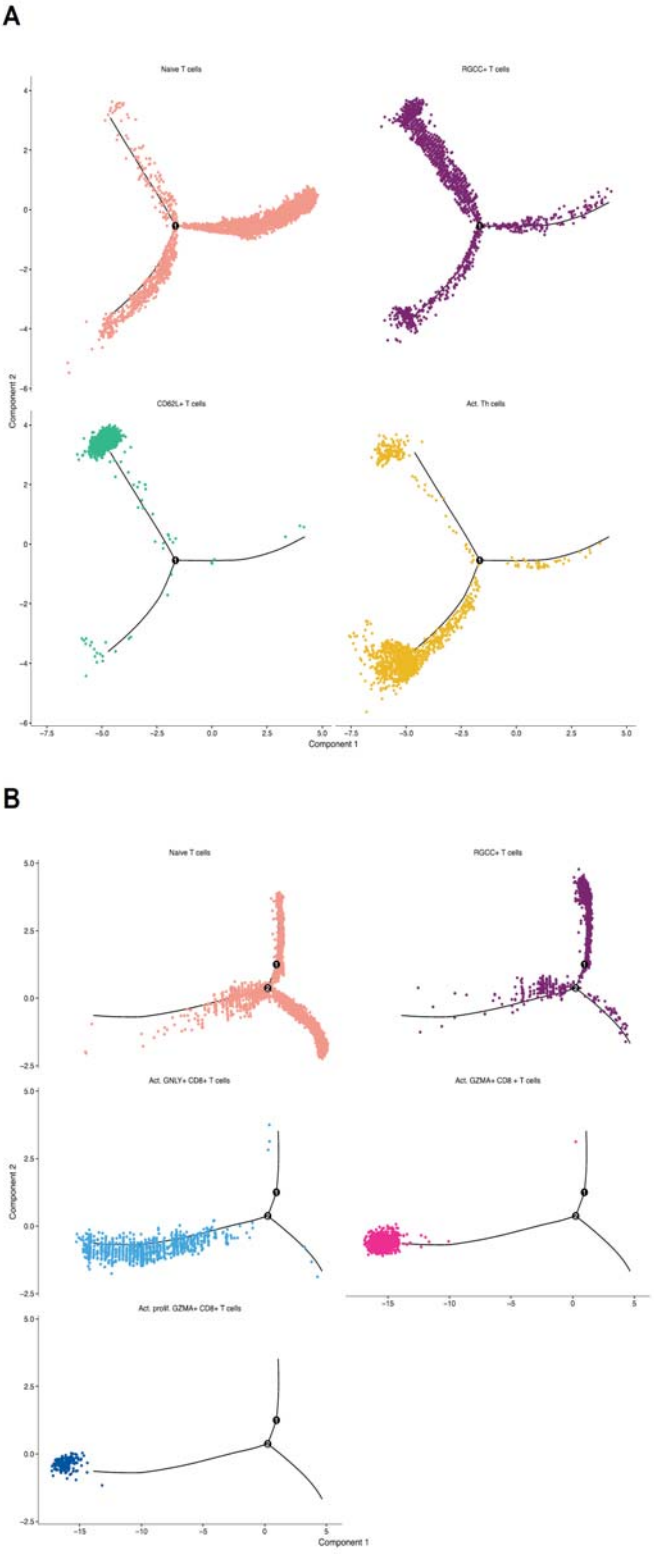
SUPPLEMENTARY DATA

Supplementary Figure S11. Related to Figure 3.
Violin plot showing the expression of *IL32* in the 12 cell clusters identified from the single-cell RNA-Seq data, displayed separately for Cases (orange) and Controls (green).



