SUPPLEMENTARY DATA

Supplementary Methods

Detailed description of the Biomarker discovery pipeline

Cross validation (k x M-fold cross-validation)

1. M-fold cross validation
   a. For an N x P dataset (S), randomly split the dataset into M-folds (datasets, S^1, …, S^M) of size N/M such that each fold contains roughly equal number of subjects from both groups (controls vs. estT1D, controls vs. newT1D and estT1D vs. newT1D).
   b. For i = 1 to M (i.e. repeat until each fold has been used as a test set)
      i. Let S^i be the test dataset and S^{-i} be the training dataset
      ii. Fit an elastic-net regularized logistic regression model (elastic-net model), $f^i(\text{input}, \alpha)$ using the glmnet R-library using a set α level ($\alpha \in [0, 1]$)
      iii. Predict group label of test data S^i using training classifier: $f^i(\text{input}=S^i, \alpha)$
      iv. Using the correct test labels, compute the error rate (# of misclassified subjects/total number of subjects), and the area under the receiver operating curve (AUC). Using an arbitrary threshold of 0.5, compute the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Note: The prevalence of the disease was not used for the computation of the PPV and NPV.
   c. Compute the cross-validation (CV) estimate of the prediction error using:

$$CV(f,\alpha) = \frac{1}{M}\sum_{i=1}^{M} L(S^i, f^{-i}(input, \alpha))$$

   where *f*=final biomarker panel (model); L=loss function (computes error rate)
2. Repeat step 1, k number of times
3. For all performance measures, average over k x M-folds
4. The final biomarker panel (*f*) is built by fitting an elastic-net model to all the data at the same α value used the CV.

For the 30-gene panel (controls vs. newT1D):

1. The classification performance measures were obtained from a 200x5-fold cross-validation (CV) setting α=0 using a 53 samples (24 controls, 29 newT1D) x 30 genes dataset.
2. The final panel was built by fitting the elastic-net model to all the data, setting α=0 (no variable selection).

For the 6-gene panel (controls vs. newT1D):

1. To build a panel with a smaller number of genes, a 200x5-fold CV was used setting α=0.65 (variable selection) using a 53 samples x 30 genes dataset.
2. The percentage of times a gene was included in the 200x5 = 1000 panels of the CV at α=0.65 was calculated. Using a frequency cut-off of 65% resulted in a 6-gene highly stable (with respect to the cross-validation folds) biomarker panel.
3. The classification performance (biased) measures were obtained by applying a 200x5-fold CV using a dataset only consisting of the 53 samples x 6 genes and setting α=0.
4. The final panel was built by fitting the elastic-net model to a 53 samples x 6 genes dataset, setting α=0. Note: The estimate of the test error of the 6-gene panel is unknown and can only be provided by using an independent set of samples.