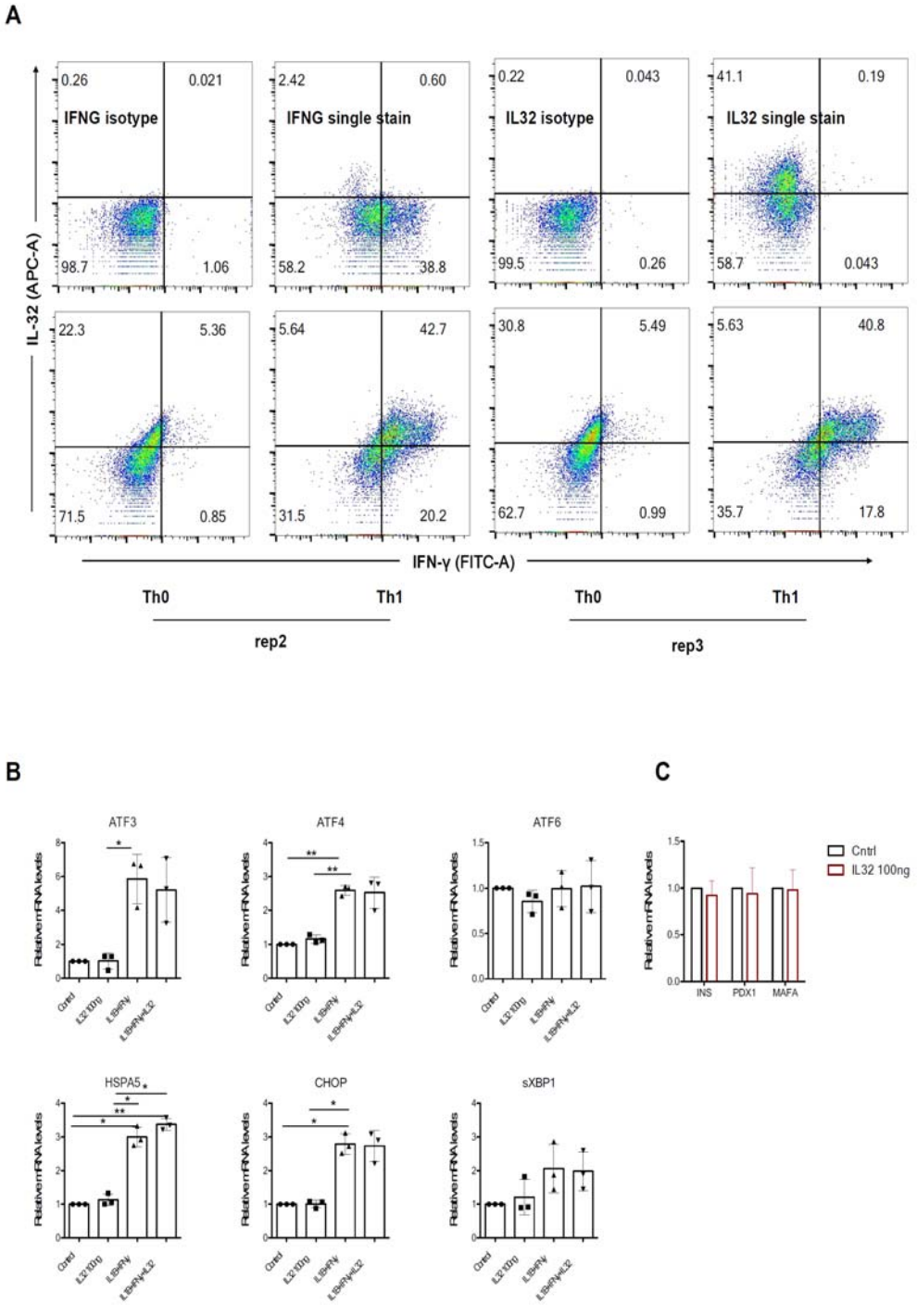
SUPPLEMENTARY DATA

Supplementary Figure S12. Related to Figure 3.
A) Trajectory in pseudotime of CD4+ specific and **B)** CD8+ specific cells along with the precursor cells plotted individually.



SUPPLEMENTARY DATA

Supplementary Figure S13. Related to Figure 4.
A) Two additional replicates of the Th0/Th1 intracellular staining data shown in **Figure 4C**. **B)** EndoC-βH1 cells were treated for 24 h with recombinant IL-32□ in presence and absence of IL-1β and IFNγ, and the expression of ER stress markers *ATF3*, *ATF4*, *ATF6*, *CHOP*, *HSPA5* and *sXBP1* was measured by RT-qPCR assay. **C)** Expression of endocrine marker genes *INS*, *PDX1* and *MAFA* was measured after treatment of EndoC-βH1 cells with 100 ng of IL-32γ for 24h. In **B-C** fold change is calculated as compared to non-treated (control) cells. Statistical significance was determined by Tukey's multiple comparisons test. * =p-value <0.05 while ** =p-value <0.01.



SUPPLEMENTARY DATA

Bulk RNA-seq data analysis

RNA-seq data processing and analysis

Of the 306 RNA-seq samples (**Supplementary Table 1**), 298 were used for the differential expression analysis because some Case samples had more than one corresponding control samples. The average sequencing depth of the samples in this study was around 51 million paired-end reads. Quality control checks were performed on the raw RNA-seq data using FastQC (version 0.10.0, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were aligned to the human reference transcriptome, Human GRCh37 assembly version 75 (http://feb2014.archive.ensembl.org/index.html), using TopHat (version 2.0.10), where the default parameters of the software were retained. On average, approximately 93% of the reads from each sample in each fraction were mapped (average overall read mapping of samples in each cell type was CD4+: 93%, CD8+: 92.6%, CD4-CD8-: 93.4%, PBMC: 93.5%). This resulted in about 89% of concordant pair alignments (CD4+: 88.9%, CD8+: 88.93%, CD4-CD8-: 89.89%, PBMC: 89.89%). The aligned reads, with a mapping quality > 10, were counted at the gene level availing the *htseq-count* function from the HTSeq package and using the overlap resolution mode of 'intersection-strict' (htseq-count version 0.6.1). The read counts of genes were normalized using the trimmed means of the M-values (TMM) method, implemented in the software package *edgeR*, which adjusts for varying sequencing depths as well as normalizes for the RNA composition. Using the biotype information, the genes were divided into coding and non-coding categories. The biotype data for each gene were retrieved from the Ensemble database , and the descriptions of biotypes were taken from Gencode (http://www.gencodegenes.org/gencode_biotypes.html).

Filtering genes using RPKM values

First, the RPKM values were calculated for each gene in each sample of the analysis, where the length of the gene was taken to be the sum of the lengths of all its known exons. Second, a max-of-means RPKM value (mmRPKM) was computed for each gene to assess the overall expression of the gene in all the samples of the analysis. As the differential expression analyses in this study usually involved two groups (e.g., cases and controls, CD4+ and PBMCs), the max-of-means of RPKM value refers to: maximum(mean("RPKM values in Group 1"), mean("RPKM values in Group 2")). Subsequently, coding genes with mmRPKM > 3 and non-coding genes with mmRPKM > 0.5 were retained. These filtering criteria usually retained about 7000–7500 coding genes and 600–700 non-coding genes.

Differential expression analysis

Differential expression analyses were conducted separately for coding and non-coding genes, using the *edgeR* package. The variance of the data was estimated using the trended dispersion method.

Post-differential analysis filtering steps (only for paired sample analyses)

The *edgeR* output of differentially expressed (DE) genes with FDR < 0.05 from the paired sample analyses were further subjected to median $\log_2$FC filtering, where DE genes with a |median $\log_2$FC| > 0.5 were retained for downstream filtering step. The final filtering step retained only those genes as DE that had more than 65% samples across all individuals regulated in the same direction (i.e., up- or down-regulated). These filtering steps were added to discard false positives that may arise due to the heterogeneity of the samples due to normal variation, which is non-related to T1D and outliers. A visual depiction of the RNA-seq data processing and analysis pipeline has been shown in **Supplementary Figure 2**.

Analysis 1: Cell fraction vs PBMC

In this analysis, the expressions of genes from each cell fraction (i.e., CD4+, CD8+ and CD4-CD8-) were compared to those of the (paired) original PBMC population of Control children. Samples

collected at all ages were included but were required to have expression data from both the fraction under analysis as well as PBMCs.