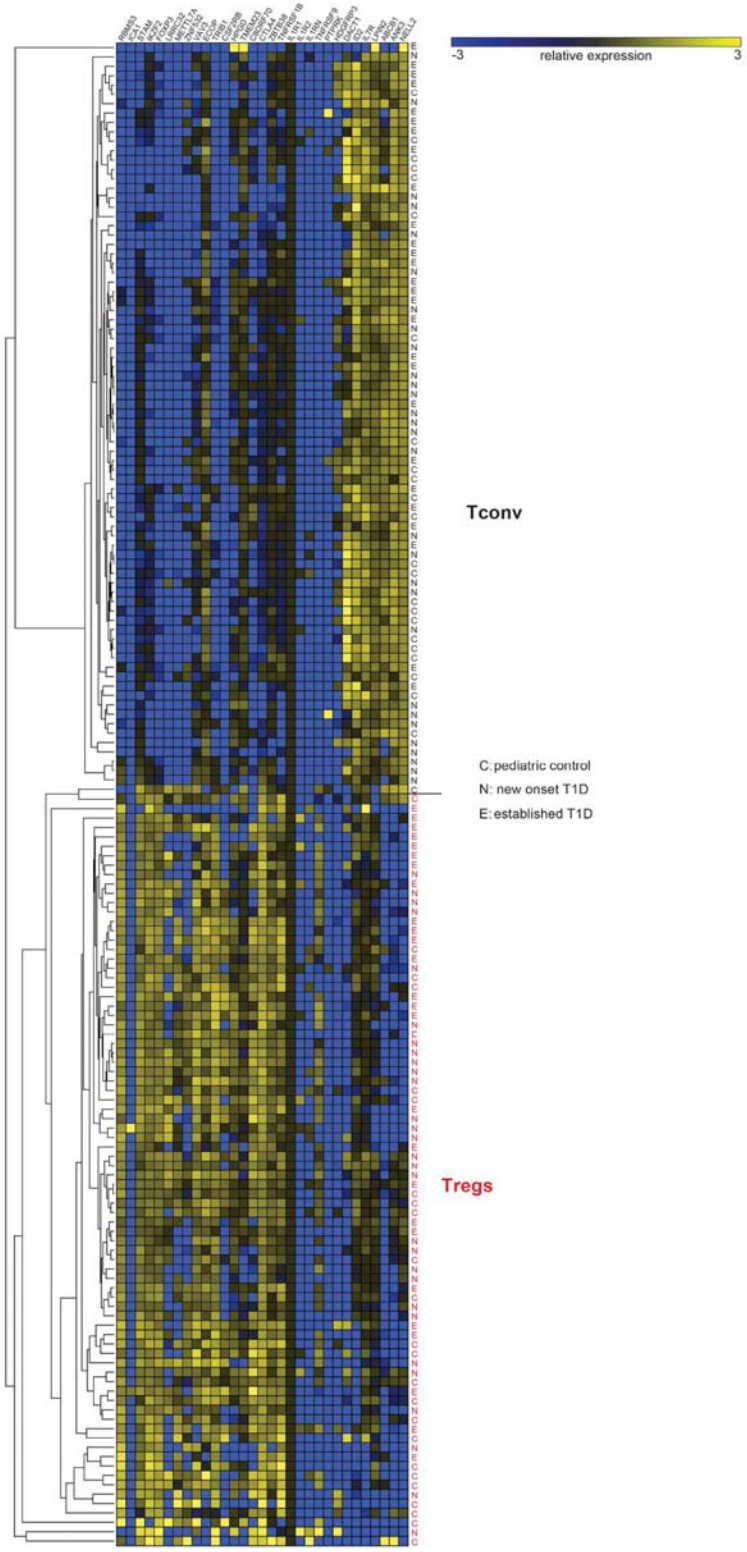
SUPPLEMENTARY DATA

Important notes:

1. An arbitrary threshold of 0.5 was used to compute the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Additional independent datasets are required to determine the optimal threshold value.
2. The CV estimate of the 6-gene panel is biased since the same dataset was used to determine the stable genes and classification performance. The test error rate can be estimated by used additional independent datasets.

SUPPLEMENTARY DATA

Supplementary Figure S1. *Heatmap of the 31 gene signature in the pediatric cohort.* Normalized expression of the 31 genes with hierarchal clustering in the pediatric cohort. Each row is a gene, with the gene symbol on the right. C: pediatric control; N: new onset T1D; E: established T1D. Pediatric controls n=24, new onset T1D n=29,established T1D n=27).



SUPPLEMENTARY DATA

Supplementary Figure S2. *Comparison of 30-gene and 6-gene panel vs. random classifiers.* (A) General biomarker panel algorithm. (B) The biomarker panel algorithm for 30 genes was applied using pediatric control and new onset T1D Treg samples with weighting according to the 30-gene panel (blue) or with assignment of random weighting (green). (C) 30-gene panel algorithm with weighting. (D) The biomarker panel algorithm for 6 genes was applied using pediatric control and new onset T1D Treg samples with weighting according to the 6-gene panel (blue) or with assignment of random weighting (green). (E) 6-gene panel algorithm with weighting.