Analysis 2: Cases versus Controls – over all timepoints
The aim of this analysis was to identify genes that are differentially expressed in children who have seroconverted to autoantibody positivity (Cases) in comparison to those who have not (Controls). Each Case child was matched to a Control child, according to date of birth, HLA-risk class, gender and country. Case samples were compared to the samples from their matched Controls that were collected at the same age. In these analyses, other than for pairing purposes, the sampling ages were not utilized.
Differential expression analysis of Cases and Controls were compared with another method. The RNA-seq data of the filtered coding and non-coding genes were modelled using generalized linear mixed effects models (GLMMs) using *glmer()* function from *lme4* package (23). A random effect is added in this model for each child's samples. GLMM with the negative binomial likelihood was fit to the data using the *MASS* package, where the dispersion values per gene were obtained from the *edgeR*. For the filtered coding and non-coding genes from all fractions, the Spearman rank correlation coefficients of the results from the two methods ranged between 0.91 and 0.96 with an average of 0.936, indicating similar ranking of the genes after FDR correction in both the methods. Further details of RNA-seq data analysis can be found in the "Supplementary Material".

Analysis 3: Cases versus Controls – 12 months before the seroconversion window
This analysis is similar to **Analysis 2** in terms of the Case versus Control analysis set-up. However, to understand gene expression changes that take place in Cases right before seroconversion, this analysis compared only those Case samples that were taken at most 12 months before seroconversion with their matched Control samples.

For comparison, the RNA-Seq data of the filtered coding and non-coding genes were also modelled using generalized linear mixed effects models (GLMMs) using the same design as explained in **Analysis 2**. A random effect is added in this model for each child's samples. The *glmer()* function from *lme4* package was used here for modelling. GLMM with the negative binomial likelihood was fit to the data using the *MASS* package, where the dispersion values per gene were obtained from the *edgeR* **Analysis 2**. For the filtered coding and non-coding genes from all fractions, the Spearman rank correlation coefficients of the results from the two methods ranged between 0.91 and 0.96 with an average of 0.936, indicating similar ranking of the genes after FDR correction in both the methods.
Differential gene clustering
To find the genes and autoantibodies (together referred to as 'features' in this section) co-regulated/co-clustering with *IL32* in each cell-fraction, *k*-means clustering, followed with Euclidean distance based co-clustering selection criteria, was performed on the expression levels of coding and non-coding differentially expressed genes (**Analysis 2**) as well as on the autoantibodies. Due to the heterogeneity of the data and the disease, the clustering was done individually on each case and its matched control. Before clustering, the RPKM expression values of each gene and expression level of each autoantibody were $\log_2$ transformed to ensure approximately normal distribution of the values, and gene-wise standardized to make the features comparable.

For each possible number of clusters (i.e., from 2 to total number of features - 1), the features were clustered using the *k*-means clustering algorithm (*kmeans* function implemented in R *stats* package). Subsequently, using the resulting classification of features into clusters along with the Euclidean distance measures between the features, a silhouette score was calculated. The optimum number of clusters was chosen to be the one with the largest silhouette score. The features were then clustered into the "optimum number of clusters" using *k*-means clustering with 20 random sets of initialization values

and sufficient iterations for convergence, where the configuration with minimum loss score was reported as the best clustering. Once clustered, the cluster containing *IL32* was considered the *IL32*-cluster with its co-regulated features.

To summarize over the *IL32*-clusters from the seven case-control pairs, a feature co-clustering with *IL32* in at least one case-control pair was considered to co-cluster with *IL32* if the median of its Euclidean distances to *IL32* across all pairs was below 2.5. This selection criteria, based on median Euclidean distance to *IL32,* ensured that only those features were considered to co-cluster with *IL32* that co-clustered with it in at least 5–6 case-control pairs (**Supplementary Table 4**).

As *IFNG*-cluster in CD8+ cells and *INS*-cluster in PBMCs were of specific interest also, the Euclidean distance-based summarization over the seven case-control pairs was repeated for these genes as well (**Supplementary Table 4**).

Transcription factor binding site analysis

Overrepresented transcription factor binding motifs on the promoters of *IL32* and its co-regulated genes were analysed with updated (2018) TRANSFAC database, using the Fmatch tool with default parameters (best supported promoter, -10,000 to +1000 bp around transcription start site) and a randomly selected gene set as a background. Afterwards the *p*-values were corrected for multiple testing using the Benjamini-Hochberg method. Results with FDR < 0.05 are presented in **Supplementary Table 4**.

Single-cell RNA-seq data processing and analysis

The Chromium single-cell 3' RNA-Seq data from four Case and four Control samples (**Supplementary Table 5**) was individually preprocessed using the Cell Ranger Single-Cell Software Suite. The reads were aligned to the human reference genome (hg19) using STAR and the data from non-cellular barcodes were filtered out. Across samples, the mean raw reads per cell varied between ~57 k to ~200 k (**Supplementary Table 5**). To identify rare cell types, the cells from different samples were pooled together using Cell Ranger's multi-library aggregation algorithm where the samples were normalized using subsampling normalization. The downsampling (subsampling normalization) of sample reads after pooling retained on average ~31 k confidently mapped reads per cell (from ~59 k raw reads per cell on average). These mapped to the median of 801 genes per cell. After the pooling, expression of 32,738 genes from 20,370 cells was obtained.

For QC analysis and further exploration of the single-cell RNA-Seq data the Seurat R package was used. Firstly, all the genes expressed in less than one cell and all the cells expressing less than 200 genes or more than 4000 genes were filtered out. Furthermore, any cells containing more than 5% of mitochondrial genes or a UMI count higher than 5000 but a gene count less than 500, were also filtered out. The latter filtering steps involved filtering of cells with high UMI count but low gene count on the basis of the gene count and UMI count relationship plots following the recommendations of Seurat tool. After these quality control filtering steps, 18,396 cells expressing 20,830 genes were retained for downstream analyses.

The filtered data were normalized using Seurat's default global-scaling normalization method, 'LogNormalize', and variation from uninteresting sources (i.e., the number of molecules detected and percentage of mitochondrial genes expressed per cell) was regressed out. To capture the heterogeneity of the single-cell data and cluster the cells, a set of highly variable genes (HVGs) was selected, whose average expression was above 0.0125, and dispersion above 0.5 resulting in ~1200 HVGs in pooled cell library. Principal component analysis (PCA) was then performed on the HVGs, and the resulting top 20 PCs were used in the graph-based clustering employed by Seurat, keeping other parameters as default.

To determine the cell type represented in each cluster, markers defining the clusters were determined via differential expression algorithm implemented in Seurat, where cells of a single cluster were compared to the cells of all other clusters combined. A gene was considered a marker of a cluster if it was

expressed in at least 25% of the cells of either of the two clusters and the log fold change between the cluster and all other clusters was at least 0.25. On average, one to five genes were used as markers for each cluster (**Supplementary Figure 8**). On the basis of these cluster-specific markers, no biological difference was found in two of the 13 clusters, which both represented cells from naive T cells. Therefore, they were merged into a single cluster and were labeled as naive T cells, resulting in a total of 12 different clusters.

Single-cell RNA-seq trajectory analysis

The QC filtered pooled cells from the Seurat analysis were ordered in pseudotime (i.e., placed along a trajectory corresponding to a type of biological transition, such as differentiation) using Monocle 2. The trajectory analysis was performed on cells specifically from CD4+ (CD62L+ T cells and Act. Th cells) and CD8+ (Act. GNLY+ CD8+ T cells, Act. GZMA+ CD8+ T cells and Act. prolif. GZMA+ CD8+ T cells) T-cell clusters, using the cell typing information from the Seurat analysis. In both CD8+ and CD4+ specific cell ordering, cells identified as naive T cells or T cells were also included. The trajectory analysis in Monocle 2 has three major steps.

In the first step, all genes expressed in at least 1% of the cells were used in a principal component analysis, whose resulting top PCs (six in the case of CD8+ and 11 in the case of CD4+ specific single-cell trajectory analyses) were used to initialize the t-SNE ordination of the cells. Then, the *dpFeature* function was used to cluster the cells defined in the 2-D t-SNE space. Finally, the differential gene expression test of all genes expressed in more than 10 cells was performed between the clusters defined in the previous step as a way to extract the genes that distinguish them from each other. The top 1000 significant genes were then selected for subsequent steps of the analysis. The second step reduced the dimensionality of the data using the feature genes from the previous step and availing technique called reverse graph embedding (RGE) implemented in DDRTree algorithm. In the final step, cells were ordered along the trajectory by performing manifold learning on the tree from the second step.